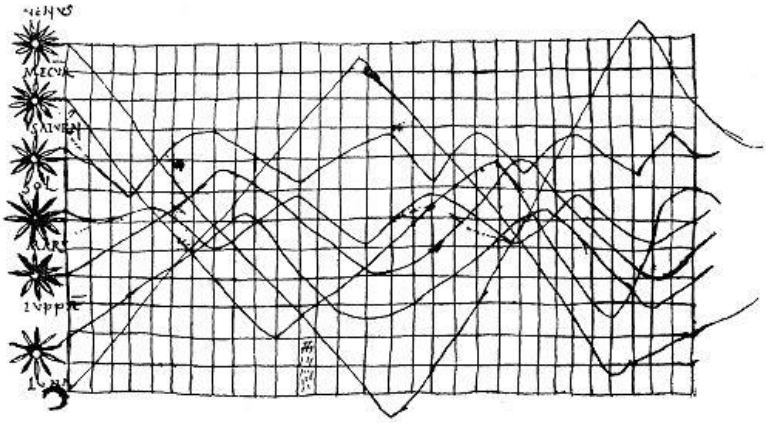
Wel en wee van grafieken



Jelke Bethlehem

Maart 2015

Inhoud

1. Waarom een grafiek? 1
2. Terug in de tijd 5
3. Ingrediënten van een grafiek 11
4. Richtlijnen voor grafieken 21
5. Grafieken voor één kwalitatieve variabele 35
6. Grafieken voor één kwantitatieve variabele 43
7. Grafieken voor twee kwalitatieve variabelen 47
8. Grafieken voor twee kwantitatieve variabelen 51
9. Grafieken voor gemengde variabelen 55
10. Tijdreeksen 57
11. Thematische kaarten 63
12. Literatuur 67

1. Waarom een grafiek?

In 2012 deed het Centraal Bureau voor de Statistiek onderzoek naar pesten op het internet. Daarbij ging het om zaken als roddel, getreiter, pesten, stalken, chantage en bedreiging. Aan een steekproef van Nederlanders werd gevraagd of ze wel eens gepest werden en op welke manier ze werden gepest. Dit onderzoek leverde allerlei interessante uitkomsten op. De vraag is dan hoe je deze uitkomsten het beste kunt presenteren. Dat kan op drie verschillende manieren.

De eerste manier is het schrijven van een tekst over de uitkomsten. Voor het onderzoek naar pesten op het internet zou die tekst er bijvoorbeeld zo uit kunnen zien:

> *In 2012 gaf 3 procent van de Nederlandse bevolking van 15 jaar of ouder aan weleens gepest te zijn via internet. Dit zijn naar schatting ruim 400 duizend personen. Jongeren worden veel meer met cyberpesten geconfronteerd dan ouderen. Van de 15- tot 18-jarigen had 10 procent in 2012 te maken met pesten via internet, tegen minder dan 1 procent van de 65-plussers. Natuurlijk maken jongeren ook meer gebruik van het internet en zijn ze actiever op sociale media, die vaak als platform voor pestgedrag dienen.*

Deze tekst bevat de belangrijkste uitkomsten van het onderzoek. Veel details staan erin. Toch is het lastig om het globale beeld eruit te halen.

Een tweede manier om de uitkomsten samen te vatten is in de vorm van een tabel. Een tabel is een numeriek overzicht waarin het vooral cijfers staan en haast geen tekst. Tabel 1.1.1 bevat de uitkomsten van het onderzoek naar cyberpesten.

Tabel 1.1.1. Slachtoffers van cyberpesten, naar leeftijd, in 2012

Leeftijd	Percentage slachtoffers
15 tot 18 jaar	10,3%
18 tot 25 jaar	7,1%
25 tot 45 jaar	3,1%
45 tot 65 jaar	2,1%
65 jaar en ouder	0,9%

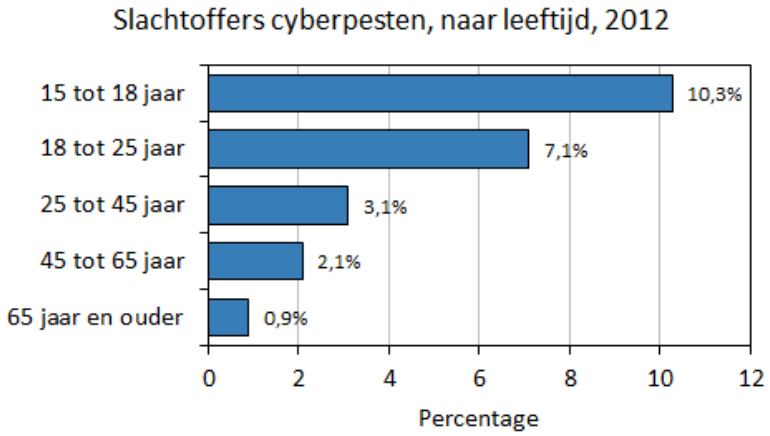
Deze kleine tabel is nog wel redelijk overzichtelijk. Zo is eenvoudig te zien in welke leeftijdscategorie cyberpesten het meest voorkomt (15 tot 18 jaar) en in welke groep het minst (65 jaar en ouder). Bij grotere en ingewikkelder tabellen is de interpretatie een stuk lastiger.

Een derde manier om de uitkomsten van een onderzoek te brengen, is met een grafiek. In een grafiek is vaak eenvoudiger te zien wat er aan de hand is. Dat maakt een grafiek geschikt voor een groot en breed publiek. Maar pas op: er zijn ook gevaren. Je kunt gegevens op allerlei manieren grafisch weergeven. Niet elke manier is goed. Het is vrij eenvoudig om mensen op het verkeerde been te zetten. Daardoor kunnen ze de verkeerde conclusies trekken. Het is daarom belangrijk dat je zorgvuldig bent bij het maken van een grafiek. En dat betekent ook dat je moet oppassen als je een grafiek van iemand anders onder ogen krijgt.

Er zijn veel verschillende typen grafieken in gebruik. Het hangt van het soort gegevens af welke type grafiek het geschiktste is om de 'boodschap' in de gegevens op een goede manier over te brengen. De sterke en zwakke punten van veel grafieken zullen aan de orde komen. Zo zullen we het hebben over cirkeldiagrammen, staafdiagrammen, histogrammen, puntenwolken, lijndiagrammen en thematische kaarten.

Voor de gegevens over cyberpesten komt het *staafdiagram* in aanmerking. Het resultaat staat in figuur 1.1.1.

Figuur 1.1.1. Een staafdiagram



In de grafiek is een duidelijke trend te zien: cyberpesten neemt af naarmate je ouder wordt. Cyberpesten komt het vaakst voor bij jongeren van 15 tot 18 jaar. Cyberpesten komt het minste voor bij ouderen vanaf 65 jaar. Dat is niet zo verwonderlijk, want ze zijn veel minder actief in de sociale media. Voor deze groep ligt het percentage onder de 1%.

De grafiek in figuur 1.1.1 bevatte maar vijf getallen (de vijf percentages). Een grafiek kan vooral zijn kracht tonen als hij veel gegevens in beeld brengt. Een grafiek kan onthullend zijn. Een grafiek kan het inzicht in de gegevens vergroten door het spontaan tonen van allerlei aspecten die in tekst of tabellen verborgen zouden zijn gebleven. Een grafiek vormt dus een effectief instrument voor het overbrengen van de boodschap die in een grote hoeveelheid gegevens opgesloten zit. De voordelen van een grafiek zijn al door Schmid (1983) op een rij gezet:

- Een grafiek brengt de boodschap in de gegevens snel en direct over. Je bent daardoor in staat om de belangrijkste aspecten van een grote hoeveelheid gegevens in één oogopslag te zien.
- Een grafiek is een krachtig instrument om een veelomvattend en samenhangend beeld te geven van de gegevens.
- Een grafiek heeft meer overtuigingskracht dan een stuk tekst of een tabel. Een grafiek doet meer dan het alleen maar het signaleren van een bepaald verschijnsel. Een grafiek laat ook daadwerkelijk zien dat dit verschijnsel zich inderdaad voordoet.
- Een grafiek kan onthullend zijn. Een grafiek vergroot het inzicht in de gegevens door het spontaan onthullen van tot dan toe verborgen gebleven aspecten. Daarmee stimuleert een grafiek analytisch denken en verder onderzoek.
- Een grafiek trekt door zijn aantrekkelijke uiterlijk (vergeleken met tekst en tabellen) eerder de aandacht van de lezer. Daardoor daagt hij lezers uit zich te verdiepen in het onderwerp.

Grafieken kun je voor twee verschillende doeleinden gebruiken: *presentatie* en *exploratie*. Bij presentatie van grafieken gaat het om het overbrengen van een boodschap door een onderzoeker aan anderen. De over te brengen boodschap staat centraal. Alles moet erop gericht zijn die boodschap zo eenvoudig en correct mogelijk over te brengen, ook bij een niet-statistisch geschoolde lezer. Dit soort grafieken vinden we bijvoorbeeld terug op de tv (in het journaal), in de krant en in het Statistisch Zakboek van het CBS.

Een tweede gebruik van grafieken is exploratie. Daarbij maakt een onderzoeker de grafiek niet voor een ander, maar voor zichzelf. Het is een hulpmiddel bij het onderzoeken van statistische gegevens. De grafiek moet helpen om onverwachte patronen en structuren aan het licht te brengen. Is er een bepaald soort samenhang? Zijn er objecten (personen, bedrijven) die zich afwijkend gedragen? Veel voorbeelden van het gebruik van grafieken voor exploratie zijn te vinden in Tukey (1977). Dit soort grafieken zullen niet vaak in publicaties terecht komen. Een onderzoeker zal er vele maken en ze misschien bijna allemaal weer weggooien.

In deze leidraad gaat het om het gebruik van grafieken voor de presentatie van statistische informatie. Het zal dus niet gaan over statistische exploratie van gegevens. Wie daarover meer wil weten, wordt verwezen naar de boeken van Tukey (1977) en Chambers et al. (1983).

'Eén plaatje zegt meer dan 1000 woorden'. Deze spreuk wordt vaak aangehaald om aan te geven hoe waardevol grafieken kunnen zijn. Er wordt dan ook nog bij gezegd dat dit een oud Chinees spreekwoord is. Dat is echter niet het geval. De spreuk is de vertaling van 'One picture is worth a 1000 words'. Die spreuk is in de 20-er jaren van de vorige eeuw bedacht om aan te geven dat het gebruik van plaatjes in advertentieteksten belangrijk is. Ook al is de spreuk geen oud Chinees spreekwoord, toch illustreert hij het principe dat één plaatje vaak beter in staat is om een boodschap over te brengen dan een groot stuk tekst.

Bovengenoemde spreuk is ook van toepassing op grafieken. Maar niet zonder meer, want er zijn goede en slechte grafieken. Een grafiek doet alleen zijn werk wanneer hij accuraat, relatief simpel, eenvoudig interpreteerbaar, overtuigend, onthullend en aantrekkelijk is. Helaas is dit vaak niet altijd het geval.

Een goede grafiek

Niet elke grafiek is een goede grafiek. Er zijn veel computerprogramma's waarmee je grafieken kunt maken. Dat kan bijvoorbeeld met statistische programmatuur zoals SPSS, SAS, Stata en R. Maar ook meer algemene software zoals het spreadsheetprogramma Excel bieden die mogelijkheid. Helaas kun je met deze software zowel goede als slechte grafieken maken. Die slechte grafieken kunnen, statistisch gezien, lang niet altijd door de beugel. Als je niet bent getraind in het doorzien van de grafische voetangels en klemmen, kun je een heel verkeerd beeld krijgen opgedrongen. Dat is vooral vervelend, omdat grafieken vaak worden gebruikt voor het overbrengen van statistische informatie aan het grote publiek. Onderzoekers, grafisch ontwerpers en computerprogrammeurs doen er daarom goed aan zich te bezinnen op wat nu eigenlijk goede grafieken zijn. Schmid (1983) geeft een aantal eisen waaraan een goede grafiek zouden moeten voldoen.

Nauwkeurigheid en aanzien

Je moet een grafiek nauwkeurig maken. Het mag niet zo zijn dat door onnauwkeurigheid of onzorgvuldigheid van de maker (mens of computer) een misleidend beeld ontstaat. Het aanzien van een grafiek bepaalt of hij belangstelling zal opwekken. Een goede grafiek doet dit vanwege zijn nette, waardige en professionele uiterlijk

Eenvoud

Hoofddoel van een grafiek moet zijn het overbrengen van de statistische boodschap in de gegevens. Dat moet zo eenvoudig mogelijk gebeuren. Daarbij moet je niet ter zake doende tekst en versiering vermijden. Dat leidt de aandacht af van de boodschap. Tufte (1983) duidt al deze overbodige poespas aan met 'chartjunk'.

Helderheid

De grafiek moet een objectieve statistische boodschap overbrengen. De figuur moet duidelijk zijn en niet tot dubbelzinnige of verkeerde interpretatie leiden. Je moet zonder al te veel moeite de juiste boodschap eruit kunnen halen. Een grafiek is bedoeld om tijd en inspanning te besparen. Het mag dus geen puzzel zijn.

Vormgeving

Er zijn veel verschillende typen grafieken. Het type grafiek dat je kiest moet in overeenstemming zijn met de over te brengen boodschap. Daarbij mag je functioneel gebruik maken van kleuren of grijstinten. Je moet optische illusies vermijden. Verschillen in waargenomen contrast moeten indicatief zijn voor verschillen in waarden. Je moet je bij het maken van de grafiek bewust zijn van de fysiologische en psychologische processen die zich in de hersenen afspelen bij het waarnemen en interpreteren van grafieken.

De hierboven genoemde eisen hangen met elkaar samen. Zo bepaalt de vormgeving het aanzien van de grafiek en de nauwkeurigheid de helderheid. Bij het maken van een grafiek moet je deze eisen daarom gezamenlijk in acht nemen.

We zullen niet ingaan op de cognitieve aspecten van de perceptie van grafieken. Meer daarover is bijvoorbeeld te vinden in Few (2004). Ook Snijkers (1985) beschrijft de cognitieve aspecten van grafieken.

Er zijn heel wat boeken geschreven over het maken van grafieken en wat daarbij mis kan gaan. Voorbeelden zijn Tufte (1983) en Wainer (1997). Een nuttige internationaal georiënteerde publicatie over het maken van grafieken is de gids van de Verenigde Naties. Zie hiervoor United Nations (2009).

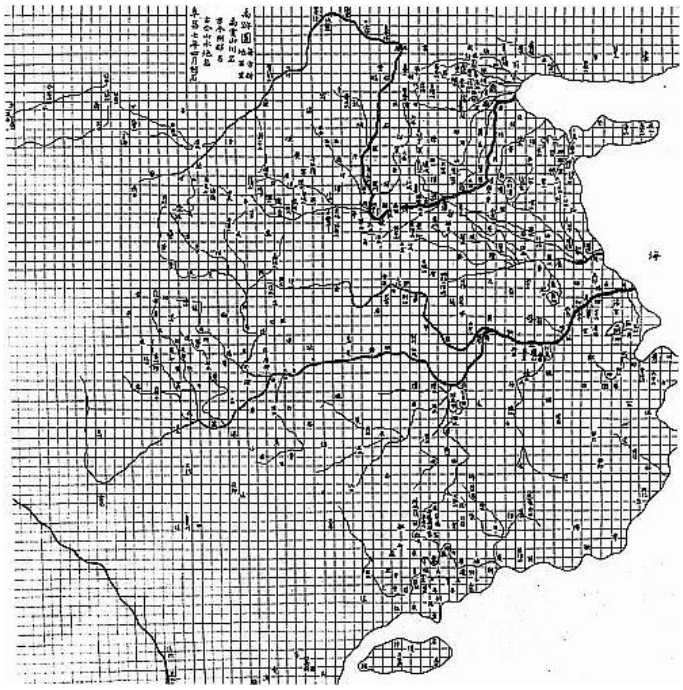
2. Terug in de tijd

Het overbrengen van informatie via tekst is lastig. Zeker wanneer de te communiceren informatie complex van karakter is, is het een bijna onmogelijke taak. Pogingen om informatie op een andere dan tekstuele wijze over te brengen gaan dan ook al terug tot in de grijze oudheid. Daarbij ging het aanvankelijk vooral om het weergeven van ruimtelijke informatie op landkaarten. Vermoedelijk werden de eerste landkaarten al zo'n 5000 jaar voor Christus gemaakt. De eerste bekende kaart dateert van ongeveer 600 voor Christus. Die kaart is in Babylonië gemaakt en wordt de *Imago Mundi* genoemd. Zie figuur 2.1. De kaart is getekend op een kleitablet en toont de stad Babylon aan de Eufraat en het gebied er omheen.

Figuur 2.1. De oudst bekende kaart: Imago Mundi

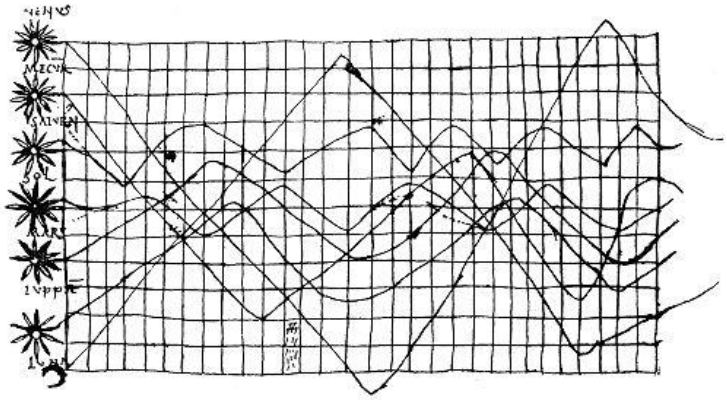
Een andere bekende oude landkaart is de *Yuji-tu-kaart*. De kaart is getekend in 1137 na Christus. Het is een behoorlijk nauwkeurige kaart van China. Tufte (1983) bespreekt de kaart in meer detail. Al die kaarten hadden gemeen dat het echte landkaarten waren en geen grafieken.

Figuur 2.2. Een kaart van China uit de 12de eeuw.



De grafieken in figuren 2.1 en 2.2 bevatten geen statistische informatie. Dat geldt ook voor de grafiek die is afgebeeld in figuur 2.3. Volgens Funkhouser (1936) is dit de eerste echte grafiek in de geschiedenis. Ook Tufte (1983) bespreekt deze grafiek.

Figuur 2.3. De oudst bekende grafiek van een tijdreeks



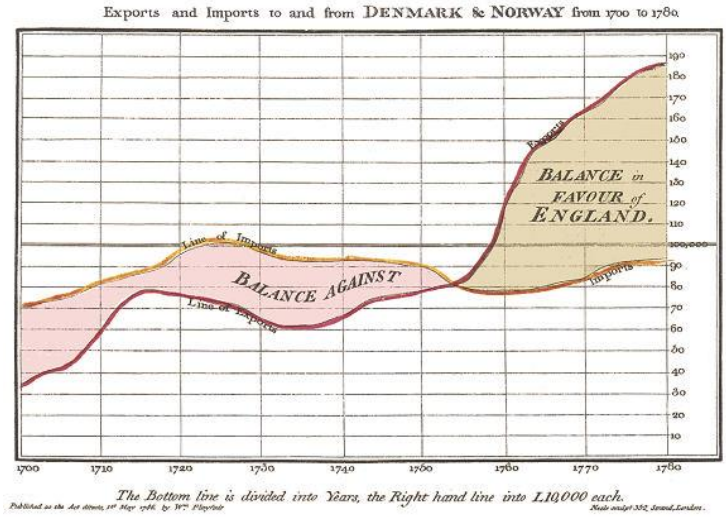
De grafiek is vermoedelijk gemaakt in de 11de eeuw in een klooster in Europa en beschrijft de inclinatie (hoogte boven de horizon) van de planeetbanen als functie van de tijd. De grafiek stond in een leerboek van een kloosterschool. Figuur 2.3 is het oudst bekende voorbeeld van een grafiek waarin veranderingen in de tijd van een verschijnsel in beeld zijn gebracht. Het is ook een unieke grafiek in de zin dat de volgende grafiek van een tijdreeks pas honderden 800 jaren later verscheen.

De eerst grafiek met statistische informatie is misschien wel de thematische kaart die de Amsterdamse uitgever Jodocus Hondius in 1607 publiceerde. De kaart maakt deel uit van de Atlas Minor, een pocket-versie van de Atlas van Mercator. De kaart toont de verspreiding van de verschillende godsdiensten. Zie figuur 2.4. Kleine symbolen op de kaart gaven aan welk godsdienst men in het betreffende gebied aanhing. In de uitvergroting zijn die symbolen wat duidelijker te zien.

Figuur 2.4.De thematische kaart van Jodocus Hondius

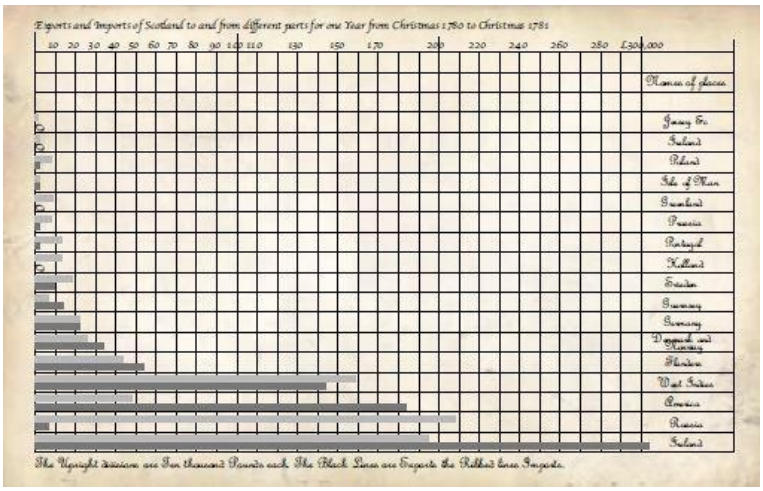
In de 18de eeuw was er een doorbraak bij het gebruik van grafieken. Toen presenteerde de Engelsman John Playfair (1759 - 1823) voor het eerst niet-ruimtelijke gegevens op een grafische manier. In 1786 publiceerde hij een boek met daarin 44 grafieken waarin hij economische variabelen in beeld bracht. Zie Playfair (1786). Van de in totaal 44 grafieken ging het in 43 gevallen om tijdreeksen. Figuur 2.5 bevat een van die grafieken. De figuur wordt ook besproken in Tufte (1983). De grafiek brengt de handel van Engeland met Denemarken en Noorwegen in beeld. Playfair gaf de voorkeur aan dit soort grafieken boven gegevens in tabelvorm omdat ze in zijn ogen de lezer beter in staat stelden om allerlei ontwikkelingen te vergelijken.

Figuur 2.5. Eén van de eerste tijdreeksen van economische variabelen



De enige grafiek in het boek van Playfair die geen tijdreeks is, blijkt een staafdiagram te zijn. Hij is gereproduceerd in figuur 2.6. Hij beschrijft de handel van een aantal landen met Schotland. De staven zijn geordend naar de omvang van de handel. Daarmee is Playfair ook de uitvinder van het staafdiagram.

Figuur 2.6. Eén van de eerste staafdiagrammen



Grafieken kunnen een krachtig instrument zijn voor exploratief onderzoek van gegevens. Een heel goed voorbeeld daarvan is de grafiek die in 1854 door John Snow werd gemaakt. Op een kaart van het centrum van Londen gaf hij alle sterfgevallen aan Cholera weer door middel van

stippen. Bovendien tekende hij de 11 beschikbare waterpompen op de kaart (de kruisjes). Deze kaart is gereproduceerd in figuur 2.6. Zie ook Tufte (1983) en Wainer (1997).

Door de verdeling van de punten te bestuderen, kwam Snow erachter dat de sterfgevallen zich vooral concentreerden rondom de waterput in Broad Street. Hij trok de conclusie dat de ziekte iets te maken had met het water dat uit deze pomp kwam. Nadat de pomp buiten gebruik was gesteld (door het verwijderen van het handvat) kwam er al snel een einde aan een epidemie die meer dan 500 slachtoffers had geëist.

Figuur 2.7. Sterfte aan Cholera in het centrum van Londen in 1864

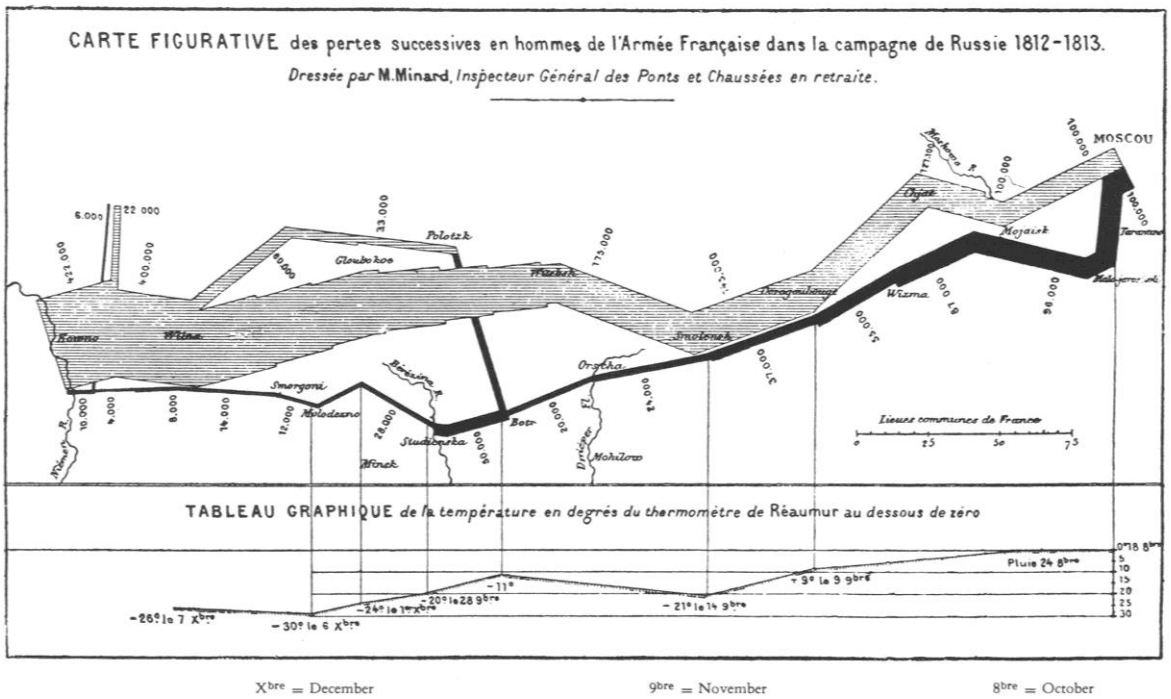
Ook het werk van de Fransman Charles Joseph Minard mag niet onvermeld blijven. Een voorbeeld is zijn grafische weergave uit 1862 van de veldtocht van Napoleon naar Rusland. De grafiek is weergegeven in figuur 2.8. Hij wordt ook besproken door Tufte (1983) en Wainer (1997).

De grafiek beschrijft de lotgevallen van het Franse leger vanaf het moment dat het de Pools-Russische grens passeert. De omvang van het leger wordt aangegeven door de dikte van de grijze band. In het begin bestaat het leger uit 422.000 man. Als het in september Moskou bereikt, is het gereduceerd tot 100.000 man. De terugtocht van het leger wordt aangegeven met de zwarte band. Onderaan de grafiek wordt ook de temperatuur tijdens de terugtocht weergegeven.

De oevers van de rivier Berezina lagen bezaaid met de lichamen van 22.000 dode soldaten nadat het leger in november bij een temperatuur van -20 graden de oversteek maakte. Toen het leger tegen het einde van het jaar de Pools-Russische grens weer bereikte, waren er nog slechts 10.000 man over van de oorspronkelijk 422.000 man.

De grafiek van Minard wordt door vele deskundigen bestempeld als één van de beste statistische grafieken die ooit is gemaakt. Dat heeft niet alleen te maken met de duidelijkheid waarmee dit tragische verhaal wordt verteld, maar ook met de rijkdom aan informatie die in één plaatje aanwezig is. Er zijn zes variabelen uitgezet: tijd, geografische positie (X - en Y-coördinaat), omvang van het leger, richting van het leger en temperatuur. Je kunt de waarden en onderlinge samenhang van de variabelen vrij eenvoudig aflezen uit de grafiek.

Figuur 2.8. De veldtocht van Napoleon



Xbre = December 9bre = November 8bre = October

Een laatste hoogtepunt uit de geschiedenis van de grafieken is de opkomst van de *beeldstatistiek*. Grondlegger daarvan was Otto Neurath (1882-1945), een Oostenrijkse filosoof, econoom en socioloog. Als lid van de Wiener Kreis spande hij zich in voor een betere wereld. Hij vond dat iedereen, en vooral de gewone man, moest kunnen begrijpen hoe de maatschappij in elkaar zat. Maar hij zag ook in dat zulke informatie voor veel mensen saai is en moeilijk te begrijpen. Daarom ontwikkelde hij een systeem van pictogrammen als bouwstenen voor grafieken. Hij noemde dat systeem ISOTYPE (International System Of TYpographic Picture Education).

Neurath had iemand nodig die zijn ideeën kon omzetten in heldere, eenvoudige beelden. Dat werd de Duitse, sociaal betrokken kunstenaar Gerd Arntz (1900-1988). Neurath vluchtte in 1934 naar Den Haag. Later kwam ook Arntz daar naar toe. Deze grafisch vormgever ontwierp meer dan 4000 symbolen voor gebruik in statistieken en optimaliseerde samen met Otto Neurath de ISOTYPE-methode voor het CBS.

Figuur 2.8. Een beeldgrafiek

Figuur 2.8 bevat een voorbeeld van een beeldstatistiek. Deze grafiek stond in 1941 op de kaft van het statistisch zakboek van het CBS. De grafiek laat de samenstelling van de beroeps-bevolking zien in 1899 en 1938. In dit geval is een poppetje als pictogram gebruikt. Elk poppetje stelt 200.000 personen voor.

3. Ingrediënten van een grafiek

In dit hoofdstuk bespreken we de verschillende ingrediënten die nodig zijn voor het maken van een grafiek. De grafiek die zo ontstaat, moet voldoen aan de in de inleiding genoemde eisen voor een goede grafiek. Aan de orde komen:

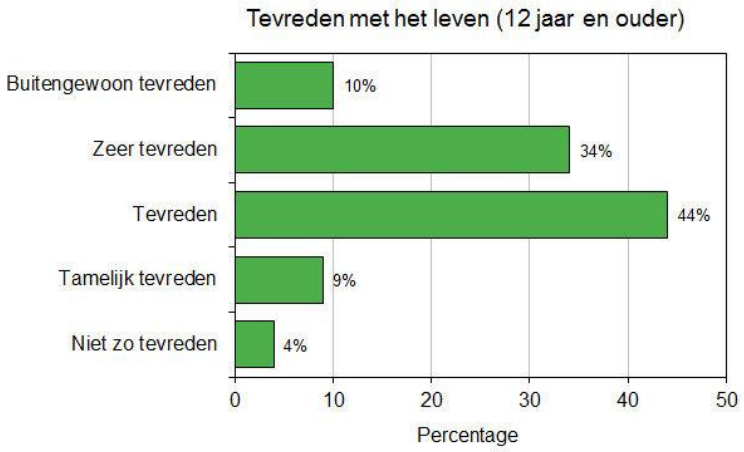
- De *data* (de echte gegevens) en de *metadata* (de informatie die nodig is om de grafiek te begrijpen).
- De noodzakelijke bouwstenen van een grafiek, zoals titel, de assen en de legenda.
- Het gebruik van kleur, grijswaarden en arcering.
- De keuze van het type grafiek, afhankelijk van waar je de grafiek voor wilt gebruiken en het soort gegevens dat je wilt weergeven.

Data en metadata

Een grafiek bestaat uit een grafische weergave van data. De data zijn de gegevens die je in een onderzoek hebt verzameld. Om uit te leggen wat er is te zien in je grafiek gebruik je metadata. We illustreren dat aan de hand van het voorbeeld in figuur 3.1. De *data* zijn de gegevens die je in de grafiek laat zien. In het voorbeeld zijn dat de staven van het staafdiagram. Die zijn de grafische vertaling van de percentages mensen in de verschillende categorieën.

De *metadata* is nodig om uit te leggen wat er precies te zien is in de grafiek en die helpt dus bij het begrijpen en interpreteren van de grafiek. De metadata in figuur 3.1 vertelt ons welke variabele is weergegeven (tevredenheid met het leven), benoemt de categorieën (buitengewoon tevreden, zeer tevreden, enz.) en legt ook uit wat de horizontale as voorstelt (percentage).

Figuur 3.1 Data en metadata van een grafiek

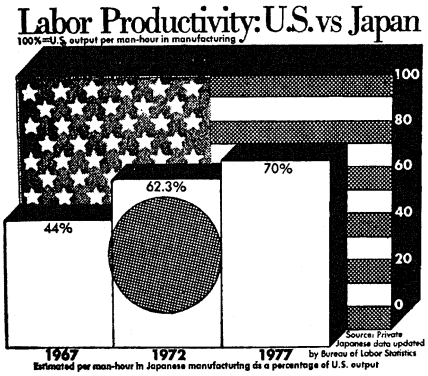


Heel globaal gaat het bij de data dus om zaken die binnen het kader (de rechthoek) zijn afgebeeld, en de metadata zijn alle zaken die er omheen te zien zijn. Figuur 3.1 bevat een uitzondering op dit principe: de verticale grijze hulplijnen behoren tot de metadata, maar ze bevinden zich wel binnen de rechthoek.

Soms bevatten grafieken nog meer dan alleen data en metadata. Het gaat dan om allerlei extra versiering. Tufte (1983) duidt deze toeters en bellen aan met *chartjunk*. Die tierlantijnen dragen er meestal niet toe bij dat de boodschap in de grafiek beter overkomt. In tegendeel, het kan die boodschap juist verhullen. Daarom moet je chartjunk altijd vermijden in je grafieken.

De linker grafiek in figuur 3.2 bevat een voorbeeld van een grafiek met chartjunk. De grafiek brengt de relatieve arbeidsproductiviteit in kaart van de Verenigde Staten ten opzichte van Japan. De grafiek bevat slechts drie getallen en veel chartjunk. In de rechter grafiek is de chartjunk verwijdert. Er zijn slechts data en metadata overgebleven. De grafiek is nu beter leesbaar.

Figuur 3.2. Arbeidsproductiviteit van de VS ten opzichte van Japan



Een manier om de hoeveelheid chartjunk meetbaar te maken is de door Tufte voorgestelde *data/ink ratio*. De waarde van deze index is gelijk aan de hoeveelheid 'inkt' die in de grafiek is verwerkt voor het weergeven van waarden (data en metadata), gedeeld door de in totaal gebruikte hoeveelheid inkt. In een goede grafiek zou deze data/ink ratio in de buurt van de 1 moeten liggen. Het zal duidelijk zijn dat voor de linker grafiek in figuur 3.2 de data/ink ratio veel groter is dan 1, terwijl hij dicht bij 1 zal liggen voor de rechter grafiek.

De bouwstenen van een grafiek

Voor het weergeven van data en metadata in een moet je een aantal bouwstenen gebruiken. Ze zijn in vrijwel elke grafiek zijn terug te vinden:

- De titel van de grafiek.
- Het data-gebied. Dit is het gebied (meestal rechthoek) waarin je de gegevens toont.
- De assen (horizontaal en/of verticaal) met daarop aangegeven het waardebereik en/of de namen van de categorieën. Maatstreepjes helpen bij het aflezen daarvan. Ook roosterlijnen kunnen helpen.
- De legenda

Bij de meeste grafieken geef je de data weer in een rechthoekig gebied in het midden van de grafiek. Welke vorm moet dit data-gebied krijgen? Staand (hoger-dan-breed, 'portrait') of liggend (breder-dan-hoog, landscape)? Tufte (1983) wijst erop dat het menselijk oog goed kan omgaan met het waarnemen van afwijkingen ten opzichte van de horizon. Daarom moet een grafiek lijken op een landschap met een horizon. De grafiek moet een soort panorama zijn. Dit betekent dat hij dus een liggende rechthoek moet worden. Een praktisch argument is bovendien dat het plaatsen van teksten (die wel breed, maar niet hoog zijn) beter tot zijn recht komt in een liggende rechthoek.

Tufte (1983) heeft een groot aantal erkend goede grafieken (89) van John Playfair uit de 17de eeuw bekeken. Dat waren bijna allemaal tijdreeksen. De breedte-hoogte verhouding van deze grafieken lag steeds tussen 1,4 en 1,8. Het waren dus liggende grafieken.

Er lijkt een voorkeur te bestaan voor liggende grafieken. Maar welke verhouding moet nu worden gehanteerd? Een aardig suggestie is de uit de vijfde eeuw voor Christus daterende *Gulden Snede*. Deze komt overeen met een breedte-hoogte verhouding van 1,618. Vooral in de architectuur is deze verhouding terug te vinden (van het Parthenon in Athene tot Le Corbusier). Dit is kennelijk een prettig ogende verhouding.

Het data-gebied is dus de rechthoek die alle data bevat. Meestal krijgt deze rechthoek een zichtbare rand. Daarmee is het voor de lezer duidelijk dat alles binnen deze rand tot de data behoort.

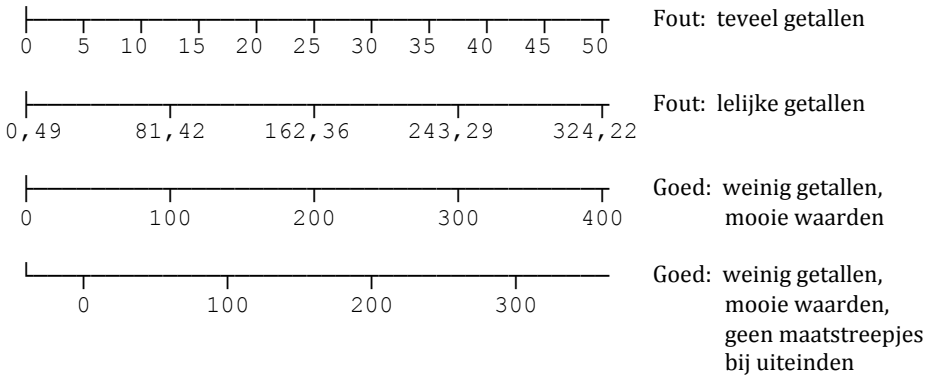
De grafiek moet een geschikte titel krijgen. Die titel moet voldoende zijn om de grafiek te kunnen begrijpen. Het mag niet zo zijn dat de lezer in de tekst moet gaan zoeken naar uitleg van de grafiek. Het ligt voor de hand om de titel boven de grafiek te zetten.

Als je in de grafiek het waardebereik van de gegevens wilt weergegeven, dan moeten de waarden goed af te lezen zijn op de as. De tekst bij de as moet duidelijk beschrijven wat die waarden precies voorstellen, inclusief de meeteenheid. Voor het correct aflezen van de juiste waarde is het belangrijk om *maatstreepjes* (*tick marks*) aan de assen toe te voegen. Daarbij moet je bedenken dat te weinig maatstreepjes niet echt helpen, maar een teveel aan maatstreepjes kan ook verwarrend werken. De verticale as van rechter grafiek in figuur 3.2 bevat 6 maatstreepjes.

Het is verstandig om even na te denken over de waarden die horen bij de uiteinden van de as. Bij welke waarde moet de as beginnen en bij welke waarde moet hij eindigen? Soms gebruikt een computerprogramma voor het maken van grafieken hiervoor de kleinste en de grootste voorkomende waarde. Dat hoeven geen mooie getallen te zijn. Dat leidt er dan toe dat ook de waarden bij de tussenliggende maatstreepjes 'lelijk' zijn. Dat bevordert niet de leesbaarheid van de grafiek. Zie het tweede voorbeeld in figuur 3.3. Er zijn twee oplossingen denkbaar:

- Handmatig een mooie beginwaarde en eindwaarde kiezen. De voorkomende waarden moeten dan wel tussen die begin- en eindwaarde liggen. Zie het derde voorbeeld in figuur 3.3.
- De maatstreepjes niet bij de uiteinden van de as zetten, maar op de positie van de mooie waarden. Zie het vierde voorbeeld in figuur 3.3

Figuur 3.3. Voorbeelden van maatstreepjes en waarden



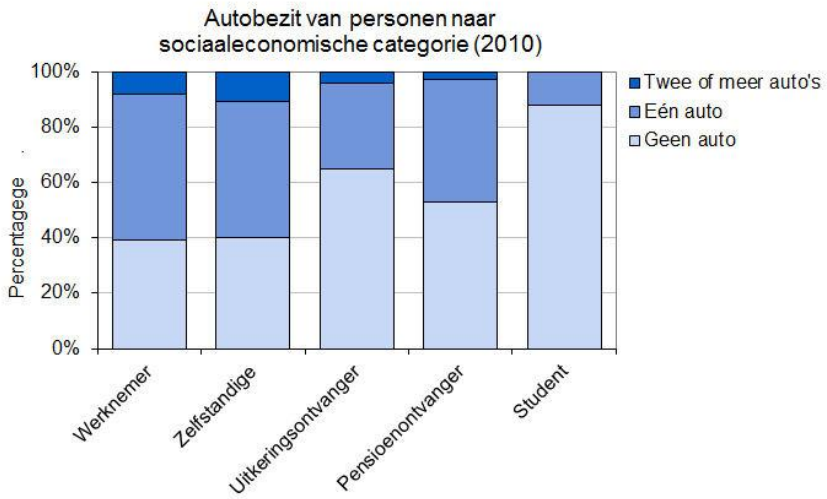
Meer over deze problematiek is te vinden in een artikel van Talbot, Lin & Hanrahan (2010).

Wat kan helpen bij het aflezen van waarden is het toevoegen van *roosterlijnen* (*grid lines*). Ook hier geldt dat je moet nadenken over het juiste aantal lijnen (niet te veel en niet te weinig). Bovendien moeten ze in de grafiek niet gaan overheersen. Daarom is het beter om ze in een niet

te opvallende kleur weer te geven. Als een voorbeeld daarvan bevat de rechter grafiek in figuur 3.2 grijze roosterlijnen.

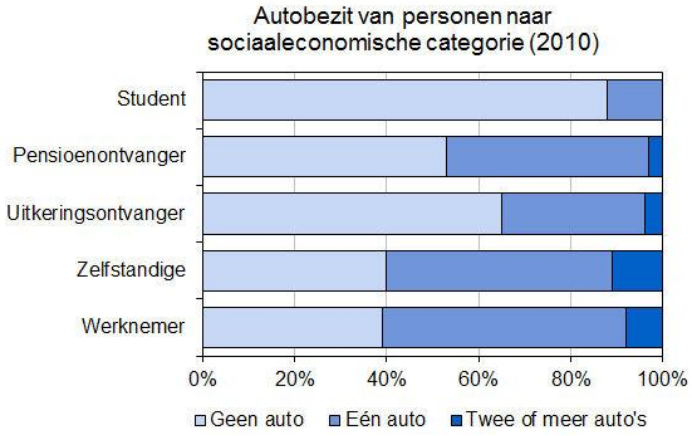
Om de leesbaarheid van de metadata optimaal te houden, is het aan te bevelen alle teksten horizontaal te zetten, en niet verticaal of schuin. Figuur 3.4 geeft een voorbeeld van hoe het niet zou moeten.

Figuur 3.4. Een grafiek met gedraaide teksten



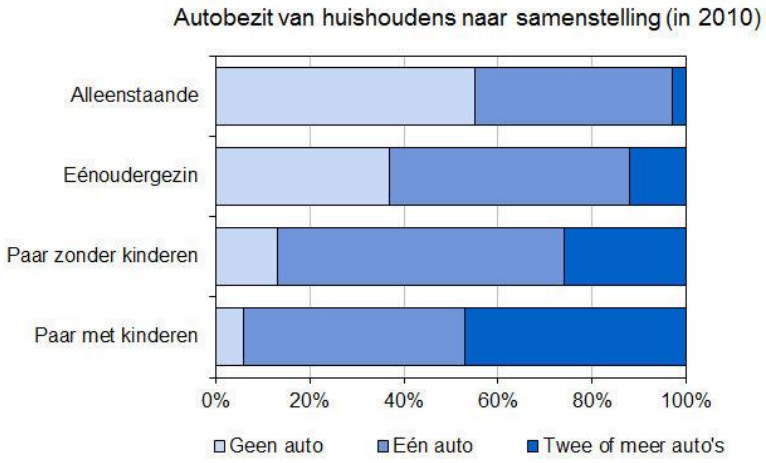
Figuur 3.5 geeft een verbeterde versie van figuur 3.4. De balken zijn nu horizontaal getekend, waardoor er ruimte komt om er horizontale teksten bij te zetten. De legenda is verplaatst van de zijkant naar de onderkant. Daarmee lopen de kleuren in de legenda weer synchroon met de kleuren in de balken.

Figuur 3.5. Een grafiek met horizontale teksten



Of het nu gaat om de staven in een staafdiagram, de lijnen in een lijndiagram of de sectoren in een cirkeldiagram, het moet altijd duidelijk zijn wat die verschillende onderdelen precies voorstellen. De kun je soms doen door er een verklarende tekst bij te zetten (zoals in figuur 3.1), maar als dat niet kan, dan moet de grafiek een legenda bevatten. Figuur 3.6 bevat een voorbeeld.

Figuur 3.6. Grafiek met een legenda



De legenda maakt deel uit van de metadata van de grafiek. De categorieën in de legenda moeten dezelfde volgorde hebben als in de grafiek. Ook de oriëntatie moet hetzelfde zijn. Daarom staan de categorieën in figuur 3.5 naast elkaar en in figuur 3.4 onder elkaar.

Verschillende typen grafieken

Er zijn heel wat verschillende typen grafieken mogelijk, maar niet elke type grafiek is geschikt voor elke situatie. De keuze van het type grafiek hangt in een concrete situatie af van het type variabele en welk aspect ervan je wilt laten zien. We onderscheiden hier twee typen variabelen: kwantitatieve en kwalitatieve variabelen.

Kwantitatieve variabelen meten meestal een omvang, waarde, hoeveelheid of duur. De waarden van dit type variabelen hebben een betekenis. Een grotere waarde betekent groter, meer of duurder en een kleine waarde betekent kleiner, minder of goedkoper. Je kunt rekenen met die waarden. Grootheden zoals het gemiddelde en het totaal hebben betekenis. Voorbeelden van kwantitatieve variabelen zijn gewicht, inkomen, lengte en leeftijd.

Als een kwantitatieve variabele een natuurlijk nulpunt heeft, dan noemen we hem ook wel een *ratio-variabele*. Een voorbeeld is de lengte van een persoon. De waarde 0 betekent dat er helemaal niets meer is. En een tweemaal zo grote waarde betekent ook tweemaal zoveel. Heeft een kwantitatieve variabele geen natuurlijk nulpunt, dan noemen we dat een *interval-variabele*. Een voorbeeld is temperatuur (in graden Celsius). De waarde 0 betekent niet dat er niets meer is. De temperatuur kan zelfs nog verder zakken. En 20 graden is niet twee keer zo warm als 10 graden.

Kwalitatieve variabelen doen niet anders dan de waarnemingen in een aantal groepen (categorieën) verdelen. De waarde is niet meer dan een etiket voor de groep waarin het desbetreffende object (persoon, huishouden, bedrijf) thuishoort. Voorbeelden van kwalitatieve variabelen zijn geslacht (man, vrouw), burgerlijke staat (ongehuwd, gehuwd, gescheiden, weduwstaat), en geloofsovertuiging (Protestants, Rooms-katholiek, Islamitisch, Joods, …). Je kunt niet met de waarden van een kwalitatieve variabele rekenen. Er bestaat niet zoiets als het gemiddelde geslacht, de gemiddelde burgerlijke staat of de gemiddelde geloofsovertuiging.

Soms is het mogelijk een natuurlijke volgorde aan te brengen in de categorieën. Een dergelijke kwalitatieve variabele noemen we een *ordinale variabele*. Een voorbeeld is de variabele opleidingsniveau. Is er geen sprake van een natuurlijke volgorde, dan is dat een *nominale variabele*. Geloofsovertuiging is een voorbeeld van een nominale variabele.

Het komt regelmatig voor dat een kwantitatieve variabele wordt omgezet in een kwalitatieve variabele. Dat kan door het waardebereik te verdelen in een aantal intervallen. Dit proces wordt ook wel *categoriseren* genoemd. Zo kun je, bijvoorbeeld, van de kwantitatieve variabele leeftijd de kwalitatieve variabele leeftijdsklasse maken. Dit indelen in klassen leidt altijd tot een ordinale variabele. Het is een manier om een grote hoeveelheid data te reduceren en samen te vatten in een simpele grafiek.

Een manier om wat orde te brengen in de veelheid van grafische technieken, is het maken van onderscheid in wat die grafieken proberen te doen. Hier maken we onderscheid in twee functies: het in beeld brengen van de verdeling en het beeld brengen van de samenhang.

Bij het in beeld brengen van de *verdeling* gaat het om het gedrag van één variabele. De grafiek moet dan duidelijk maken welke waarden voorkomen, welke waarden veel of weinig voorkomen, of er sterk afwijkende waarden zijn, enz.

Bij het in beeld brengen van de *samenhang* gaat het om de samenhang tussen twee variabelen. Vallen waarden van de ene variabele samen met de waarden van een andere variabele? Is het misschien mogelijk de ene variabele uit de andere te voorspellen?

In tabel 3.1 zijn een aantal typen grafieken opgesplitst naar enerzijds het type variabele (kwantitatief of kwalitatief) en anderzijds naar functie (tonen van verdeling of samenhang). We zullen al deze grafieken behandelen. Achter elk type grafiek staat het nummer van het hoofdstuk.

Tabel 3.1. Typen grafieken

Type variabele	Functie	
	Verdeling	Samenhang
Kwalitatief	Cirkeldiagram (5) Staafdiagram (5)	Samengesteld staafdiagram (7) Stapeldiagram (7) Mozaïekdiagram (7) Cirkeldiagrammen (7)
Kwantitatief	Spreidingsdiagram (6) Snorrendoos (6) Histogram (6)	Puntenwolk (8) Tijdreeks (10) Thematische kaart (11)
Gemengd		Spreidingsdiagrammen (9) Snorrendozen (9)

Merk op dat de tijdreeks een speciaal geval is. Daarbij is de ene variabele de in beeld te brengen variabele en de andere variabele is de tijd. Ook de thematische kaart is een speciaal geval. Daarbij is de andere variabele de geografische positie.

Het gebruik van kleur

Je kunt kleur op verschillende manieren gebruiken in een grafiek. Als kleur alleen maar bedoeld is om de grafiek wat op te vrolijken en hij heeft geen functie bij de interpretatie, dan is er sprake van chartjunk. Dit gebruik van kleur moet je vermijden.

Kleur kan echter ook een functioneel aspect van een grafiek zijn. Daarbij gebruik je kleur om de statistische boodschap in de data beter over te brengen of de structuur in de gegevens beter weer te geven. Bijvoorbeeld, Zeileis, Hornik & Murrell (2009) bespreken een aantal technieken voor het perceptueel juist gebruik van kleur.

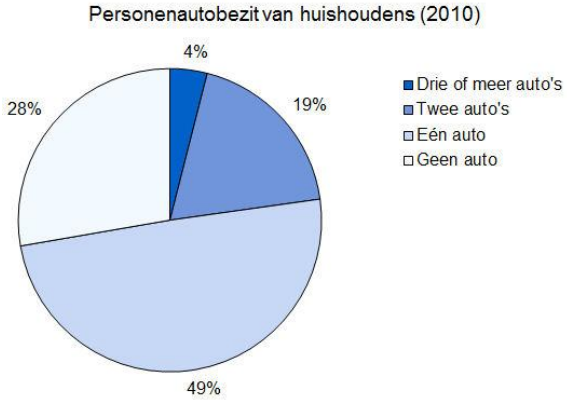
Zoals al eerder is aangegeven, zijn er twee soorten variabelen: kwalitatieve variabelen en kwantitatieve variabelen. Elk type variabele heeft zijn eigen gebruik van kleur.

Wanneer je de verschillende waarden van een kwantitatieve variabelen in een grafiek wilt weergegeven, dan kun je daarbij heel goed gebruik maken van één kleur, waarbij verschillende intensiteiten van dezelfde kleur de verschillende waarden aangeven. Je zou opeenvolgende waarden bijvoorbeeld kunnen weergeven in een steeds donker wordende tint. Figuur 3.7 bevat een voorbeeld waarin de kleur steeds lichter is naarmate de waarde hoger is. We noemen dit een *sequentieel palet*. Je kunt er natuurlijk ook voor kiezen om steeds donkerder wordende tinten te nemen. Het ligt voor de hand je hierbij te laten leiden door wat de variabele meet.

Figuur 3.7. Tinten van dezelfde kleur voor een kwantitatieve variabele

Als een kwantitatieve variabele elke mogelijke waarde in een bepaald interval kan aannemen (een continue variabele), dan zal ook elke kleurtint mogelijk moeten zijn. In de praktijk is dit niet te doen. Ook zal het lastig zijn om verschillende tinten te onderscheiden. Het is dan beter om de kwantitatieve variabele omzetten in een (ordinale) kwalitatieve variabele. Dan kun je volstaan met een beperkt aantal kleurtinten. Figuur 3.8 bevat daarvan een voorbeeld. Er zijn vier categorieën gemaakt voor het aantal auto's dat een huishouden kan bezitten: geen auto, één auto, twee auto's en drie of meer auto's. Naarmate het aantal auto's in een huishouden toeneemt, wordt de tint donkerder.

Figuur 3.8. Eén kleur voor een kwantitatieve variabele



In principe zou je een grafiek als in figuur 3.8 ook wel kunnen maken met verschillende grijstinten. Bedenk dan dat door het gebruik van één echte kleur de data zich beter onderscheiden van de metadata in de grafiek (assen, maatstreepjes, waarden, as-titels, legenda, enz.). Bovendien trekt een grafiek als in figuur 3.8 wat meer de aandacht.

Soms wil je een kwantitatieve variabele gebruiken om een saldo van twee variabelen aan te geven. Het is dan belangrijk om snel te kunnen zien of dat saldo positief of negatief is. Denk bijvoorbeeld aan het saldo van inkomsten en uitgaven, het saldo van instroom en uitstroom van inwoners van een gemeente en winst of verlies bij de verkiezingen.

In het geval van saldo's zijn twee kleuren nodig: een kleur voor de positieve waarden en een contrasterende kleur voor negatieve waarden. Figuur 3.9 geeft een voorbeeld van een kleurenpalet dat je voor een saldovariabele zou kunnen gebruiken. Het ligt voor de hand om

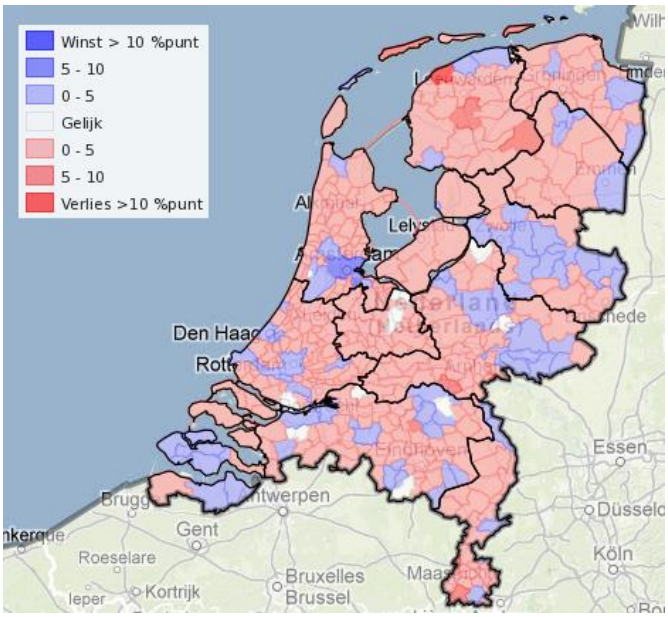
blauw te gebruiken voor positieve saldo's en rood voor negatieve saldo's. Dit noemen we een *divergent palet.*

Figuur 3.9. Kleuren voor het weergeven van saldo's

Het gebruik van rood en groen voor een saldovariabele wordt afgeraden. Veel mensen zien het verschil tussen deze twee kleuren niet goed.

Figuur 3.10 bevat een voorbeeld van het weergeven van de positieve en negatieve waarden. Het gaat hier om winst en verlies van de PvdA bij de verkiezingen voor de Provinciale Staten in 2011. In de legenda kun je zien dat er drie categorieën zijn voor winst, drie voor verlies en nog een neutrale categorie (geen verandering). Dit voorbeeld maakt nog eens duidelijk hoe krachtig een goede grafiek kan zijn. In één oogopslag is te zien waar in Nederland de PvdA heeft gewonnen (Amsterdam) en waar deze partij ernstig heeft verloren (Noord-Friesland).

Figuur 3.10. Winst en verlies van de PvdA bij de verkiezingen voor de Provinciale Staten in 2011



Bij een kwalitatieve variabele is het gebruik van kleur anders. De waarden van een dergelijke variabele hebben geen intrinsieke betekenis. Ze zijn alleen maar bedoeld om groepen van elkaar te kunnen onderscheiden. Het zijn slechts etiketten. Voorbeelden van kwalitatieve variabelen zijn geslacht, burgerlijke staat, geloofsovertuiging en de partij waarop men heeft gestemd.

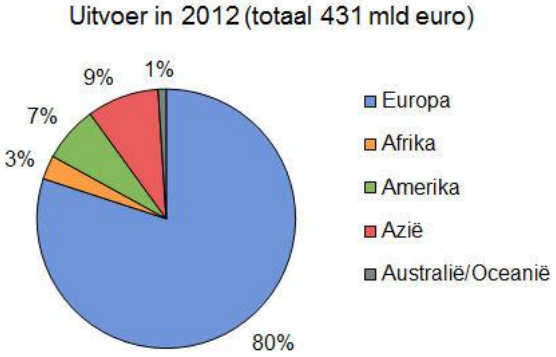
De verschillende categorieën van een kwalitatieve variabele hebben meestal geen natuurlijk volgorde. Daarom kun je de waarden van een dergelijke variabele ook niet in een oplopende of aflopende tint van dezelfde kleur weergeven (aangezien dat een ordening zou suggereren). Je moet verschillende kleuren gebruiken om verschillende categorieën aan te duiden. De perceptie van de intensiteit van die kleuren moet voor elk categorie hetzelfde zijn. Figuur 3.11 toont een voorbeeld van een dergelijk palet.

Figuur 3.11. Kleuren voor het weergeven van een kwalitatieve variabele

In figuur 3.12 is dit palet toegepast op een kwalitatieve variabele die de uitvoer van Nederland naar de verschillende werelddelen beschrijft. Hiervoor is gebruik gemaakt van een cirkeldiagram. De verschillende continenten hebben geen natuurlijke volgorde en daarom zijn ze aangeduid met verschillende kleuren.

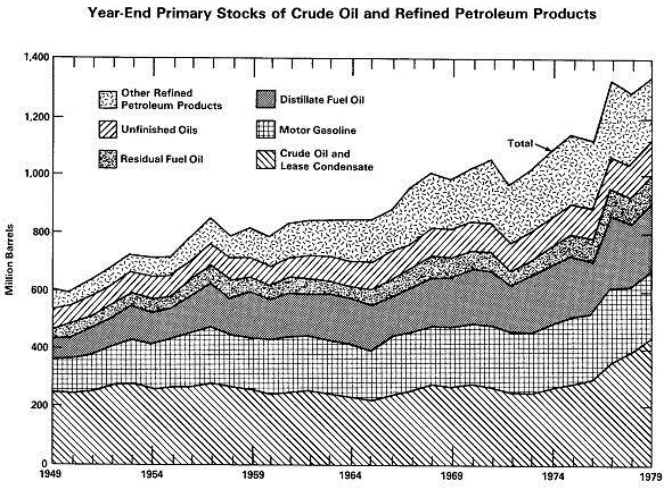
Op de website *colorbrewer2.org* van Cynthia Brewer is het mogelijk om diverse kleurenpaletten voor kwantitatieve variabelen, saldi en kwalitatieve variabelen uit te proberen. Deze paletten zijn uitgetest op proefpersonen. Sommige van die paletten zijn ook bruikbaar voor kleurenblinden.

Figuur 3.12. Kleur voor een kwalitatieve variabele



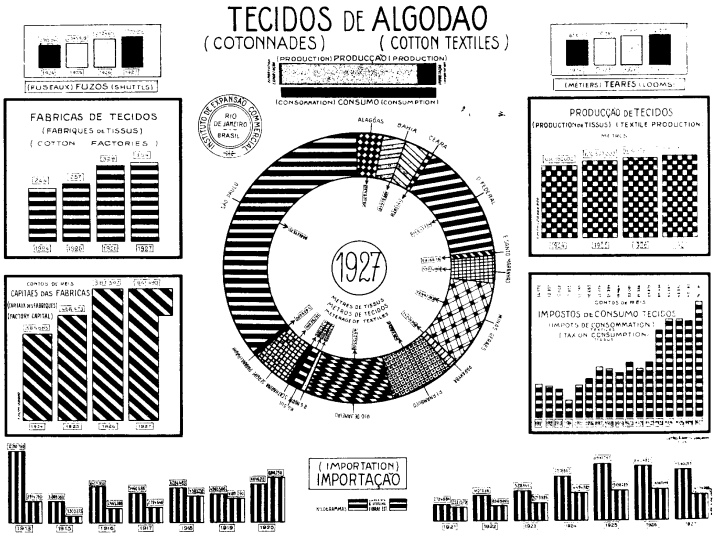
De vraag zou kunnen opkomen of je verschillende categorieën ook niet kunt weergeven met verschillende soorten arceringen in zwart/wit. Theoretische zou dat misschien wel kunnen, maar in de praktijk is het toch vaak lastig om dat goed te doen. Je moet immers voorkomen dat de verschillende arceringen bij de lezers de perceptie van intensiteitsverschillen oproepen. De uitdaging is om arceringen te gebruiken die geen intensiteitsverschillen hebben en die toch duidelijk van elkaar te onderscheiden zijn.

Figuur 3.13. Het gebruik van arcering voor een kwalitatieve variabele



Arceringen kunnen leiden tot een onrustig beeld. Soms lijkt het wel of delen van een grafiek gaan vibreren. Er kunnen zelfs moiré-effecten optreden. Die interferentiepatronen kunnen ontstaan als je verschillende lijnenpatronen over elkaar heen legt. De schijnbare beweging in de grafiek leidt af van de boodschap die de grafiek probeert over te brengen. Figuur 3.14 bevat een voorbeeld van een grafiek met zulke hinderlijke effecten.

Figuur 3.14. Hinderlijk gebruik van arcering

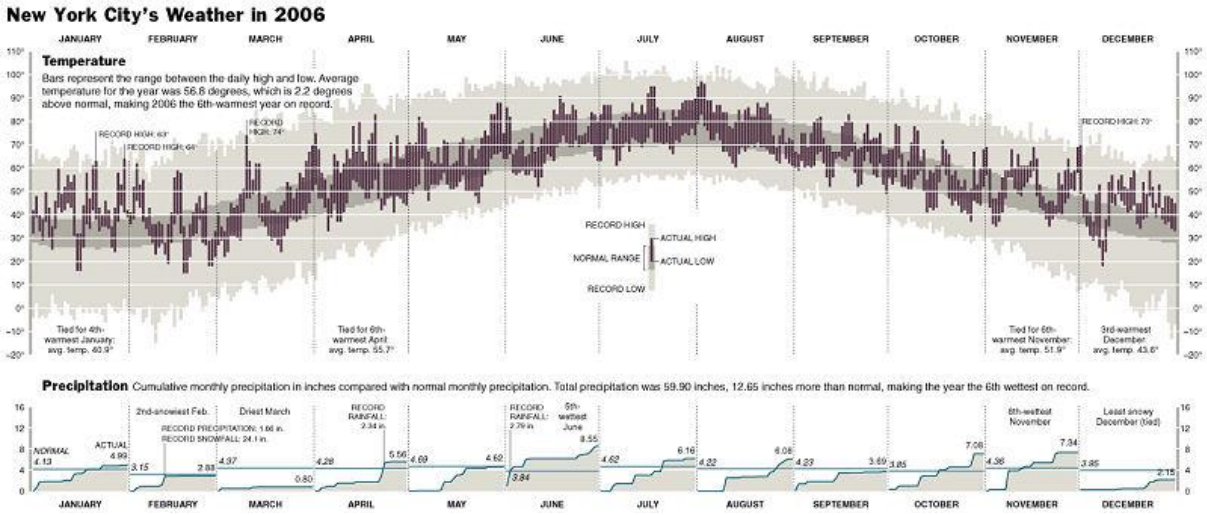


4. Richtlijnen voor grafieken

Grafieken zijn vooral een waardevol instrument voor het overdragen van informatie die bevat zit in grote hoeveelheden getallen. Een goed verbeeld daarvan is figuur 4.1. Deze grafiek beschrijft het weer in New York in het jaar 2006.

De grafiek bevat gegevens over temperatuur en neerslag. Bovendien kun je de feitelijke temperaturen over 2006 vergelijken met wat je normaal zou kunnen verwachten (de donkergrijze band) en met de laagste en hoogste temperaturen ooit (de lichtgrijze band). De grafiek is in feite gebaseerd op meer dan 2000 getallen. Ondanks de hoge dichtheid aan informatie, is de grafiek goed leesbaar. Het is niet alleen mogelijk om globale trends te observeren (het is in de zomer warmer dan in de winter), maar ook kleine details (3 augustus was de warmste dag van het jaar). In de grafiek is duidelijk te zien wanneer het weer afweek van het normale patroon (in december was het ongewoon warm). En ook is het mogelijk om allerlei vergelijkingen op detailniveau te maken (op 1 december was het warmer dan op 5 april).

Figuur 4.1. Het weer in New York in 2006



Een grafiek kan dus een zeer krachtig communicatiemiddel zijn voor het overbrengen van een statistische boodschap. Voor een groter publiek is een grafiek bovendien vaak aantrekkelijker en leesbaarder dan een tabel vol getallen. Daarom is het niet verwonderlijk dat in veel publicaties gebruik wordt gemaakt van grafieken. Het is echter niet allemaal rozengeur en maneschijn.

Het maken van grafieken is niet zonder risico's. Een (bewuste of onbewuste) verkeerde vormgeving van een grafiek kan er al heel snel toe leiden dat je de verkeerde conclusies trekt. Zeker in de media, waar vormgevers vaak de grafieken maken, en niet de statistici, zie je veel voorbeelden van foute grafieken. Het is daarom van belang om je bij het maken van grafieken aan een aantal elementaire richtlijnen te houden. In de rest van dit hoofdstuk zullen we een aantal richtlijnen bespreken en illustreren aan de hand van voorbeelden. Die richtlijnen zijn grotendeels gebaseerd op het werk van Tufte (1983), Schmid (1983) en Wainer (1997).

Richtlijn 1: Rommel niet met de assen

Om de omvang van allerlei verschijnselen uit een grafiek te kunnen aflezen heb je assen nodig. Die assen horen voorzien te zijn van maatstreepjes en bij die maatstreepjes horen waarden te staan. Bij het maken van een grafiek moet je die assen op een correcte manier tekenen. Doe je

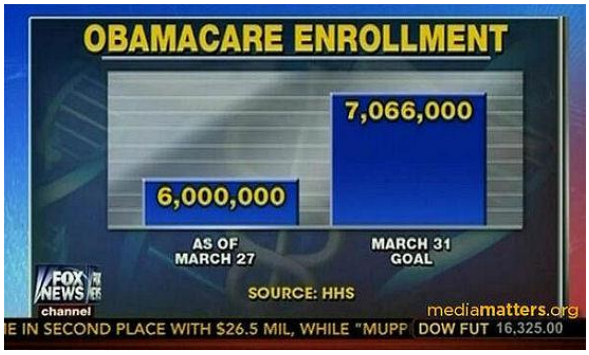
dat niet dan leidt de grafiek al gauw tot verkeerde interpretatie. In de praktijk zie je twee soorten fouten regelmatig terug komen:

- *De as begint niet bij 0*. Als de as een omvang, hoeveelheid, waarde, duur, aantal of percentage aanduidt, dan is er een natuurlijk nulpunt. De as moet dan bij dit nulpunt beginnen. Als dat niet het geval is dan lijken allerlei verschijnselen groter dan ze in werkelijkheid zijn. Er is dus sprake van overdrijving. De grafiek is suggestief.
- *Meer dan één as*. Als je in een grafiek verschillende verschijnselen wilt vergelijken, dan moeten je die verschijnselen allemaal op dezelfde as afzetten. Als elke verschijnsel een eigen as krijgt met een eigen maatstreepjes en eigen waarden, dan is een correcte vergelijking niet meer mogelijk.

Een eerste voorbeeld van gerommel met de as komt van *Fox News*, een van de grote nieuwszenders in de Verenigde Staten. De zender is de stem van de rechtse, conservatieve Amerikanen. In de berichtgeving wordt geen mogelijkheid onbenut gelaten om de regering van president Barack Obama in een kwaad daglicht te zetten.

Het was een van de belangrijke doelen van Obama om het systeem van gezondheids-verzekeringen te hervormen. Daarvoor is een nieuwe wet ingevoerd: de *Patient Protection and Affordable Care Act*. De wet verplicht iedere Amerikaan een zorgverzekering af te sluiten. En de zorgverzekeraars mogen niemand weigeren. In de volksmond wordt deze wet wel *Obamacare* genoemd. De Republikeinen zijn vel tegen de wet. Daarom probeerde Fox News de invoering van Obamacare af te schilderen als een volledige flop. De regering Obama mikte erop dat op 31 maart 2014 zo'n 7 miljoen Amerikanen zich moesten hebben aangemeld voor de nieuwe zorgverzekering. Dus probeerde Fox News op 27 maart 2014 met de linker grafiek in figuur 4.2 aan te geven dat dit doel bij lange na niet was gehaald.

Figuur 4.2. Fox News over Obamacare



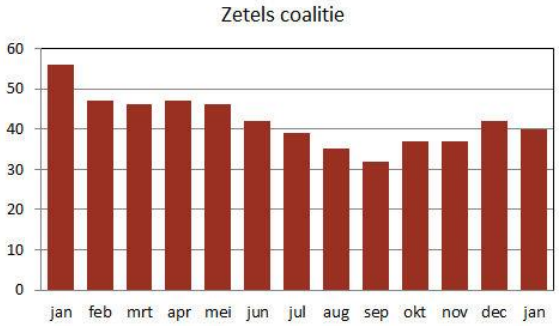
Het verschil tussen de twee staven in het staafdiagram is inderdaad bijzonder groot. Maar een blik op de cijfers leert dat in feite het verschil helemaal niet zo heel groot is: er zijn al 6 miljoen aanmeldingen en dat is 1 miljoen minder dan de doelstelling van ruim 7 miljoen. De staven weerspiegelen dus helemaal niet de werkelijke cijfers. Dat komt omdat de verticale as niet bij 0 begint. Dat is niet eenvoudig vast te stellen omdat die verticale as helemaal geen schaalverdeling heeft. Daarmee is dit een bijzonder slechte grafiek. Als je niet goed oplet, wordt je al snel op het verkeerde been gezet.

Er was veel kritiek op deze grafiek van Fox News. Dat leidde ertoe dat Fox de grafiek een paar dagen later (op 1 april 2014) corrigeerde. De rechter grafiek in figuur 4.2 is de gecorrigeerde grafiek. Er is nu een keurige schaalverdeling bij de verticale as. En die schaalverdeling begint,

zoals het hoort, bij 0. Het verschil tussen beide staven is veel minder groot. De grafiek geeft nu een veel realistischer beeld van de situatie.

Een tweede voorbeeld van gerommel met een as komt uit het tv-programma *EenVandaag*. Dit programma doet regelmatig een politieke peiling. Die peiling heet *De Stemming*. Op 29 januari 2014 kwam presentator Gijs Rademaker met een nieuwe peiling. Tijdens de uitzending was ook de grafiek linksonder in figuur 4.3 te zien. Die grafiek gaf het verloop in de peilingen weer van het aantal zetels van beide coalitiepartijen (VVD en PvdA) samen.

Figuur 4.3. EenVandaag peilt de steun voor de coalitie



Ook deze grafiek heeft een verticale as die niet bij 0 begint. Als je heel goed kijkt, dan blijkt die as te beginnen bij 30 (zetels) in plaats van bij 0. Daardoor lijkt het verloop in de tijd van het aantal zetels veel spectaculairder dan het in werkelijkheid is. Je zou zelfs de indruk kunnen krijgen dat de coalitie in september bijna helemaal geen zetels meer had. Van augustus naar september lijkt het aantal zetels meer dan gehalveerd!

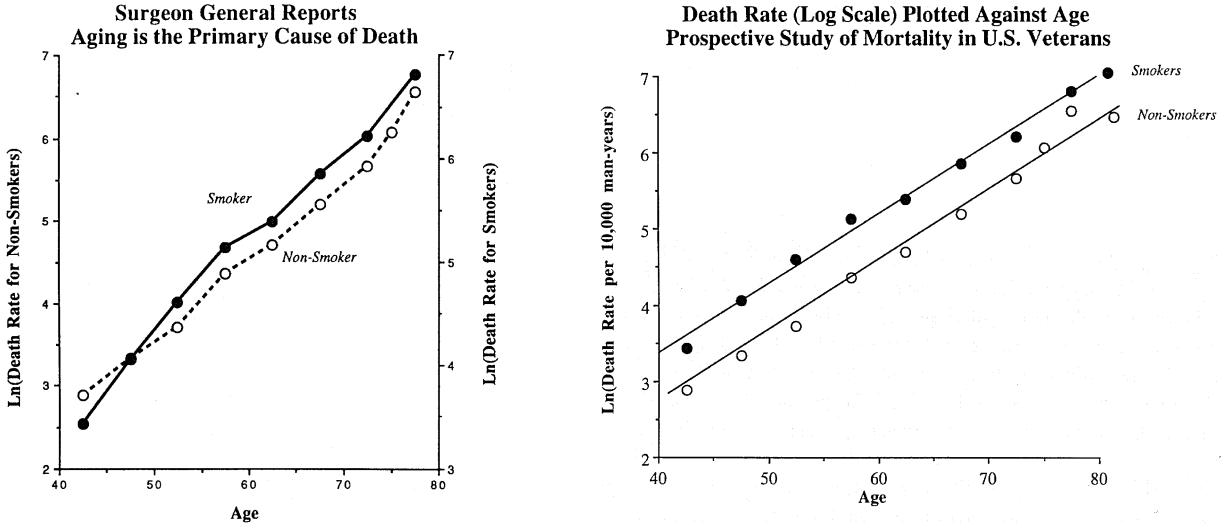
Gelukkig staan de juiste zetelaantallen wel genoemd onder de grafiek. Dan kun je in ieder geval vaststellen dat de grafiek niet klopt. Van augustus naar september liep het aantal zetels terug van 35 naar 32. Dat is een achteruitgang (van 9%), maar zeker geen halvering.

De rechter grafiek in figuur 4.3 is de verbeterde versie van de grafiek. Hij heeft een verticale as die wel bij 0 begint. Het verloop in de tijd is nu wat rustiger. Het is echter wel een getrouwere weergave van de cijfers.

Wainer (1997) beschrijft een voorbeeld van de tweede manier van rommelen met assen: verschillende assen voor verschillende verschijnselen. Het voorbeeld gaat over het in beeld brengen van de uitkomsten van een onderzoek naar de mogelijke relatie tussen sterfte en roken. Figuur 4.4 toont twee varianten van een grafiek om de gegevens weer te geven. In beide gevallen is op de horizontale as de leeftijd afgezet en op de verticale as de logaritme van het aantal sterfgevallen op 10.000 personen per leeftijdsklasse. Een hogere waarde van deze grootheid betekent dus een grotere sterftekans.

De linker grafiek in figuur 4.4 geeft de informatie weer zoals de tabaksindustrie dat vast graag wil. Uit de figuur blijkt dat de sterftekans toeneemt met de leeftijd. Dat is niet verbazing-wekkend. Maar belangrijker is dat er geen verschil lijkt te zijn tussen rokers en niet-rokers. Er zit echter een addertje onder het gras. De grafiek maakt gebruik van twee verschillende verticale assen. De as aan de linkerkant is gebruikt voor het uitzetten van de gegevens van de niet-rokers, en de as aan de rechterkant voor de gegevens van de rokers. Merk op dat de schaalverdeling op beide assen ten opzichte van elkaar is verschoven.

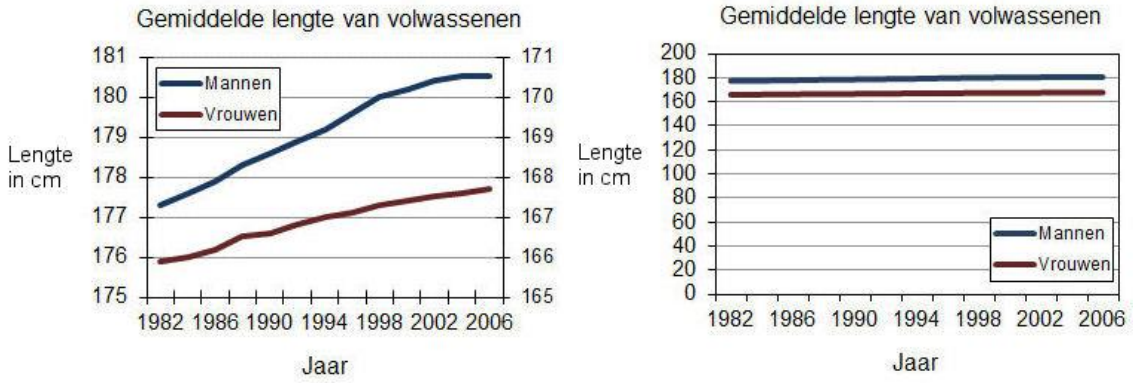
Figuur 4.4. Verband tussen sterfte en leeftijd voor rokers en niet-rokers



Als je voor beide groepen gegevens dezelfde schaalverdeling gebruikt, dan krijg je als resultaat de rechter grafiek in figuur 4.4. Dan wordt ineens duidelijk dat het sterftecijfer binnen een bepaalde leeftijdscategorie voor rokers toch aanzienlijk hoger is dan dat voor niet-rokers. De conclusie moet zijn dat door het gebruik van dubbele assen het toch wel aanzienlijk lastiger is geworden om de juiste conclusie uit de grafiek te trekken.

In een laatste voorbeeld van gerommel met de assen is er zelfs sprake van dubbel gerommel met de assen. De grafiek is afkomstig uit het Webmagazine van het CBS (van 17 januari 2008) en toont de toename in de tijd van de gemiddelde lengte van mannen en vrouwen. Het gaat om de linker grafiek in figuur 4.5.

Figuur 4.5. Dubbel gerommel met de assen



Een eerste blik op de linker grafiek roept de indruk op van een sterke toename van de gemiddelde lengte. Dat geldt vooral voor mannen (blauwe lijn) en in iets mindere mate voor vrouwen (rode lijn). Verder wekt de grafiek de suggestie dat vrouwen gemiddelde maar half zo lang zijn als mannen.

Er zijn twee problemen met de linker grafiek. In de eerste plaats zijn er twee verticale assen. De linker as is voor de mannen en de rechter as moet je gebruiken voor de vrouwen. De twee assen lopen niet synchroon. Daardoor kun je mannen en vrouwen niet met elkaar vergelijken. In de tweede plaats is voor beide assen het natuurlijke nulpunt niet vermeld. De linker verticale as begint bij 175 en de rechter bij 165.

In de rechter grafiek van figuur 4.5 zijn beide defecten gerepareerd. In die grafiek is te zien dat de lengte van mannen en vrouwen nog steeds wel toeneemt, maar veel langzamer. Er zijn ook nog steeds verschillen tussen mannen en vrouwen, maar die zijn veel minder groot. Het beeld is realistischer.

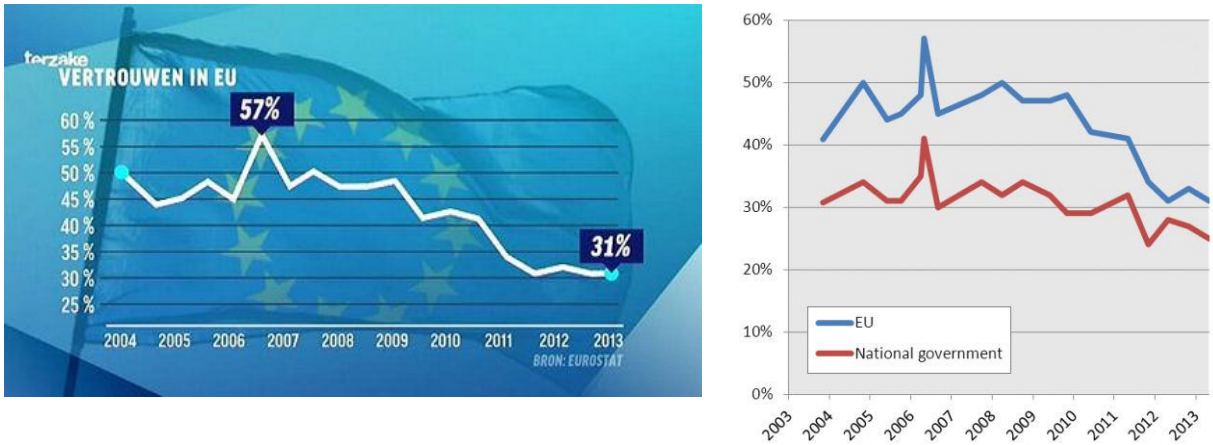
Het zou natuurlijk kunnen dat je vooral de verschillen tussen mannen en vrouwen in meer detail wilt onderzoeken. Dan is het voor de beeldvorming toch beter om eerst te beginnen met het globale plaatje in de rechter grafiek, waarna je eventueel kunt inzoomen via de linker grafiek. Maar dan moet je wel beide assen synchroon laten lopen. Oftewel: beperk je tot één verticale as.

Richtlijn 2: Presenteer de gegevens in de juiste context

Hoe interpreteer je de gegevens in een grafiek op een correcte manier. Dan moet je niet alleen een correcte grafiek tot je beschikking hebben, maar die grafiek moet de gegevens ook in de juiste context weergeven. Als je niets weet over die context, dan kun je het verschijnsel ook niet goed duiden. Je kunt bijvoorbeeld meer context krijgen door de ontwikkeling van het verschijnsel in de loop van de tijd te beschouwen. Of je kunt het verschijnsel vergelijken met andere, verwante verschijnselen.

Een eerste voorbeeld van gebrek aan context komen we tegen in een item van het tv-programma *Terzake* van de Belgische omroep VRT op 23 januari 2014. In het item wilde men duidelijk maken dat de Europeaan zo langzamerhand zijn vertrouwen in de Europese politiek verliest. De linker grafiek in figuur 4.6 werd gebruikt ter onderbouwing hiervan. Het vertrouwen in de EU lijkt inderdaad snel te dalen.

Figuur 4.6. Gebrek aan context



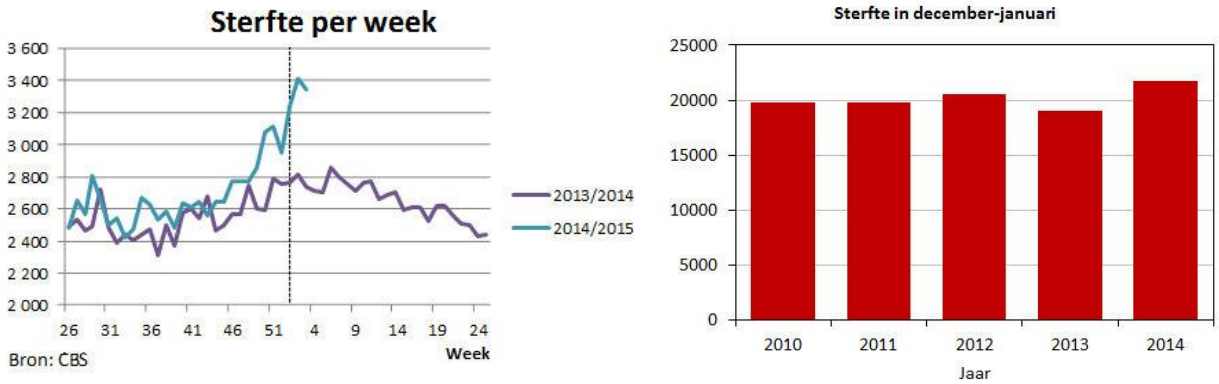
Wat onmiddellijk opvalt is dat de verticale as begin bij 25% in plaats van bij 0%. Door een stuk van de as weg te laten, lijken stijgingen en dalingen altijd spectaculairder dan ze in werkelijkheid zijn.

Eelco Hartveld van de Universiteit van Amsterdam heeft de grafiek opnieuw gemaakt, maar nu wel met een correcte verticale as. Het resultaat is de rechtergrafiek in figuur 4.6. Het vertrouwen in de Europese Unie daalt nog steeds, maar niet meer zo snel. Bovendien is nog een tweede lijn opgenomen in de grafiek en die geeft het vertrouwen weer in de nationale overheid. Zo krijgt de grafiek wat meer context. Twee dingen vallen dan op. In de eerste plaats heeft men kennelijk meer vertrouwen in de Europese Unie dan in de nationale overheid. In de tweede plaats dalen het vertrouwen in Europese Unie en nationale overheid ongeveer op dezelfde manier. Daaruit zou je kunnen afleiden dat de daling van de vertrouwen in de Europese Unie wel eens een

andere oorzaak kan hebben dan de Europese Unie zelf. Kortom, zo dramatisch is het niet met de Europese Unie.

Een tweede voorbeeld van gebrek aan context is een grafiek van het Centraal Bureau voor de Statistiek. Het gaat hier om de linker grafiek in figuur 4.7. De grafiek maakte deel uit van een bericht waarin melding werd gemaakt van een relatief hoge sterfte in de winter van 2014/2015. Inderdaad, als je naar de grafiek kijkt, dan lijkt de sterfte omstreeks de jaarwisseling wel bijna twee keer zo hoog dan in 2013/2014.

Figuur 4.7. Gebrek aan context in een sterftegrafiek



In de grafiek staat het aantal sterfgevallen per week. De blauwe lijn geeft de sterfte aan in de huidige winterperiode (2014/2015). De paarse lijn duidt de sterfte aan in de vorige winterperiode (2013/2014). Als we die twee lijnen met elkaar vergelijken dan is de sterfte in de eerste weken van 2015 inderdaad veel groter dan in de eerste weken van 2014. Maar let: de grafiek zet je op het verkeerde been.

In de eerste plaats begint de verticale as niet bij 0, maar bij 2000. Daardoor worden de verschillen nogal opgeblazen. Het verschil lijkt groter dan het is. Om een eerlijker beeld te geven van de veranderingen, is het beter de as bij 0 te laten beginnen.

In de tweede plaats wordt de sterfte in de huidige winterperiode alleen vergeleken met de sterfte in de vorige winterperiode. Wat zegt dat? Misschien was die sterfte wel atypisch. Misschien is de huidige sterfte niet hoger, maar de vorige sterfte juist lager. Je moet de cijfers dus beter in de context plaatsen.

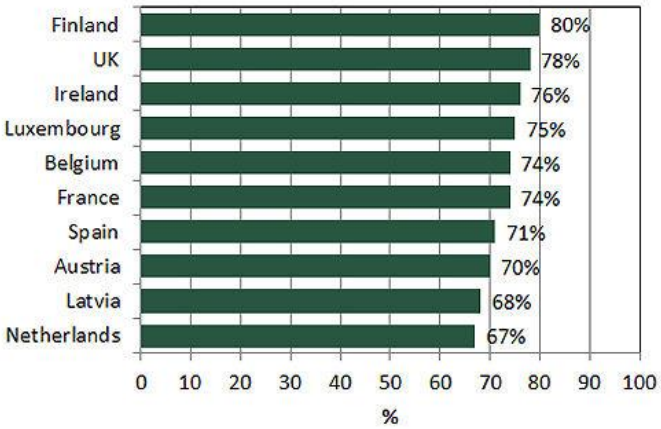
Dezelfde CBS-cijfers zijn gebruikt om een andere grafiek te maken. Dat is de rechter grafiek in figuur 4.7. Voor vijf opeenvolgende jaren is steeds de totale sterfte berekend in de laatste vier weken van dat jaar en de eerste drie weken van het volgende jaar. Daarmee worden week-op-week fluctuaties weggewerkt. Verder begint de verticale as keurig bij 0.

Er ontstaat nu een heel ander beeld. Er is sprake van een lichte stijgende trend in de winter-sterfte. Wat daarvan de oorzaak is? Dat moeten de deskundigen maar uitzoeken. Misschien heeft het wel te maken met de toenemende vergrijzing. Of misschien neemt de bevolking wel toe, waardoor er ook meer sterfte is.

Wat ook opvalt is dat de sterfte in de vorige winter afwijkt van het algemene patroon. De winter in 2013/2014 is lager dan verwacht. Het was dus niet zo'n goed idee om de sterfte in de huidige winter te vergelijken met die van vorig jaar. Al met al is de ontwikkeling in de wintersterfte minder spectaculair dan de grafiek van het CBS suggereert.

Ten slotte nog een derde voorbeeld van een verkeerde context. Op 23 december 2014 maakte een journalist van *The Guardian* een foto van een poster ergens op straat in de Engelse stad Leeds. Op de poster stond een grafiek. Zie de linker grafiek in figuur 4.8. De poster was gemaakt door de Engelse Rail Industry en wil tonen dat de Engelsen toch wel heel tevreden zijn met hun spoorwegen. Ze scoren het hoogst met 78% tevreden reizigers.

Figuur 4.8. Tevredenheid met de spoorwegen



De gegevens voor de grafiek zijn afkomstig uit de Eurobarometer. Dat is een peiling die regelmatig in alle landen van de EU wordt gehouden In 2013 is er gevraagd naar de tevredenheid met allerlei aspecten van het reizen met de trein. Meer cijfers zijn te vinden in het rapport "Europeans' satisfaction with rail services", Flash Eurobarometer 382a.

Het eerste wat opvalt aan de grafiek is dat de lengtes van de staven niet overeenkomen met de percentages. Zo is de staaf voor Duitsland minder dan half zo groot is als de staaf voor Engeland. En de staaf van Italië (39%) is wel erg veel kleiner dan die van Frankrijk (74%). De staaf van Italië zou minsten half zo groot moeten zijn als die van Frankrijk. Het probleem wordt veroorzaakt door een verticale as die niet bij 0 begint. Die verticale as is ook nog eens onzichtbaar, maar na enig gepuzzel blijkt dat die as start bij 30%.

Een ander probleem van de grafiek is dat de gegevens uit hun context zijn gehaald. In de grafiek staan maar vier landen weergegeven. Hoe zou het zijn met de andere landen die niet in de grafiek staan? Als je dan het rapport van de Eurobarometer erbij neemt, dan blijkt dat Engeland helemaal niet het beste land is. Finland scoort nog hoger met 80%. Verder scoren Ierland (76%) en Luxemburg (75%) maar net onder Engeland. Een eerlijker beeld krijg je door deze landen ook in de grafiek op te nemen. De verbeterde versie van de grafiek is de rechter grafiek in figuur 4.8. Er is nu een natuurlijk nulpunt op de as en de scores van de 10 beste landen zijn vermeld. Je kunt nu de score van Engeland beter beoordelen: het land is niet het beste, maar hoort wel tot de landen die het goed doen. Een iets genuanceerder oordeel dus.

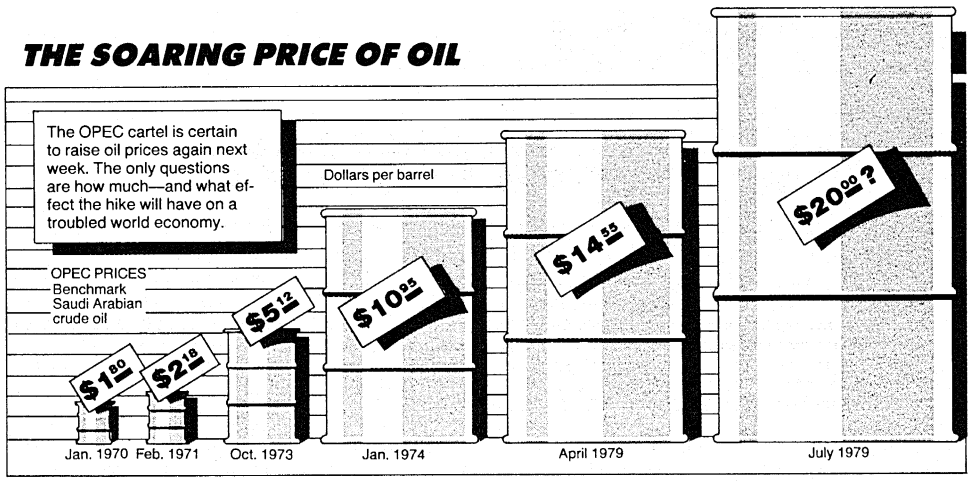
Richtlijn 3: Wees voorzicht met symbolen

Het is de essentie van een grafiek dat gegevens op een visuele manier wordt weergegeven. Er wordt dus een visuele metafoor gemaakt van numerieke gegevens. Die metafoor kan bijvoorbeeld een staaf (in een staafdiagram) zijn of een sector (van een cirkeldiagram). Dat is alleen zinvol als de oorspronkelijke eigenschappen van de gegevens in de metafoor behouden blijven. Dit betekent bijvoorbeeld dat de natuurlijke rangorde van de waarden van klein naar groot gehandhaafd moet blijven. Maar het betekent ook dat de onderlinge verhoudingen gehandhaafd moeten blijven. Is een waarde twee maal zo groot als een andere waarde dan moet dit ook in de visuele metafoor zo worden weergegeven.

Vormgevers van grafieken vinden soms staafdiagrammen of een cirkeldiagrammen wat saai. Om de grafiek 'leuker' te maken vervangen ze dan die staven of diagrammen door symbolen die wat aantrekkelijker zijn, en misschien ook passen bij het onderwerp van de grafiek. Dat kan, maar je moet wel oppassen dat de onderlinge verhoudingen niet worden aangetast, zodat je een verkeerd beeld krijgt van de omvang van de in beeld gebrachte verschijnselen.

Een eerste voorbeeld van dit probleem is afkomstig uit een aflevering van het tijdschrift Newsweek in 1979. Schmid (1983) besprak dit voorbeeld al eerder. Het gaat om de grafiek in figuur 4.9. Die grafiek brengt de stijging van de olieprijs in beeld in de periode van 1970 t/m 1979. Het gaat in feite om een staafdiagram waarvan de staven de vorm van olievaten hebben gekregen. De hoogte van de vaten is evenredig gemaakt met de olieprijzen. De ontwerper heeft echter over het hoofd gezien dat de argeloze gebruiker van de grafiek zal kijken naar de oppervlakte van de vaten en niet naar de hoogte. En de verhoudingen tussen de oppervlaktes komen niet overeen met de verhoudingen tussen de olieprijzen. De oppervlaktes nemen veel sneller toe. Daardoor krijgt de gebruiker onbewust de indruk van een veel snellere stijging van de olieprijzen dan in werkelijkheid het geval was.

Figuur 4.9. De stijging van de olieprijs



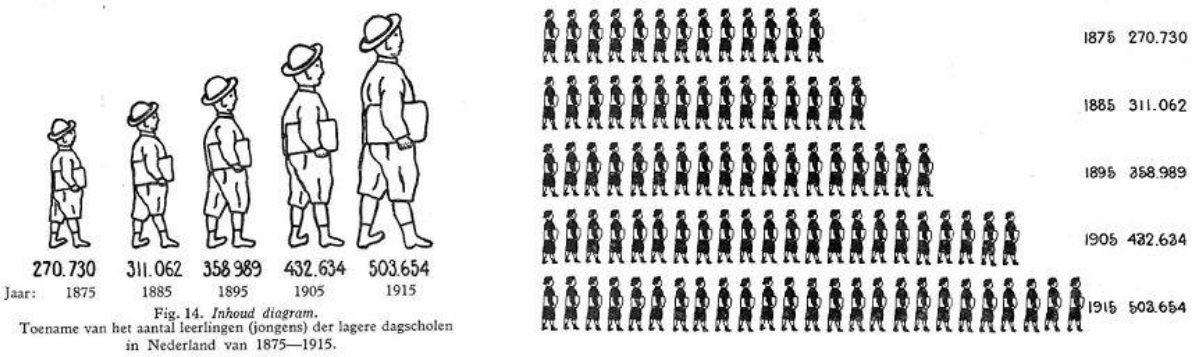
Tufte (1983) heeft voor dit soort grafieken de *lie factor* (Ned: *liegfactor*) ingevoerd. Deze is gedefinieerd als de door de grafiek gesuggereerde waarde gedeeld door de werkelijke waarde. In grafiek 4.7 is de verhouding tussen de olieprijzen in de jaren 1970 en 1979 bijvoorbeeld gelijk aan 20 / 1,80 = 11,11, terwijl de verhouding van de oppervlaktes van de overeenkomstige olievaten ongeveer gelijk is aan 78. De liegfactor heeft hier dus een waarde van 7. In werkelijkheid kan de liegfactor misschien zelfs nog wel groter zijn dan 7. De olievaten

suggereren immers een driedimensionale vorm. Als de argeloze gebruiker die meeweegt in zijn beoordeling van de grafiek, kan de liegfactor zelfs oplopen tot boven de 500.

Merk overigens op dat er nog wel wat meer valt aan te merken op de grafiek. Zo is er duidelijk geen sprake van een gelijkmatige schaalverdeling op de horizontale-as. De afstand tussen de verschillende waarden weerspiegelt niet de tijd die is verlopen tussen de metingen. Tussen de eerste vaten zit 1 jaar verschil, tussen het vierde en vijfde vat jaar zit 5 jaar, en tussen de laatste twee vaten zit maar 3 maanden. Hierdoor krijg je een verkeerde indruk van de ontwikkeling in de tijd.

De problemen met het gebruik van symbolen in plaats van staven zijn al lang bekend. De hoogleraar B.G. Escher (stiefbroer van de bekende graficus M.C. Escher) schrijft in 1924 het boekje 'De methodes de grafische voorstelling'. Daarin staat de linker grafiek in figuur 4.10. Het gaat om de toename van het aantal jongens op de 'lagere dagscholen'. Dat neemt toe, maar niet zo hard als die grafiek suggereert. Het poppetje voor 1915 is 3,3 keer zo groot als het poppetje van 1875, terwijl het aantal leerlingen slechts met een factor 1,9 is toegenomen. Dat is een liegfactor van 1,7. Escher noemt ook nog een ander probleem van de linker grafiek: mensen zouden kunnen denken dat de leerlingen in de loop van de tijd groter zijn geworden.

Figuur 4.10. Het aantal jongens der lagere dagscholen

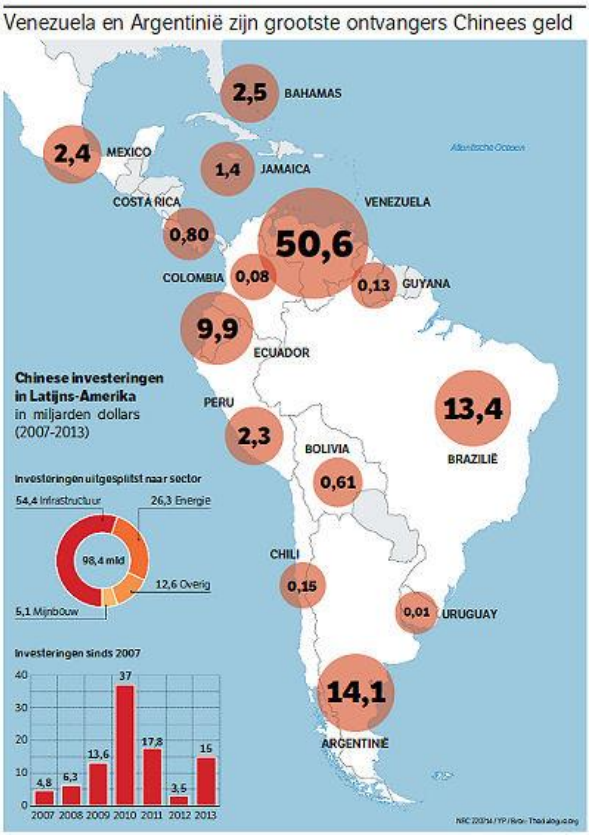


Als je dan toch met symbolen wilt werken, doe het op de manier van de beeldstatistieken (zie hoofdstuk 2). Geef de omvang van het verschijnsel niet aan door het symbool een bepaalde grootte te geven, maar door het symbool een aantal maal te herhalen. Zo kun je bijvoorbeeld de rechter grafiek in figuur 4.10 krijgen.

Symbolen kun je ook gebruiken in andere typen grafieken. Een voorbeeld daarvan is de thematische kaart. Een dergelijke grafiek stond op 22 juli 2014 in het NRC. Hij is overgenomen in figuur 4.11. Het gaat om de Chinese investeringen in Zuid-Amerika. Op de thematische kaart is te zien welke landen allemaal geld van China krijgen en om welke bedragen het gaat. Helaas is die kaart niet goed vorm gegeven, waardoor je op het verkeerde been werd gezet.

Een belangrijk probleem met thematische kaarten is 'area bias'. Die kan zich voordoen als je de waarde van een variabele in een geografisch gebied wilt tonen door het gebied in te kleuren. Gebieden met een groot oppervlak vallen echter meer op dan gebieden met een klein oppervlak, ook al heeft dat gebied een lagere waarde. Dit is vooral een probleem bij het weergeven van aantallen of totalen. Je kunt het probleem van de 'area bias' omzeilen door de gebieden niet in te kleuren, maar door symbolen in de gebieden te tekenen waarvan de grootte overeenkomt met de waarde van de variabele. Bij de grafiek van de Chinese investeringen heeft de NRC (terecht) gekozen voor zo'n symbolenkaart. Als symbool is voor de cirkel gekozen.

Figuur 4.11. De Chinese investeringen in Zuid-Amerika

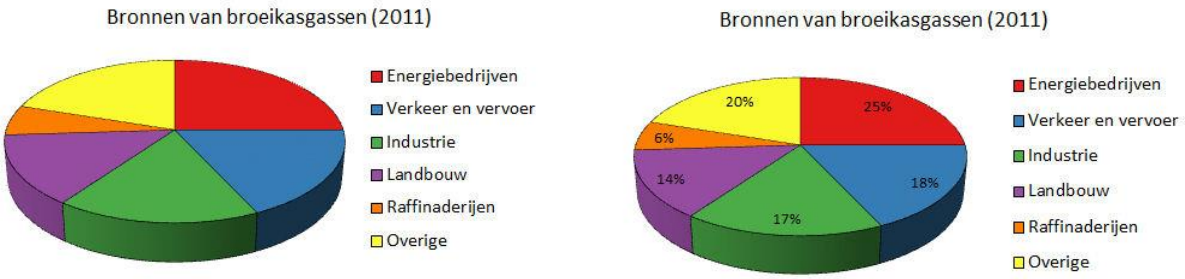


Er zijn grote verschillen tussen de oppervlakten in de krant en de correcte oppervlakten. Zo zijn de investeringen in Venezuela (50,6 miljard dollar) 5060 keer zo groot als in Uruguay (0,01 miljard dollar). Dus zou de cirkel voor Venezuela 5060 keer zo groot moeten zijn als de cirkel voor Uruguay. Dit is echter niet zo. De cirkel van Venezuela is maar 7,4 keer zo groot als de cirkel van Uruguay. Daarmee is de liegfactor in deze situatie maar liefst gelijk aan ruim 680.

Richtlijn 4: Geen driedimensionale grafieken

Grafieken in de media soms gemaakt door grafisch vormgevers die niet altijd veel verstand van statistiek hebben. Ze vinden grafieken vaak saai en pogen die daarom op allerlei manieren op te vrolijken. Dat kan bijvoorbeeld met chartjunk, met leuke symbolen in plaats van staven, en ook door er een driedimensionaal perspectief aan te geven. Dat laatste leidt misschien wel tot leukere plaatjes, maar niet tot beter interpreteerbare grafieken.

Figuur 4.12. Cirkeldiagrammen met een driedimensionaal perspectief



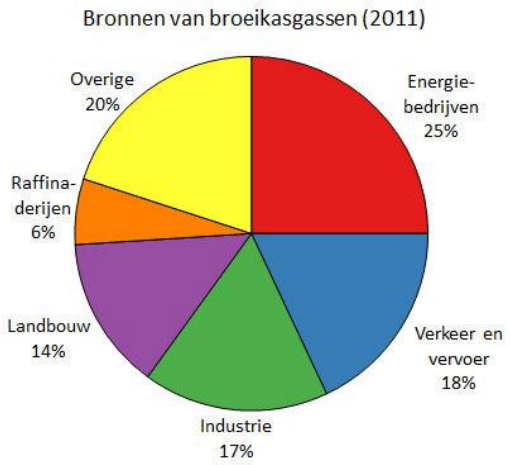
In figuur 4.12 is een driedimensionaal perspectief aangebracht in een cirkeldiagram. Zo wordt het cirkeldiagram dan met recht een 'taartdiagram'. Door dit perspectief vervormen de cirkelsectoren. De vervorming hangt af van de plaats in de grafiek. De sectoren vooraan en achteraan worden korter en breder en de sectoren aan de zijkanten worden langer en smaller. Daardoor is het lastig de verschillende sectoren met elkaar te vergelijken. De vergelijking wordt nog meer vertroebeld doordat bij sommige sectoren de zijkant wel zichtbaar is (zoals bij de industrie) en bij andere sectoren niet (zoals bij de energiebedrijven).

In de rechter grafiek in figuur 4.12 zijn de percentages in de sectoren gezet. Dat helpt om de vergelijking te vereenvoudigen. Het wordt dan echter vooral percentages vergelijken. De rol van het cirkeldiagram zelf wordt minder.

Bij de linker grafiek in figuur 4.13 zijn geen percentages aangegeven. Dat maakt de vergelijking erg moeilijk. Is de (groene) sector voor de industrie nu groter of kleiner dan de (blauwe) sector voor verkeer en vervoer? En hoe zit het met de (rode) vector voor de energiebedrijven?

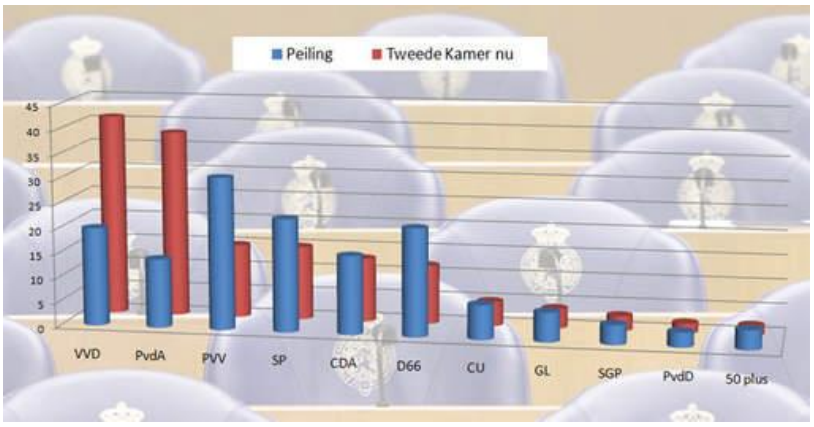
Je kunt maar beter een staafdiagram maken dan een cirkeldiagram. Als je toch een cirkeldiagram wilt, vermijdt dan in ieder geval het driedimensionaal perspectief. Als voorbeeld bevat figuur 4.13 de tweedimensionale versie van de driedimensionale cirkeldiagrammen in figuur 4.12. Om de leesbaarheid te bevorderen, zijn zowel de percentages als de namen van de categorieën bij de sectoren gezet. Merk ook op dat er maar zes categorieën zijn. Dat is volgens deskundigen het maximum aantal categorieën dat je redelijkerwijs in een cirkeldiagram kunt tonen.

Figuur 4.13. Cirkeldiagrammen zonder driedimensionaal perspectief



We geven nog een tweede voorbeeld van een grafiek met een driedimensionaal perspectief. Het gaat om een 3D-versie van een staafdiagram. De cijfers zijn afkomstig van een politieke peiling van Maurice de Hond. Veel media nemen de uitkomsten van deze peiling klakkeloos over. Soms proberen die media die uitkomsten nog wat 'op te leuken'. Dat gebeurde ook op zondag 28 oktober 2013. De Persgroep bedacht dat het wel aardig zou kunnen zijn om een grafiek van de uitkomsten te maken. Tot de Persgroep horen kranten als Trouw en het AD. En dus verscheen op de websites van beide kranten de grafiek in figuur 4.14.

Figuur 4.14. Staafdiagram met driedimensionaal perspectief

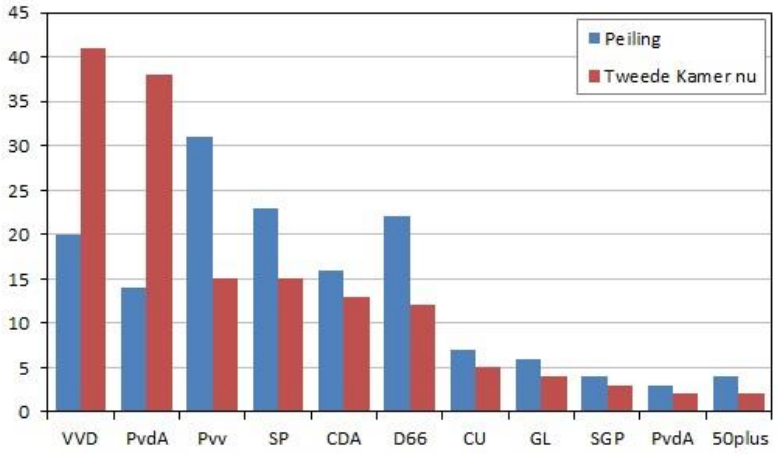


Het belangrijkste probleem van 3D-grafieken is dat je er de informatie slecht uit kunt aflezen. Het is bijna niet mogelijk om de correcte lengtes van de staven af te lezen. Probeer maar eens te bepalen wat in deze peiling de lengte van de staaf van de PvdA is (de blauwe staaf). Het is bijna niet te doen. Je moet dan in je hoofd vanaf de top van de staaf een schuine lijn naar achteren trekken (evenwijdig aan de korte lijnen links bij de as) en dan kijken waar hij de achterwand raakt.

Een ander probleem van grafieken met een driedimensionaal perspectief is dat soms staven voor een groot deel achter andere staven verdwijnen. Daardoor is het nog eens extra moeilijk om de lengte van de staven te bepalen. Ook in de grafiek hierboven zijn sommige staven maar amper te zien.

Er is natuurlijk een simpele oplossing voor het goed weergeven van de gegevens in de grafiek hierboven. Dat is een samengesteld staafdiagram zoals in figuur 4.15. Je kunt nu de lengtes van alle staven goed aflezen. De roosterlijnen spelen daarbij een nuttige rol. En ook kun je de lengtes van verschillende staven goed met elkaar vergelijken.

Figuur 4.15. Staafdiagram zonder driedimensionaal perspectief



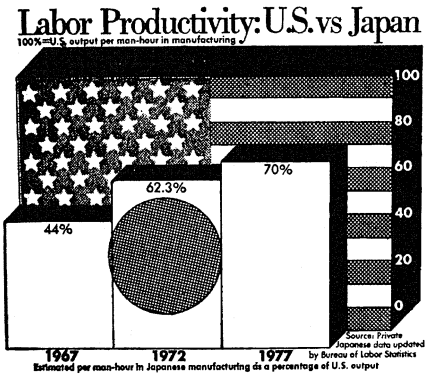
Het samengesteld staafdiagram in figuur 4.15 is misschien wel wat minder spectaculair dan de 3D-grafiek in figuur 4.14, maar je kunt de 'boodschap' in de cijfers wel veel duidelijker aflezen.

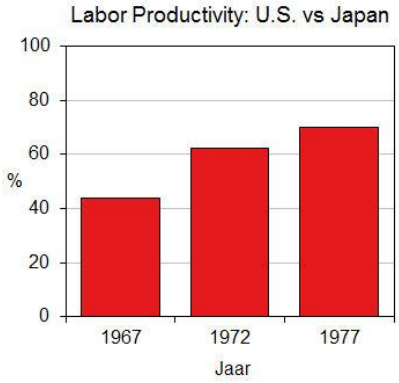
Richtlijn 5: Geen chartjunk

Om een grafiek goed te kunnen lezen, zijn twee dingen nodig: de data en de metadata. Zonder metadata weet je niet wat de data voorstellen en kun je dus ook niet de grafiek begrijpen. De grafiek moet ook niet meer bevatten dan data en metadata. Alle zaken in de grafiek die niet tot de data of metadata behoren, worden door Tufte (1983) chartjunk genoemd. Het is overbodige poespas. Het maakt de grafiek misschien wel leuker, maar niet beter leesbaar. Chartjunk leidt de lezer alleen maar af van de boodschap in de grafiek. Tot de chartjunk horen bijvoorbeeld te fijnmazige en te dikke roosterlijnen, overbodige extra teksten, exotische tekstfonts, extra plaatjes en veel te wilde arcering.

Het eerste voorbeeld van een grafiek met veel chartjunk zijn we al tegengekomen in hoofdstuk 3. Die grafiek en de verbeterde versie zijn herhaald in figuur 4.16. De grafieken tonen de relatieve arbeidsproductiviteit van de Verenigde Staten ten opzichte van Japan. De grafiek bevat slechts drie getallen (44%, 62,3% en 70%). De rest is chartjunk. In de rechter grafiek is alle chartjunk verwijdert. Er zijn slechts data en metadata overgebleven. De grafiek is nu veel beter leesbaar.

Figuur 4.16. Arbeidsproductiviteit van de VS ten opzichte van Japan



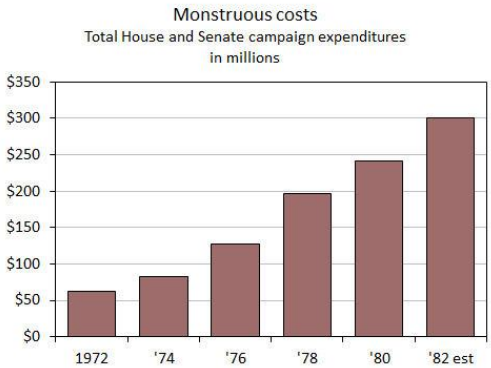


Om de hoeveelheid chartjunk in een grafiek meetbaar te maken is door Tufte de *data/ink ratio* ingevoerd. De waarde van deze index is gelijk aan de hoeveelheid 'inkt' die in de grafiek is gebruikt voor het weergeven van de echte informatie (data en metadata), gedeeld door de totale hoeveelheid gebruikte inkt. In een goede grafiek moet de data/ink ratio in de buurt van de 1 liggen. Het zal duidelijk zijn dat voor de linker grafiek in figuur 4.16 de data/ink ratio veel groter is dan 1.

De linker grafiek in figuur 4.17 is gemaakt door de Amerikaanse grafisch vormgever Nigel Holmes. Hij is juist een voorstander van het toevoegen van allerlei versiering aan een grafiek. Hij denkt dat mensen zich een dergelijke grafiek beter zullen herinneren. De grafiek van Holmes in figuur 4.17 is inderdaad nogal opvallend. Het is echter de vraag of de lezers van de grafiek zich de correcte boodschap zullen herinneren.

Behalve dat de grafiek heel wat chartjunk heeft, is er nog wel wat meer mis. De grafiek is in feite een staafdiagram waarvan de staven zijn vervangen door lange tanden. Daarbij is vergeten dat hogere tanden ook breder zijn geworden. Dus de oppervlaktes van de tanden komen niet overeen met de waarde van het verschijnsel (kosten van de verkiezingscampagne). Verder loopt de horizontale as niet horizontaal, maar schuin omhoog. Daardoor lijkt de stijging van de campagnekosten groter dan die in werkelijkheid is.

Figuur 4.17. De kosten van de verkiezingscampagnes in de V.S.



Bij de rechter grafiek in figuur 17 is alle chartjunk weggehaald. Er blijft een minimalistische grafiek over met alleen maar data en metadata. Misschien is de grafiek een beetje saai, maar wel duidelijk.

5. Grafieken voor één kwalitatieve variabele

In dit en de volgende hoofdstukken bespreken we verschillende soorten grafieken. We doen dat op een systematische manier. Aan de ene kant maken we onderscheid tussen kwalitatieve variabelen en kwantitatieve variabelen. En aan de andere kant behandelen we grafieken voor het in beeld brengen van de verdeling van één variabele en grafieken voor het tonen van de samenhang tussen twee variabelen.

In dit hoofdstuk gaat het om het in beeld brengen van de verdeling van één *kwalitatieve variabele*. Zo'n variabele doet niet anders dan de waarnemingen in een aantal groepen (categorieën) verdelen. Voorbeelden van kwalitatieve variabelen zijn geslacht (man, vrouw), burgerlijke staat (ongehuwd, gehuwd, gescheiden, weduwstaat) en geloofsovertuiging (Protestants, Rooms-katholiek, Islamitisch, Joods, …). Je kunt niet met de waarden van een kwalitatieve variabele rekenen. Er bestaat niet zoiets als het gemiddelde geslacht, de gemiddelde burgerlijke staat of de gemiddelde geloofsovertuiging.

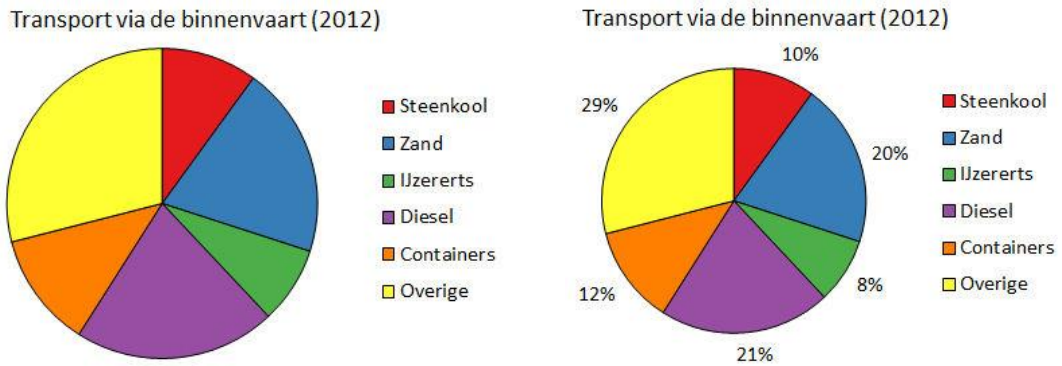
Soms is het mogelijk een natuurlijke volgorde aan te brengen in de categorieën. Een dergelijke kwalitatieve variabele noemen we een *ordinale variabele*. Een voorbeeld is de variabele opleidingsniveau. Is er geen sprake van een natuurlijke volgorde, dan noemen we dat een *nominale variabele*. Geloofsovertuiging is een voorbeeld van een nominale variabele.

Bij het in beeld brengen van de verdeling gaat het om het gedrag van één variabele. De grafiek moet duidelijk maken welke waarden voorkomen, welke waarden veel of weinig voorkomen, of er sterk afwijkende waarden zijn, enz. Het enige wat je met een kwalitatieve variabele kunt doen is de aantallen waarnemingen per categorie bepalen en die met elkaar vergelijken. We zullen hier twee technieken bespreken die dat doen: het cirkeldiagram en het staafdiagram.

Cirkeldiagram

Het cirkeldiagram is misschien wel een van de populairste grafieken in de media. Een andere naam voor het cirkeldiagram is het taartdiagram. Bij dit type grafiek verdeel je een cirkel (de taart) in evenveel sectoren (taartpunten) als er categorieën zijn. De oppervlakte van deze sectoren neem je evenredig aan de aantallen waarnemingen in de categorieën. Figuur 5.1 bevat twee varianten van cirkeldiagram van de getransporteerde goederen door de binnenvaart in 2012. De cijfers zijn afkomstig van het CBS.

Figuur 5.1. Cirkeldiagram zonder/met percentages bij de sectoren



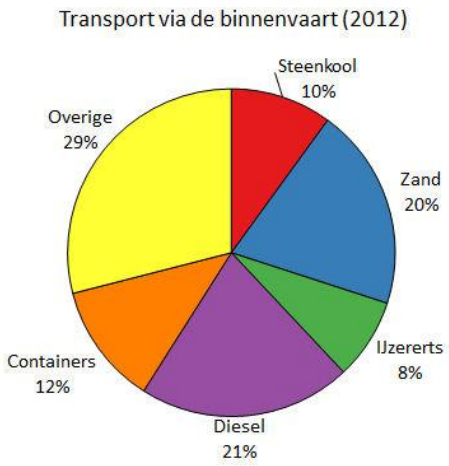
Het linker cirkeldiagram in figuur 5.1 heeft een aantal tekortkomingen. In de eerste plaats is het lastig om af te lezen hoe groot de sectoren zijn. Het is ook moeilijk sectoren onderling met elkaar te vergelijken. Is de blauwe sector (zand) nu groter of kleiner dan de paarse sector (Dieselolie)?

Je kunt dit probleem oplossen door de percentages erbij te zetten, zoals in de rechter grafiek in figuur 5.1 is gedaan. Dat lukt niet altijd. Als er een aantal kleine sectoren is, kan er onvoldoende ruime zijn voor de percentages.

Om de verschillende sectoren van elkaar te kunnen onderscheiden is het nodig om verschillende kleuren te gebruiken. Die kleuren hebben verder geen betekenis. Ze zijn niet functioneel. Het is belangrijk dat de ene kleur niet meer de aandacht trekt dan de andere kleur. Daarom moeten alle kleuren ongeveer even fel (verzadigd) zijn.

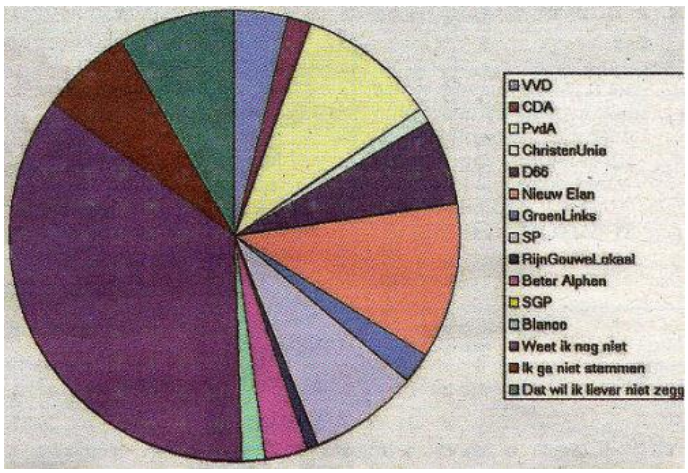
Om te kunnen bepalen welke sector bij welke groep hoort, moet je een legenda bij de grafiek zetten. Figuur 5.1 bevat daarvan een voorbeeld. Een nadeel daarvan is dat je bij het bekijken van het cirkeldiagram steeds met je ogen heen en weer moet tussen de cirkel en de legenda. Dat maakt het bekijken van de grafiek wat onrustig. Een andere oplossing kan zijn om niet alleen de percentages bij de sectoren te zetten, maar ook de namen van de groepen. Je krijgt dan zoiets als in figuur 5.2.

Figuur 5.2. Cirkeldiagram zonder legenda



Een wel heel slecht voorbeeld van een cirkeldiagram stond in september 2013 in de regionale krant *Alphen.cc*. Het was een prognose voor de gemeenteraadsverkiezingen die in november 2013 in Alphen a/d Rijn, Boskoop en Rijnwoude zouden worden gehouden. Zie figuur 5.3.

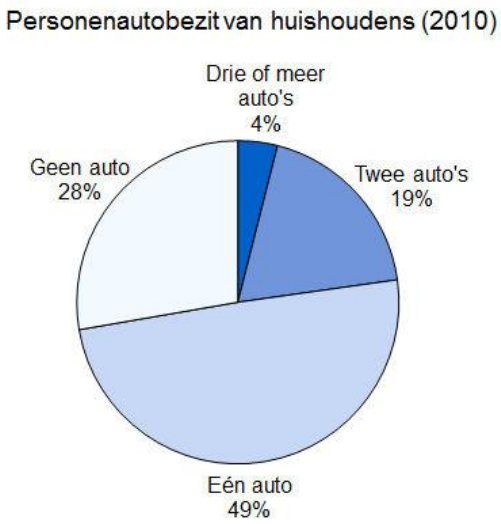
Figuur 5.3. Cirkeldiagram van de prognose van de gemeenteraadsverkiezingen



Er is van alles mis met deze grafiek. Zo zijn er heel veel sectoren. En ze zijn moeilijk uit elkaar te houden omdat de kleuren nog al op elkaar lijken. Verder zijn nergens percentages te vinden. Dat maakt het erg lastig om de verschillende partijen met elkaar te vergelijken.

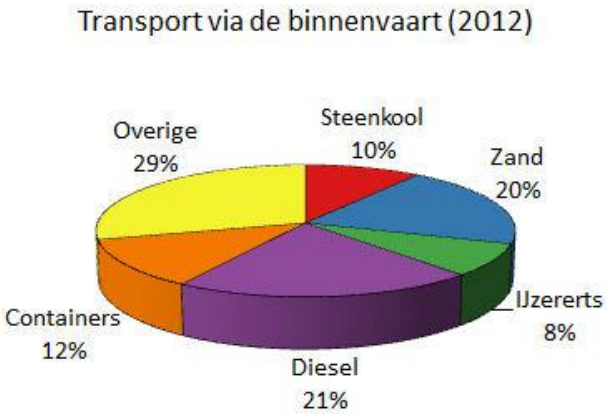
Het cirkeldiagram heeft dus tekortkomingen. Beperk het gebruik ervan daarom zoveel mogelijk. Zo adviseert Statistics Canada, het statistisch bureau van Canada, om alleen een cirkeldiagram te maken als er hooguit zes verschillende groepen zijn, die groepen niet allemaal ongeveer even groot zijn, en het doel is om aan te geven welk aandeel een groep heeft in het grote geheel. Figuur 5.4 bevat een cirkeldiagram dat hier redelijk aan voldoet. Het gaat hier om het aantal auto's per huishouden. De cijfers zijn afkomstig van het CBS en gaan over 2013. Er zijn slechts vier groepen. Ook staan de namen bij de sectoren, zodat een legenda niet nodig is. Omdat het aantal auto's per huishouden een natuurlijke volgorde heeft (geen auto's, één auto, twee auto's en drie auto's of meer), zijn er tinten van één kleur gebruikt en niet vier verschillende kleuren.

Figuur 5.4. Cirkeldiagram van het aantal auto's per huishouden



Er zijn nog twee variaties op het cirkeldiagram die regelmatig langskomen in de media. De eerste is het cirkeldiagram met driedimensionaal perspectief. Als voorbeeld bevat figuur 5.5 de 3D-versie van de grafiek in figuur 5.2.

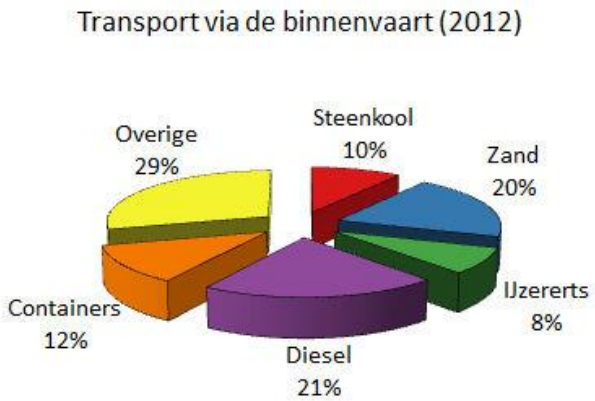
Figuur 5.5. Een cirkeldiagram in driedimensionaal perspectief



Door het driedimensionaal perspectief vervormen de sectoren. De sectoren aan de voorkant en achterkant zijn kort en breed (zoals die van Diesel en Steenkool), terwijl sectoren aan zijkant smal en lang zijn (zoals die Containers en IJzererts). Daardoor is het nog moeilijker om sectoren met elkaar te vergelijken. Verder lijken sectoren aan de voorkant groter omdat ook de zijkant van het cirkeldiagram zichtbaar is (zoals bij Diesel). Daardoor lijkt zo op het oog de sector van Diesel groter dan die van overige. De percentages wijzen echter anders uit. Het zal duidelijk zijn dat het gebruik van cirkeldiagrammen met een driedimensionaal perspectief sterk moet worden afgeraden.

Figuur 5.6 bevat een cirkeldiagram waarin de verschillende sectoren uit elkaar zijn getrokken. Ze noemen dit in het Engels een *exploding pie chart*. De taart is als het ware in punten gesneden en die punten zijn uit elkaar geschoven. Dit vertroebelt de interpretatie van de grafiek nog meer. Het dient geen enkel doel.

Figuur 5.6. Een cirkeldiagram met uit elkaar geschoven sectoren



Figuur 5.6 is een slecht cirkeldiagram vanwege het driedimensionaal perspectief en de uit elkaar getrokken sectoren. Ook als je het driedimensionaal perspectief verwijdert, blijft het moeilijk om het overblijvende grafiek met uit elkaar getrokken sectoren te interpreteren. Er is misschien één situatie waarin het wel acceptabel is, en dat is de situatie waarin je één sector heel veel nadruk wilt geven. Dan kun je alleen die sector uit de cirkel wegtrekken.

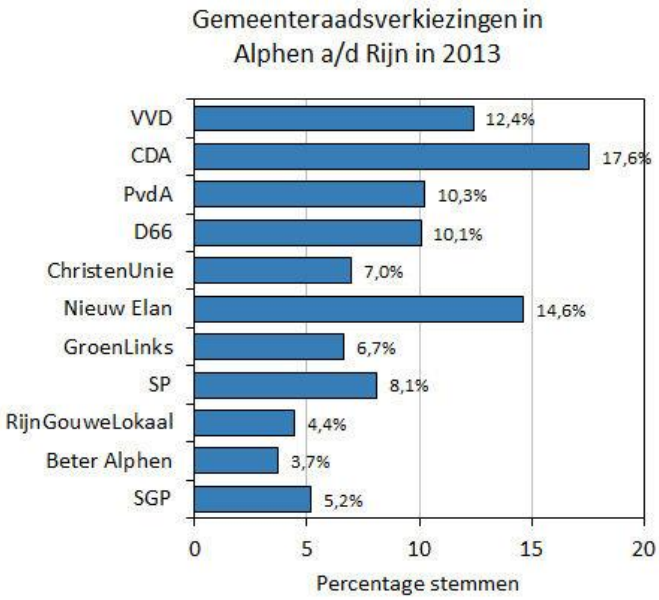
Samenvatting

- Het aantal sectoren in het cirkeldiagram mag niet groter zijn dan zes.
- Een cirkeldiagram is alleen zinvol als je wilt tonen welk aandeel de groepen in het geheel hebben. Voor onderlinge vergelijking van categorieën is een cirkeldiagram niet geschikt.
- Als er sprake is van een nominale variabele, gebruik dan verschillende kleuren voor de verschillende categorieën. Alle kleuren moeten even helder zijn.
- Als er sprake is van een ordinale variabele, gebruik dan verschillende tinten van dezelfde kleur voor de verschillende categorieën.
- Zet de aantallen of percentages bij de sectoren.
- Indien mogelijk, zet dan de namen van de categorieën bij de sectoren. De legenda kan dan verdwijnen.
- Begin het cirkeldiagram met de eerste categorie op ’12 uur’ en vul de cirkel met de wijzers van de klok mee. Als er geen natuurlijke ordening is, kun je overwegen de categorieën in oplopende of aflopende omvang te ordenen.

Staafdiagram

In de meeste gevallen werkt een staafdiagram beter dan een cirkeldiagram. Het is eenvoudiger om een staafdiagram te interpreteren. Ook kan een staafdiagram goed uit de voeten met veel meer categorieën. We leggen uit waaraan een goed staafdiagram moet voldoen. Dat doen we aan de hand van de uitslag van de gemeenteraadsverkiezingen in Alphen a/d Rijn, Boskoop en Rijnwoude in november 2013. Er waren daar verkiezingen vanwege de fusie van de drie gemeenten. Figuur 5.7 bevat het staafdiagram met de uitslag van die verkiezingen.

Figuur 5.7. Een goed staafdiagram



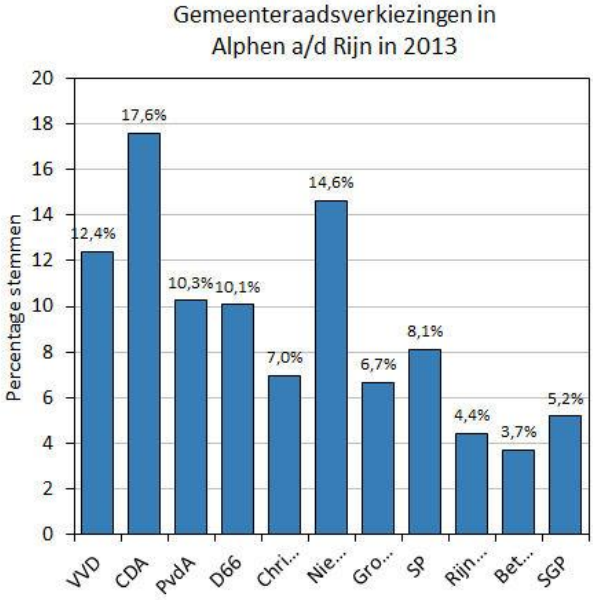
De lengtes van de staven weerspiegelen de percentages stemmen voor de partijen. Duidelijk is te zien dat het CDA de grootste partij is geworden, gevolgd door de lokale partij Nieuw Elan. Een andere lokale partij, Beter Alphen, heeft de minste stemmen gekregen.

Merk op dat de verschillende staven duidelijk van elkaar zijn gescheiden. Die tussenruimte helpt bij het lezen van de grafiek. Bovendien voorkom je zo verwarring met een histogram (zie hoofdstuk 6). Volgens Wallgren et al. (1996) moeten staven duidelijk breder zijn dan de tussenruimten. In de voorbeelden is de tussenruimte ongeveer 50% van de breedte van de staaf.

Omdat de staven los van elkaar zijn getekend, is het ook niet nodig om ze verschillende kleuren te geven. Eén kleur volstaat voor alle staven. Het is zelfs denkbaar om in plaats van kleur een grijstint te gebruiken. Met een kleur kun je echter de data beter onderscheiden van de metadata.

Je ziet vaak staafdiagrammen met verticale staven. Het is echter beter om de staven horizontaal te tekenen. Daarmee voorkom je verwarring met een histogram (die uit verticale staven bestaat). Bovendien kun je allerlei teksten beter kwijt. Figuur 5.8 bevat een voorbeeld van de problemen die je kunt krijgen bij een staafdiagram met verticale staven. Sommige tekst staat horizontaal en andere tekst verticaal. Voor de namen van de staven is onvoldoende ruimte. Daarom zijn die schuin gezet en afgekapt.

Figuur 5.8. Een niet zo goed staafdiagram



Er is één uitzondering op de regel van de horizontale staven. Als je het verloop van een verschijnsel in de tijd wilt weergeven, dan is het wel goed om de staven verticaal te tekenen. Dat komt omdat het gebruikelijk is om de tijd op de horizontale as af te zetten.

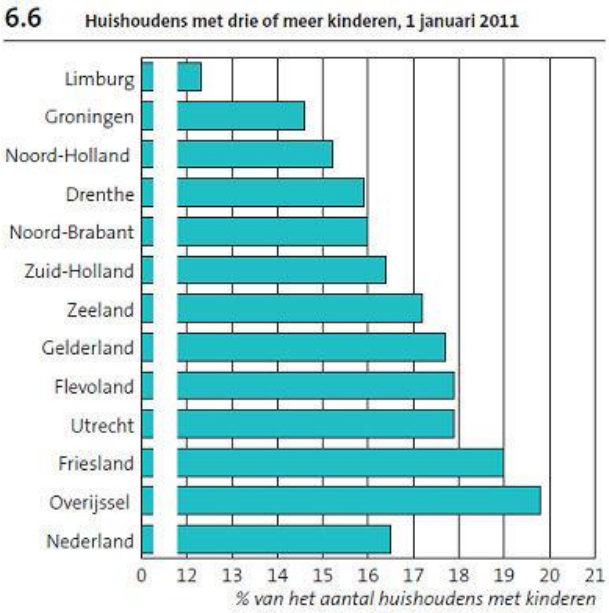
Als de lengte van een staaf een omvang, hoeveelheid, waarde, duur, aantal of percentage aanduidt, dan is er een natuurlijk nulpunt. De as moet dan bij dit nulpunt beginnen. Als dat niet het geval is dan lijken allerlei verschijnselen groter dan ze in werkelijkheid zijn. Er is dus sprake van overdrijving.

De linker grafiek in figuur 5.9 geeft een voorbeeld waarbij met de lengtes van de staven is gerommeld. De grafiek is afkomstig uit een Statistisch Jaarboek van het CBS (2012). Dit staafdiagram geeft het percentage gezinnen aan met drie of meer kinderen (als percentage van alle gezinnen met kinderen).

Het natuurlijke nulpunt staat wel vermeldt aan het begin van de horizontale as, maar daarna is een stuk van de as er tussenuit geknipt. Dit is aangegeven met een witte balk. In de praktijk kom je ook andere manieren tegen om zo'n knip in de as aan te geven:

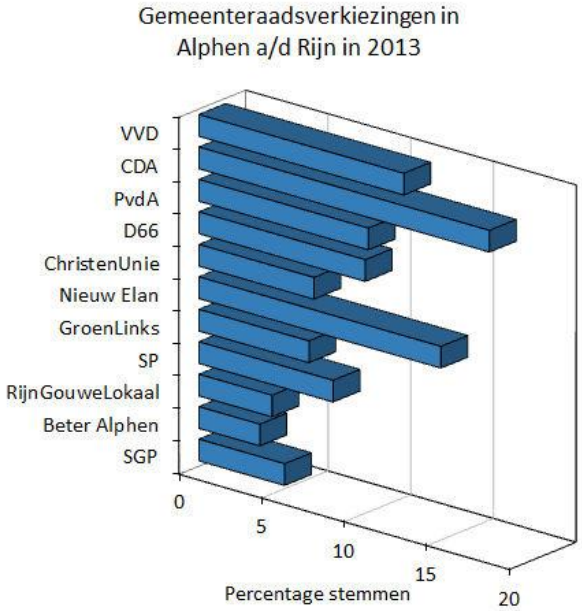
Door de knip in de as lijken de verschillen tussen de provincies groter dan ze in werkelijkheid zijn. Een voorbeeld: Het verschil tussen Limburg en Groningen is wel erg groot. Het percentage gezinnen met drie of meer kinderen lijkt maar een derde van het percentage in Groningen. In werkelijkheid is dit niet het geval. Het percentage voor Limburg ligt tussen de 12 en 13 procent. En het percentage voor Groningen ligt tussen 14 en 15 procent. Dus de verschillen zijn in feite veel kleiner.

Figuur 5.8. Een niet zo goed staafdiagram



Ten slotte bevat figuur 5.8 nog een voorbeeld van een staafdiagram met een driedimensionaal perspectief. Het is weer de uitslag van verkiezingen in Alphen a/d Rijn, Boskoop en Rijnwoude die in beeld is gebracht. Omdat de staven ergens in de ruimte zweven, is het lastige om de lengtes van die staven om een correcte manier af te lezen. Zo'n 3D-staafdiagram lijkt misschien wel cool, maar het is niet functioneel. Niet doen dus.

Figuur 5.8. Een niet zo goed staafdiagram



Samenvatting

- Teken de staven van het staafdiagram horizontaal.
- Als de staven opeenvolgende tijdstippen voorstellen, teken de staven dan wel verticaal. De horizontale as geeft dan de tijd aan.
- Teken de staven niet tegen elkaar aan, maar zorg voor tussenruimte. Die tussenruimte mag niet groter zijn dan de breedte van de staven. Een redelijke tussenruimte is 40% of 50% van de breedte van de staaf.
- De horizontale as moet bij 0 beginnen. Vermijd een knip in de as.
- Als de groepen geen natuurlijke volgorde hebben, kunnen je de staven ordenen van klein naar groot (of van groot naar klein).
- Geef alle staven dezelfde kleur (of grijstint).
- Eventueel kun je in of bij elke staaf de bijbehorende waarde zetten.
- Roosterlijnen kunnen helpen bij het aflezen van de lengte van de staven. Teken ze niet te opvallend.
- Vermijdt allerlei onnodige toeters en bellen (‘chart junk’).
- Vermijd driedimensionaal perspectief.

6. Grafieken voor één kwantitatieve variabele

In dit hoofdstuk bespreken drie grafiektypen die de verdeling van een kwantitatieve variabele in beeld brengen. Het gaat hierbij dus om variabelen die een omvang, gewicht, duur of waarde meten. De grafieken die aan bod komen, zijn het spreidingsdiagram, de snorrendoos en het histogram. Deze grafieken kunnen antwoord geven op vragen als:

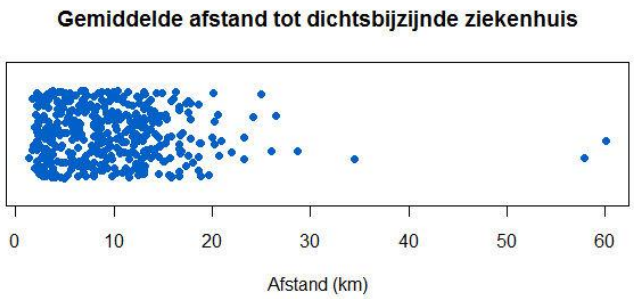
- Welke waarden komen wel en niet voor?
- Welke waarden komen veel voor en welke waarden weinig?
- Is er sprake van groepen die zich afwijkend gedragen.
- Is er sprake van extreem afwijkende waarden (uitschieters)?

We gebruiken de grafieken om een bestand met gegevens over gemeenten in beeld te brengen. Voor elke gemeente is de gemiddelde afstand van de inwoners tot de dichtstbijzijnde huisarts en het dichtstbijzijnde ziekenhuis berekend. Het gaat om de gegevens van 431 gemeenten die in 2010 door het CBS zijn verzameld.

Spreidingsdiagram

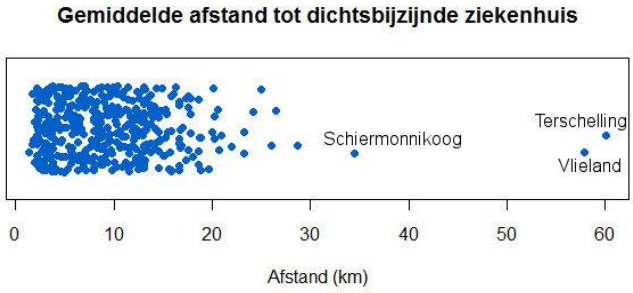
Het *spreidingsdiagram (*Eng: *strip plot* of *one-dimensional scatterplot*) brengt de verdeling in zijn meest pure vorm in beeld. Op een horizontale as wordt een schaalverdeling aangebracht en op die schaalverdeling worden de waarden als individuele punten afgezet. Om te voorkomen dat punten met dezelfde waarde over elkaar heen vallen, waardoor ze niet goed meer zichtbaar zijn, is verticale ruis toegevoegd. In het Engels noemen ze dit *jitter*. Overlappende punten worden zo uit elkaar getrokken. Figuur 6.1 bevat een spreidingsdiagram van de gemiddelde afstand tot het dichtstbijzijnde ziekenhuis voor alle 341 gemeenten.

Figuur 6.1. Een spreidingsdiagram



Uit de grafiek kun je aflezen dat voor het overgrote deel van de gemeenten de afstand tot het dichtstbijzijnde ziekenhuis onder de 15 kilometer ligt. Er zijn echter wel uitschieters. Twee heel duidelijke uitschieters zijn gemeenten met een afstand van ongeveer 60 kilometer. En dan is er ook nog een gemeente met een afstand van ongeveer 35 kilometer tot het ziekenhuis. Om beter te kunnen begrijpen wat er aan de hand is, kun je de namen van de bij drie uitschieters zetten. Dat is gebeurd in figuur 6.2.

Figuur 6.2. Een spreidingsdiagram met benoemde uitschieters

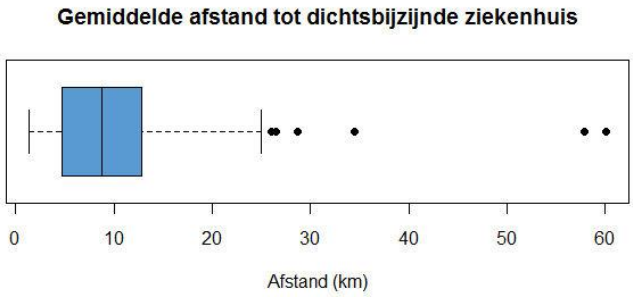


Uit deze tweede grafiek blijkt dat de uitschieters de drie Waddeneilanden Terschelling, Vlieland en Schiermonnikoog zijn. Aangezien er op die eilanden geen ziekenhuizen zijn, is het niet zo verbazingwekkend dat het dichtstbijzijnde ziekenhuis ver weg is.

Snorrendoos

Het spreidingsdiagram geeft de verdeling in zijn meeste pure vorm weer. Alle individuele waarden zijn in de grafiek terug te vinden. Een eerste stapje naar een wat meer schematische vorm is de snorrendoos. Snorrendoos is de Nederlandse vertaling van box-and-whisker plot. Soms noemen we deze grafiek ook wel een box plot. De term is oorspronkelijk bedacht door Tukey (1977). In een snorrendoos karakteriseer je de verdeling door een doos waar aan beide zijden snorharen uitsteken. Een voorbeeld van een snorrendoos is te zien in figuur 6.3. In deze grafiek zijn dezelfde gegevens gebruikt als in figuren 6.1 en 6.2.

Figuur 6.3. Een snorrendoos



De snorrendoos bestaat uit een rechthoekige doos. De doos geeft het gebied aan waarbinnen precies de middelste helft (50%) van de waarnemingen ligt. De horizontale streep in de doos duidt de mediaan aan, de middelste waarde. Van de linker- en rechterzijkant van de doos lopen lijnen (de 'snorharen') naar de zogenaamde 'naburige waarden'. Dit zijn de waarnemingen die nog net binnen een afstand van 1,5 maal de lengte van de doos van de linker resp. rechterkant van de doos liggen. De linker snorhaar geeft het gebied aan waarin de kleinste 25% van de waarden ligt. En de rechter snorhaar toont waar de 25% grootste waarden liggen.

Alle waarnemingen die nog verder weg liggen dan de snorharen, worden apart getekend. Dit zijn extreme, misschien wel atypische waarden. Ze worden ook wel uitschieters genoemd.

Uit de grafiek kun je afleiden dat de middelste waarde omstreeks 9 kilometer is. Dit betekent dat de voor de helft van de gemeenten de gemiddelde afstand tot het dichtstbijzijnde ziekenhuis kleiner dan 9 kilometer is. En voor de andere helft van de gemeenten is het meer dan 9 kilometer.

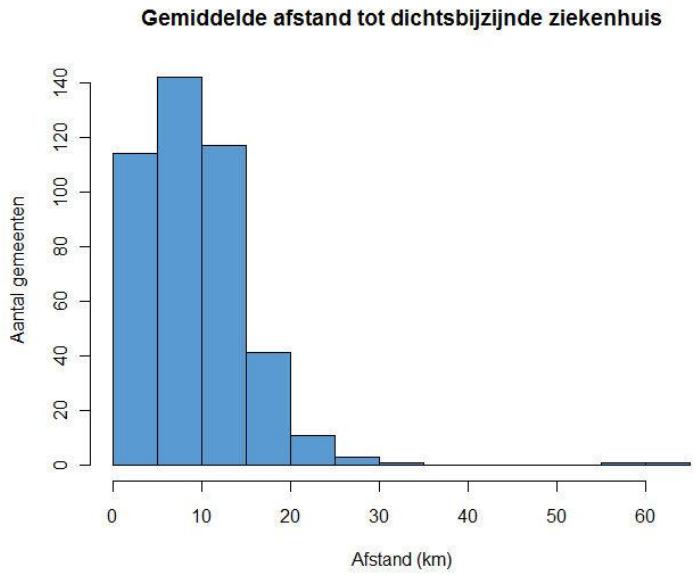
De rechter snorhaar is langer dan de linker snorhaar. Daaruit kun je concluderen dat er meer spreiding van waarden is aan de rechterkant van de verdeling. De snorrendoos duidt 6 waarden aan als uitschieters. Dat zijn niet alleen de drie al eerder genoemden Waddeneilanden Terschelling, Vlieland en Schiermonnikoog, maar ook Ameland (29 kilometer), Harlingen (27 kilometer) en Nijefurd (26 kilometer). Eventueel kun je ook in een de snorrendoos de uitschiters benoemen door de namen erbij te zetten.

Histogram

De traditionele manier om de verdeling van een kwantitatieve variabele weer te geven is het histogram. Om een histogram te kunnen tekenen moet je het waardebereik van de variabele eerst in een aantal klassen verdelen, en voor elke klasse het aantal waarden daarin tellen. Vervolgens verdeel je een horizontale as in intervallen die overeenkomen met deze klassen. Boven elk interval teken je een kolom waarvan de oppervlakte evenredig is met het aantal waarnemingen in die klasse. De kolommen moeten je zo tekenen dat ze elkaar raken. Het is immers een continue verdeling die je hier weergeeft.

Figuur 6.4 bevat een voorbeeld van een histogram. Dezelfde gegevens zijn gebruikt als in figuren 6.1 t/m 6.3. Er is sprake van een verdeling met één top. Die top ligt in de buurt van de 9 kilometer. De verdeling is niet symmetrisch. De verdeling loopt aan de rechterkant verder door dan aan de linkerkant. Hij is scheef naar rechts. Dat betekent dat er veel kleine waarden zijn en weinig grote waarden. De uitschieters zijn in dit histogram minder goed zichtbaar.

Figuur 6.4. Een histogram



Een punt van overweging is het aantal intervallen dat je kiest. Zijn er heel veel kleine intervallen, dan vertroebelt veel lokale variatie het algemene beeld van de verdeling. En als je weinig intervallen neemt, dan ontstaat er wel een heel globaal beeld, waarbij interessante details onopgemerkt kunnen blijven. Je kunt eventueel een vuistregel hanteren die adviseert om het aantal intervallen ongeveer gelijk te nemen aan de wortel uit het aantal waarden (met een minimum van 5 en een maximum van 20). Die regel is toegepast in figuur 6.3. Het aantal intervallen is 20 en dat is ongeveer gelijk aan de wortel uit 431 (het aantal gemeenten). Als je een enorm grote hoeveelheid gegevens in beeld wilt brengen (zeg, meer dan een miljoen waarden), dan zouden die 20 intervallen wel eens wat mager kunnen blijken te zijn. Er worden

in de literatuur ook andere recepten voorgesteld om het aantal intervallen te bepalen. Zie bijvoorbeeld Sturges (1926) en Scott (1979).

Je tekent een histogram altijd met verticale staven. Er is echter één uitzondering: de *bevolkingspiramide*. Die bestaat in feite uit twee histogrammen: één voor de mannen en één voor de vrouwen. Die twee histogrammen zijn gekanteld en tegen elkaar aan geplakt. Zo is de leeftijdsopbouw van de mannen beter vergelijkbaar met die voor de vrouwen. Figuur 6.4 bevat een voorbeeld.

Figuur 6.4. Een bevolkingspiramide

Samenvatting

- Teken de waarde-as van het spreidingsdiagram en de snorrendoos horizontaal.
- Als de variabele een omvang, gewicht, duur of waarde meet, laat de as van het spreidingsdiagram, de snorrendoos en het histogram dan bij 0 beginnen. Vermijd een knip in de as.
- Teken de staven van het histogram verticaal.
- Teken de staven van het histogram tegen elkaar aan, dus zonder tussenruimte.
- Geef alle punten van het spreidingsdiagram en de snorrendoos en alle staven van het histogram dezelfde kleur (of grijstint).
- Roosterlijnen kunnen helpen bij het aflezen van waarden in de grafiek. Teken ze niet te opvallend.
- Vermijdt allerlei onnodige toeters en bellen ('chart junk').
- Vermijd driedimensionaal perspectief.

7. Grafieken voor twee kwalitatieve variabelen

Er zijn grafieken die een mogelijke samenhang tussen variabelen in beeld kunnen brengen. In dit hoofdstuk gaat het om de samenhang tussen kwalitatieve variabelen. Dat zijn dus variabelen die een indeling in verschillende categorieën maken. We bespreken in dit hoofdstuk enkele typen grafieken die gebaseerd zijn op de al eerder besproken staafdiagrammen en cirkeldiagrammen.

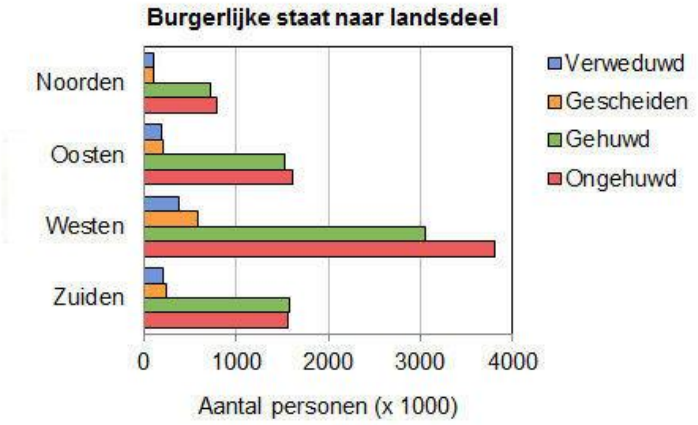
De mogelijkheden voor het onderzoek naar de samenhang tussen kwalitatieve variabelen zijn beperkt. Dit heeft te maken met de al eerder genoemde onmogelijkheid om met de waarden van kwalitatieve variabelen berekeningen uit te voeren.

In de voorbeelden kijken we naar een mogelijk verband in Nederland tussen burgerlijke staat (ongehuwd, gehuwd, gescheiden, verweduwd) en landsdeel (noorden, oosten, zuiden, westen). De gegevens zijn afkomstig van het CBS.

Samengesteld staafdiagram

Bij het *samengesteld staafdiagram* teken je voor elke categorie van de ene kwalitatieve variabele een staafdiagram van de andere variabele. Al die staafdiagrammen worden onder elkaar getekend in dezelfde figuur. Figuur 7.1 bevat een voorbeeld van een dergelijke grafiek. Het gaat hier om de verdeling per landsdeel van de verschillende categorieën van burgerlijke staat.

Figuur 7.1. Samengesteld staafdiagram



Ook hier is het weer beter de staven horizontaal te tekenen in plaats van verticaal. Daarmee voorkom je verwarring met histogrammen. Bovendien is er zo meer ruimte om teksten horizontaal neer te zetten.

Merk op dat in deze grafiek van het principe is afgestapt om de staven los van elkaar te tekenen. Binnen een categorie van de andere variabele zijn de staven tegen elkaar aan 'geplakt'. Dit is gedaan om de verschillende categorieën van de andere variabele beter te kunnen onderscheiden.

In de grafiek is een staafdiagram van de categorieën van burgerlijke staat getekend voor elk landsdeel. Het kan natuurlijk ook andersom: een staafdiagram van de verdeling over de landsdelen binnen elke categorie van burgerlijke staat.

Zien de verschillende staafdiagrammen er anders uit, dan is er sprake van samenhang tussen de twee variabelen. Hoe die samenhang dan in elkaar zit, kun je nagegaan door nader onderzoek van de staafdiagrammen. Sommige aspecten kun je wel goed aflezen uit de grafiek en andere

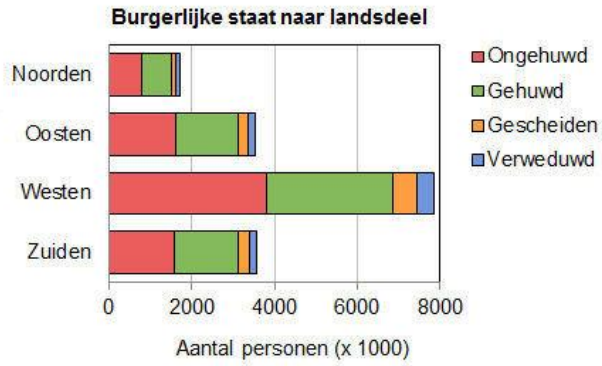
aspecten niet. Zo is moeilijk te zien hoe groot de omvang van de bevolking is in elk landsdeel. Daarvoor zou je immers alle staven van het staafdiagram voor dat landsdeel op elkaar moeten stapelen. Zo op het oog lijkt het westen de meeste inwoners te hebben, maar welk landsdeel komt op de tweede plaats? Het oosten of het zuiden?

Ook kun je moeilijk aflezen of een bepaalde burgerlijke staat relatief gezien over- of ondervertegenwoordigd is in een landsdeel. De vraag of het percentage gescheiden personen in het westen nu groter of kleiner is dan het percentage in het zuiden, is niet zonder meer te beantwoorden. De absolute omvangen van de verschillende categorieën burgerlijke staat zijn wel goed af te lezen. Zo wonen in het westen de meeste ongehuwden.

Stapeldiagram

Een andere manier om de staafdiagrammen van de ene variabele voor elk categorie van de andere variabele in beeld te brengen, is het *stapeldiagram*. Hierbij zet je de staven niet onder elkaar maar tegen elkaar aan. Figuur 7.2 geeft een voorbeeld van een dergelijk stapeldiagram.

Figuur 7.2. Een stapeldiagram

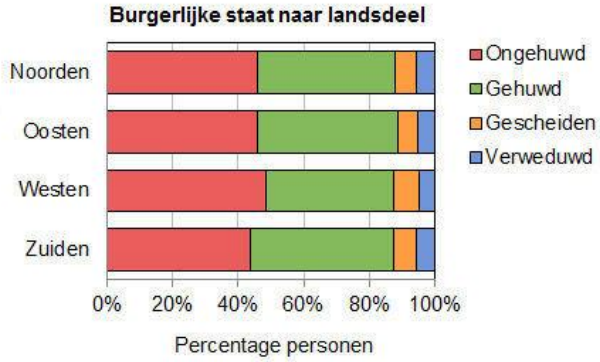


De staven zijn zo getekend dat hun lengtes overeenkomen met de aantallen inwoners in de diverse landsdelen. Op hun beurt zijn de staven verdeeld in segmenten die verdeling over de vier categorieën van burgerlijke staat in de juiste verhouding weergeven.

Wat valt er te zien aan een stapeldiagram? Net zoals bij een simpel staafdiagram is hier duidelijk te zien welke categorie van de ene variabele (hier: landsdeel) het grootste is en welke het kleinste. Verder kun je ook aardig inschatten welke categorie van de andere variabele (hier: burgerlijke staat) relatief goed of slecht is vertegenwoordigd binnen een categorie van de ene variabele. Zo is te zien dat er maar weinig gescheiden personen en verweduwde persoon in het noorden zijn, maar de totale bevolking is ook klein. Het vergelijken van de verdeling van burgerlijke staat tussen twee landsdelen is vrij lastig. Daarvoor is het samengesteld staafdiagram meer geschikt.

Een manier om wat meer inzicht te krijgen in dit laatste aspect, kun je ook een stapeldiagram tekenen waarbij de staven tot 100 zijn gepercenteerd. Dat betekent dat alle staven even lang worden (100%), maar binnen de zijn de onderlinge verhoudingen van de andere variabele nog steeds goed weergegeven. In voorbeeld staat in figuur 7.3.

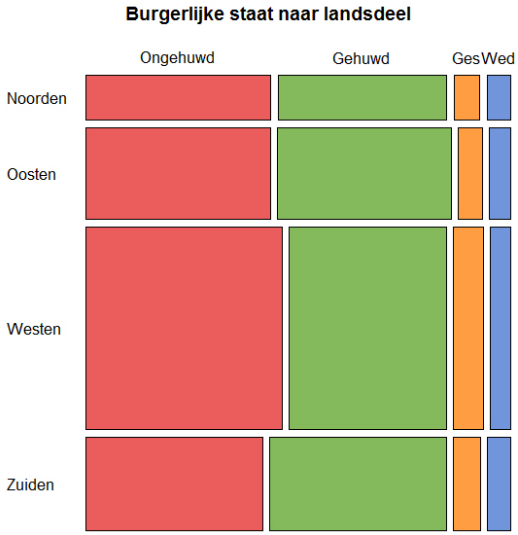
Figuur 7.3. Een stapeldiagram, met staven die optellen tot 100%



Nu kun je ook de relatieve vertegenwoordiging van de categorieën van burgerlijke staat binnen de landsdelen beter beoordelen. Zo zijn er in het westen relatief veel ongehuwden en relatief weinig gehuwden. Het relatieve aandeel gescheiden en verweduwde personen is in het noorden niet anders dan in andere landsdelen.

Figuur 7.4 toont een aangepaste versie van de het stapeldiagram die het *mozaïekdiagram* (Eng: mosaic plot) wordt genoemd.

Figuur 7.4. Een mozaïekdiagram

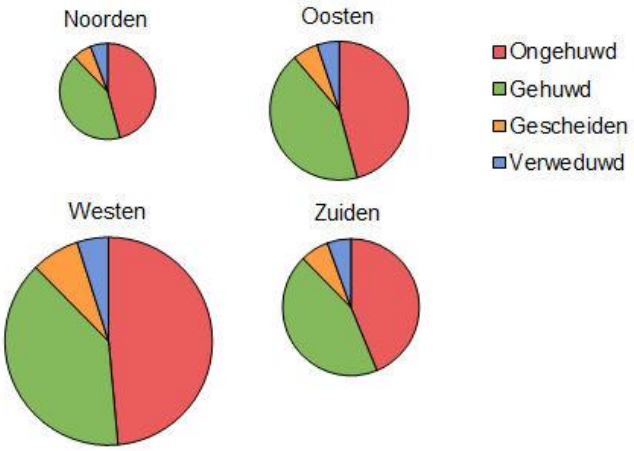


De horizontale banden voor de verschillende landsdelen, zijn nu niet meer even dik zoals bij het stapeldiagram in figuur 7.3, maar evenredig aan het aantal inwoners in de landsdelen. Zo is duidelijk te zien dat in het westen de meeste mensen wonen (dikke band) en in het noorden de minste (dunne band).

Cirkeldiagrammen

In hoofdstuk 5 is het cirkeldiagram besproken als alternatief voor het staafdiagram. Ook voor het onderzoek van de samenhang tussen twee kwalitatieve variabelen kun je eventueel cirkeldiagrammen gebruiken. De meest simpele aanpak is om de staven in het stapeldiagram in figuur 7.3 te vervangen door de bijbehorende cirkeldiagrammen. Maar een wat informatievere optie is om met de grootte van de cirkels de omvang van de categorieën van de andere variabele aan te geven. Dat is gebeurd in figuur 7.5.

Figuur 8.5. Cirkeldiagrammen



De cirkels zijn zo getekend dat hun oppervlaktes evenredig zijn met de aantallen inwoners in de bijbehorende landsdelen. Uit de grafiek kun je niet alleen aflezen wat kleine landsdelen zijn en wat grote, maar ook hoe de verdeling over de verschillende burgerlijke staten binnen de landsdelen is. De grafiek in figuur 7.5 is daarmee wel wat te vergelijken met het mozaïekdiagram. In tegenstelling tot het mozaïekdiagram is het hier wat lastiger om de omvangen van bepaalde categorieën van burgerlijke staat voor verschillende landsdelen met elkaar te vergelijken.

Samenvatting

- Teken de staven van het samengesteld staafdiagram of stapeldiagram horizontaal.
- Als de staven opeenvolgende tijdstippen voorstellen, teken de staven dan wel verticaal. De horizontale as geeft dan de tijd aan.
- Teken de staven van de ene variabele binnen een categorie van de andere variabele tegen elkaar aan. Er is wel tussenruimte tussen de verschillende categorieën van die andere variabele.
- Als er sprake is van een nominale andere variabele binnen de categorieën van de ene variabele, gebruik dan verschillende kleuren voor de verschillende categorieën. Alle kleuren moeten even helder zijn.
- Als er sprake is van een ordinale andere variabele binnen de categorieën van de ene variabele, gebruik dan verschillende tinten van dezelfde kleur voor de verschillende categorieën.
- De horizontale as moet bij 0 beginnen. Vermijd een knip in de as.
- Als de groepen geen natuurlijke volgorde hebben, kun je in een stapeldiagram de staven van ordenen van klein naar groot (of van groot naar klein).
- Roosterlijnen kunnen helpen bij het aflezen van de lengte van de staven. Teken ze niet te opvallend.
- Vermijdt allerlei onnodige toeters en bellen (‘chart junk’).
- Vermijd driedimensionaal perspectief.

8. Grafieken voor twee kwantitatieve variabelen

In dit hoofdstuk gaat het om het in beeld brengen van de samenhang tussen twee kwantitatieve variabelen. Het gaat hierbij dus om variabelen die een omvang, gewicht, duur of waarde meten. Er komt maar één grafiek aan bod komen en dat is de *puntenwolk*.

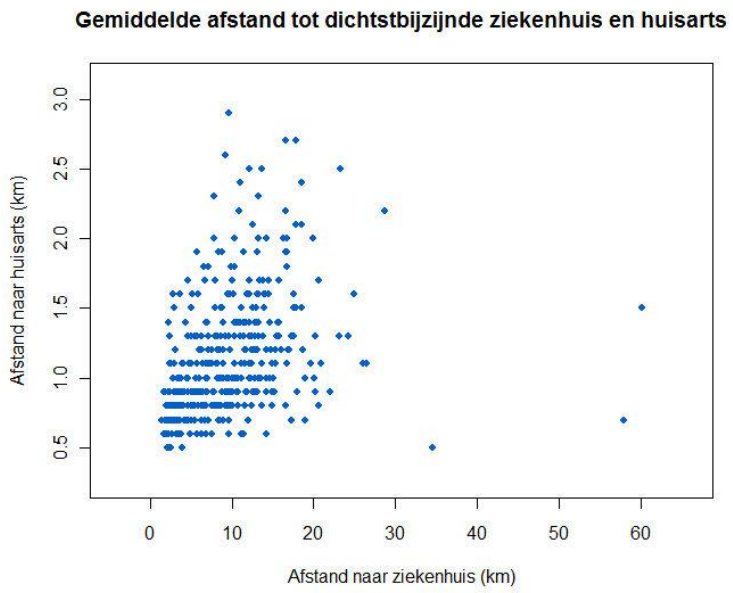
We gebruiken dit type grafiek om een bestand met gegevens over gemeenten in beeld te brengen. Voor elke gemeente is de gemiddelde afstand van de inwoners tot de dichtstbijzijnde huisarts en het dichtstbijzijnde ziekenhuis berekend. Het gaat om de gegevens van 431 gemeenten die in 2010 door het CBS zijn verzameld.

Puntenwolk

Voor de grafische analyse van de samenhang tussen twee kwantitatieve variabelen is de puntenwolk de meest geëigende grafische techniek. Bij de puntenwolk vat je elk waarnemingspaar van beide variabelen op als de coördinaten van een punt in een rechthoekig assenstelsel. Zijn er duidelijke patronen te herkennen in de puntenwolk, dan is er sprake van een bepaalde vorm van samenhang. Deze moet je dan verder onderzoeken en verklaren. De meest extreme vorm van samenhang is die waarbij alle punten op een rechte lijn liggen. In dat geval kun je de waarde van de ene variabele exact voorspellen uit de waarde van de andere variabele. Maar je kunt ook andere aspecten ontdekken in een puntenwolk. Zo zal het vrij simpel zijn om uitschieters te detecteren. En ook afwijkende groepjes waarnemingen zullen duidelijk opvallen. Heeft de puntenwolk de vorm van een ongestructureerde sneeuwbui, dan is er geen samenhang.

Figuur 8.1 bevat de een voorbeeld van een puntenwolk. Het betreft gegevens over de 431 gemeenten in Nederland (in 2010). Op horizontale as staat de gemiddelde afstand van de inwoners tot het dichtstbijzijnde ziekenhuis. Op de verticale as staat de gemiddelde afstand van de inwoners tot de dichtstbijzijnde huisarts.

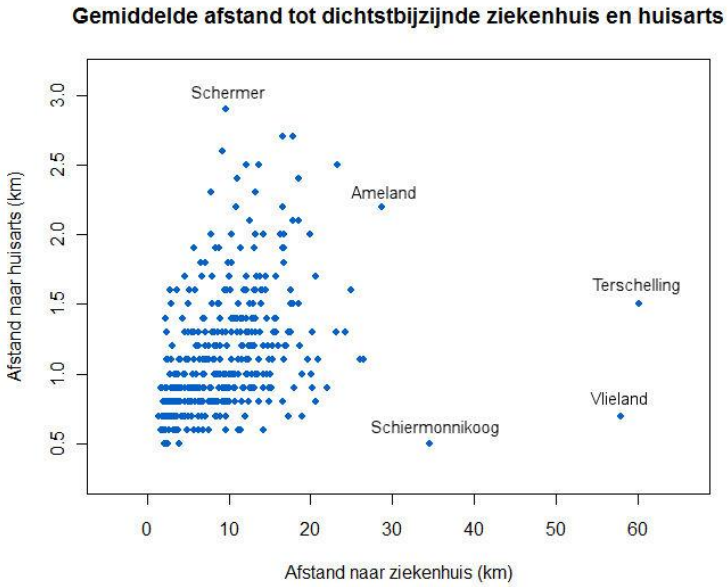
Figuur 8.1. Een puntenwolk



Er blijkt een grote groep gemeenten te zijn met een kleine afstand tot het ziekenhuis (minder dan 20 kilometer) en een kleine afstand tot de huisarts (minder dan 1,5 kilometer). Er zijn echter ook uitzonderingen, zoals de twee uitschieters aan de rechterkant van de grafiek. Het kan

informatief zijn op die punten te benoemen door er een tekst bij te zetten. In dit geval zou dat de naam van de gemeente kunnen zijn. In figuur 8.2 is dat gedaan.

Figuur 8.2. Een puntenwolk met benoemde uitschieters



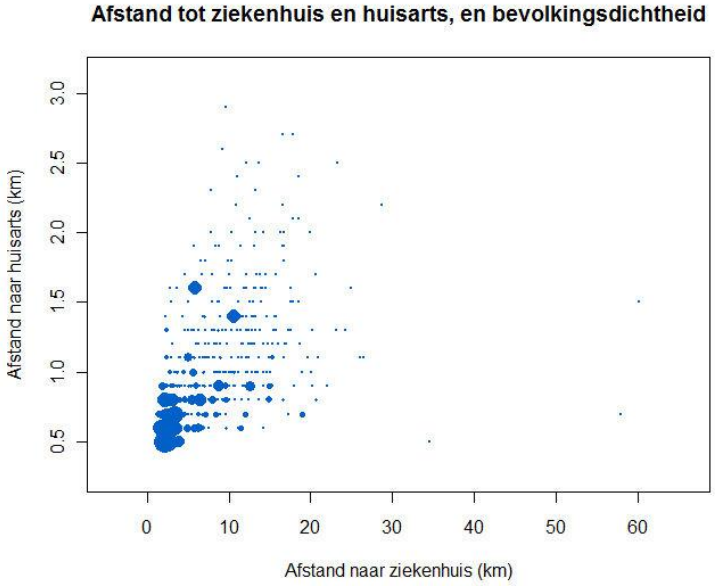
Uit figuur 8.2 wordt de speciale positie van de Waddeneilanden duidelijk. Het mag geen verbazing wekken dat de afstand naar het ziekenhuis daar groot is. De afstand naar de huisarts is niet bijzonder groot, want er wonen huisartsen op de eilanden. De gemeente Schermer in Noord-Holland is een uitschieter in de zin dat de gemeente niet ver van een ziekenhuis ligt, maar de gemiddelde afstand naar de huisarts is met 3 kilometer wel erg groot.

Bij de assen staan een beperkt aantal 'mooie' waarden vermeld. Merk op dat er marges zitten tussen de uiterste waarden en de rand van de grafiek. Daarmee voorkom je dat punten op die randen komen te liggen en daardoor minder goed zichtbaar zijn.

Het spreidingsdiagram is een krachtige techniek voor het weergeven van de samenhang tussen twee kwantitatieve variabelen. Er zijn allerlei manieren om de functionaliteit van dit type grafiek nog verder uit te breiden. Zie Bethlehem (1987) voor een overzicht. Hier beschrijven twee manieren om nog een derde variabele in de grafiek op te nemen.

Als de derde variabele een kwantitatieve variabele is, dan kun je de waarde weergegeven in de omvang van het punt. Een grote waarde leidt tot een grote cirkel en een kleine waarde tot een kleine cirkel. Figuur 8.3 toont daarvan een voorbeeld. De derde variabele is hier de bevolkingsdichtheid in de gemeente. Duidelijk is te zien dat de gemeenten met een hoge bevolkingsdichtheid zich linksonder in de grafiek bevinden. Kennelijk is de gemiddelde afstand tot ziekenhuis en huisarts daar klein.

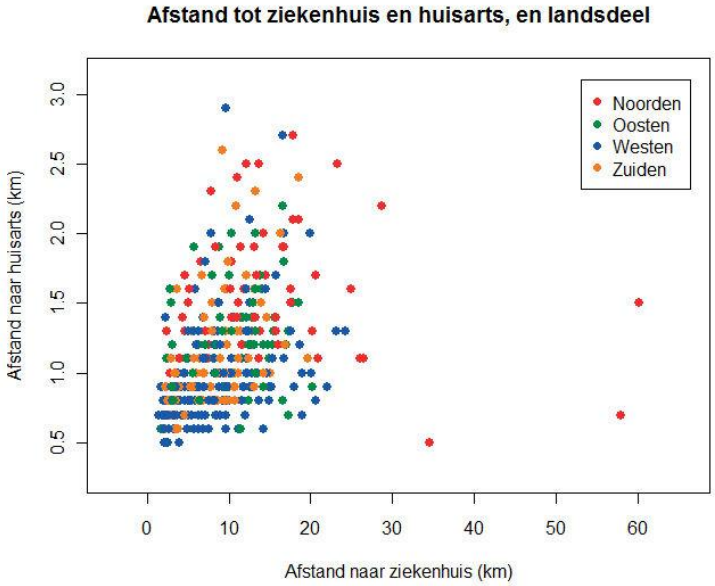
Figuur 8.3. Een puntenwolk met een derde (kwantitatieve) variabele



Een nadeel van de techniek in figuur 8.3 is dat grote cirkels kleine cirkels kunnen verbergen. Dat wordt soms wel *occlusie* of *overplotting* genoemd. Als dat veel gebeurt in een grafiek, kan dit de juiste interpretatie van de grafiek vertroebelen.

Het is ook mogelijk om als derde variabele een kwalitatieve variabele te gebruiken. Dat kan door de punten de kleur te geven van de overeenkomstige categorie. Als voorbeeld is in figuur 8.4 als derde variabele het landsdeel genomen waarin de gemeente ligt.

Figuur 8.4. Een puntenwolk met een derde (kwalitatieve) variabele



Uit de grafiek valt nu af te lezen dat gemeenten met een grote afstand tot het ziekenhuis zich in het noorden van het land bevinden. De blauwe stippen concentreren zich linksonder in de grafiek. Dat zijn de gemeenten in het westen van het land.

Samenvatting

- Als de variabele een omvang, gewicht, duur of waarde meet, laat de desbetreffende as dan bij bij 0 beginnen. Vermijd een knip in de as.
- Zet een beperkt aantal 'mooie' schaalwaarden bij beide assen.
- Neem een marge in acht tussen de buitenste punten en de randen van de grafiek.
- Gebruik bij een simpele puntenwolk van twee variabelen één kleur om de data van de metadata te onderscheiden.
- Overweeg om uitschieters in puntenwolk te benoemen.
- Roosterlijnen kunnen helpen bij het aflezen van waarden in de grafiek. Teken ze niet te opvallend.
- Vermijdt allerlei onnodige toeters en bellen ('chart junk').

9. Grafieken voor gemengde variabelen

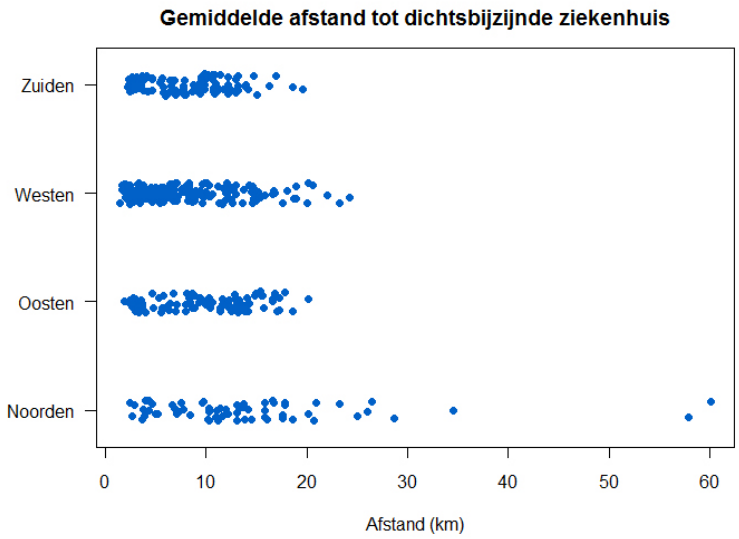
Voor de analyse van de samenhang tussen een kwantitatieve en een kwalitatieve variabele bestaan geen specifieke grafische technieken. Vaak wordt gebruik gemaakt van al eerder besproken technieken. Daarvoor kies je grafische techniek voor het weergeven van de verdeling van een kwantitatieve variabele en die techniek pas je vervolgens toe voor elke categorie van de kwalitatieve variabele. Hier beperken we ons tot twee typen grafieken: *spreidingsdiagrammen* en snorrendozen.

Spreidingsdiagrammen

Om de samenhang tussen een kwantitatieve variabele en een kwalitatieve variabele weer te geven, gaan we uit van het spreidingsdiagram zoals dat als is besproken in hoofdstuk 6. De uitbreiding is dat je in één grafiek de spreiding van de waarden voor elke categorie van de kwalitatieve variabele tekent.

Figuur 9.1 bevat een voorbeeld. De populatie bestaat uit alle gemeenten in Nederland (in 2010). Voor elke gemeente is de gemiddelde afstand (van de inwoners) tot het dichtstbijzijnde ziekenhuis bepaald. Dat is de kwantitatieve variabele. Als kwalitatieve variabele is het landsdeel gebruikt waarin de gemeente ligt.

Figuur 9.1. Spreidingsdiagrammen (met jitter)



De afstand tot het ziekenhuis is op de horizontale as afgezet en het landsdeel op de verticale as. Aan de punten is enige ruis (jitter) toegevoegd om te voorkomen dat ze over elkaar heen vallen.

De spreiding is in het noorden van het land duidelijk groter dan in de andere landsdelen. Ook hier zijn de al eerder genoemde uitschieters weer te zien. Het gaat om de Waddeneilanden Terschelling en Vlieland en in mindere mate Schiermonnikoog.

Ook hier kan het weer helpen bij de interpretatie om de uitschieters te benoemden door de namen van de gemeenten erbij te zetten.

Snorrendozen

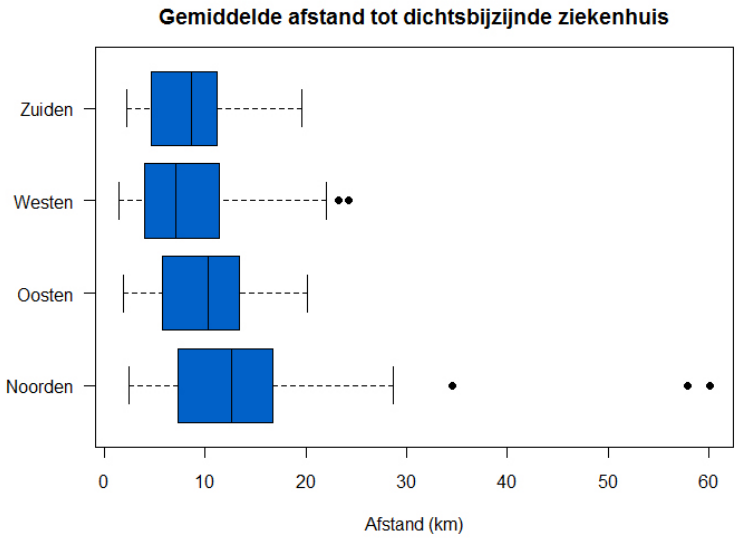
Een tweede grafische techniek die in aanmerking komt voor het weergeven van de samenhang tussen een kwalitatieve en een kwantitatieve variabele, is de snorrendoos. De snorrendoos is al

als techniek geïntroduceerd in hoofdstuk 6. Bij een snorrendoos is goed te zien welk waarden allemaal voorkomen, of de verdeling symmetrisch is en of er uitschieters zijn.

Voor het weergeven van de samenhang teken je in één grafiek voor elke categorie van de kwalitatieve variabele een snorrendoos voor de verdeling van de andere (kwantitatieve) variabele. Door al die snorrendozen in één grafiek te tekenen met dezelfde schaalverdeling, kun je de verdelingen onderling vergelijken.

Figuur 9.2 toont een voorbeeld. Dezelfde gegevens zijn gebruikt als voor het voorbeeld met de spreidingsdiagrammen. Voor elk landsdeel is er een snorrendoos voor de gemiddelde afstand tot het dichtstbijzijnde ziekenhuis.

Figuur 9.3. Snorrendozen



De afstanden tot het ziekenhuis zijn in het noorden wat groter dan in de andere landsdelen. In het oosten zijn de afstanden wat groter dan in het westen en zuiden. In het noorden zijn er drie uitschieters (Terschelling, Vlieland en Schiermonnikoog). Ook in het westen lijken er twee uitschieters te zijn: Korendijk en Wieringermeer. Het kan informatief zijn om de namen van de uitschieters in de grafiek te vermelden.

Samenvatting

- Teken de waarde-as van het spreidingsdiagram horizontaal.
- Als de variabele een omvang, gewicht, duur of waarde meet, laat de as van de spreidingsdiagrammen en de snorrendozen dan bij 0 beginnen. Vermijd een knip in de as.
- Geef alle punten van het spreidingsdiagram dezelfde kleur (of grijstint), tenzij je nog een derde (kwalitatieve) variabele wilt tonen. Geef ook alle snorrendozen dezelfde kleur.
- Roosterlijnen kunnen helpen bij het aflezen van waarden in de grafiek. Teken ze niet te opvallend.
- Overweeg om uitschieters te benoemen.
- Vermijdt allerlei onnodige toeters en bellen ('chart junk').
- Vermijd driedimensionaal perspectief.

10. Grafieken voor tijdreeksen

Onderzoek is er vaak opgericht om veranderingen in de tijd in kaart te brengen. Twee typen grafieken kunnen hiervoor worden gebruikt: het *lijndiagram* en het *staafdiagram*. Beide typen hebben hun voor- en nadelen. Je kunt lijndiagrammen en staafdiagrammen gebruiken voor het tonen van veranderingen in de tijd van zowel kwalitatieve als kwantitatieve variabelen.

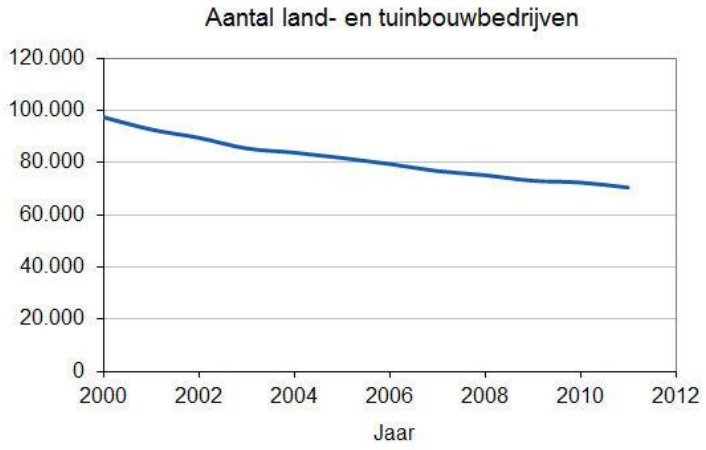
Als je verandering van een variabele in de tijd in beeld wilt brengen, dan gaat het in feite om de samenhang tussen twee variabelen: de desbetreffende variabele en de variabele tijd. Omdat je analyse vaak gericht zal zijn om de waarden van een variabele vanuit de tijd te verklaren, is het gebruikelijk om de tijd op de horizontale as te zetten en de variabele op de verticale as.

Uiteraard gelden voor beide assen de gebruikelijke richtlijnen, zoals een beperkt aantal mooie waarden bij de assen en het aangeven van een natuurlijk nulpunt als dat er is. Voor de tijd-variabele is er vaak geen natuurlijk nulpunt.

Een tijdreeks van een kwantitatieve variabele

De meest voor de hand liggende manier om een tijdreeks weer te geven is met een *lijndiagram*. De grafiek bevat voor elk tijdstip de waarde van de variabele. En die punten zijn met elkaar verbonden door een lijn. Figuur 10.1 bevat een voorbeeld. Het gaat hier om het in beeld brengen van de afname door de tijd heen van het aantal land- en tuinbouwbedrijven. De punten zijn in de grafiek niet zichtbaar. Desgewenst zou dat wel kunnen door de punten met een speciale symbool (cirkeltje, blokje) weer te geven, maar dit leidt af van het patroon dat in de lijn zit.

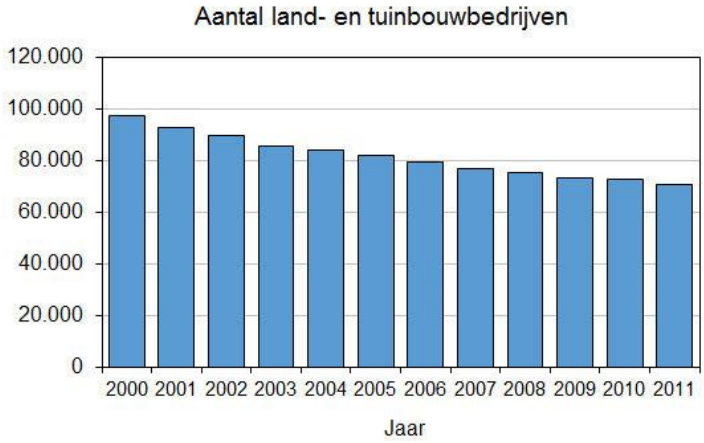
Figuur 10.1. Een lijndiagram



Door het verbinden van de punten door de lijn geef je impliciet aan dat de ontwikkelingen tussen twee tijdstippen geleidelijk verlopen.

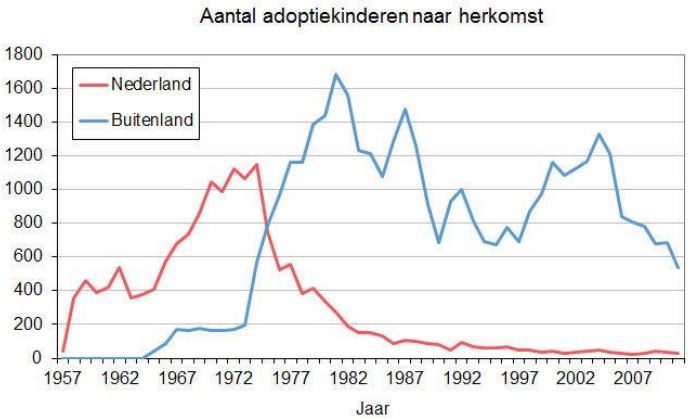
Als de variabele is gemeten op discrete tijdstippen, en dat moet worden benadrukt, dan kun je overwegen om het lijndiagram te vervangen door een *staafdiagram*, zoals in figuur 10.2. Dit staafdiagram is beter om de stand van zaken op een bepaald moment af te lezen, terwijl het lijndiagram zich vooral leent om juist de ontwikkelingen in de tijd te bestuderen.

Figuur 10.2. Een staafdiagram



Het is vrij simpel om in een lijndiagram verschillende variabele tegelijk tegen de tijd af te zetten. Als het gaat om een beperkt aantal variabelen, dan blijft interpretatie van de grafiek eenvoudig. In figuur 10.3 is het aantal adoptiekinderen door de tijd heen weergegeven. De grafiek bevat twee variabelen: aantal adoptiekinderen afkomstig uit Nederland zelf en het aantal adoptiekinderen dat uit het buitenland komt.

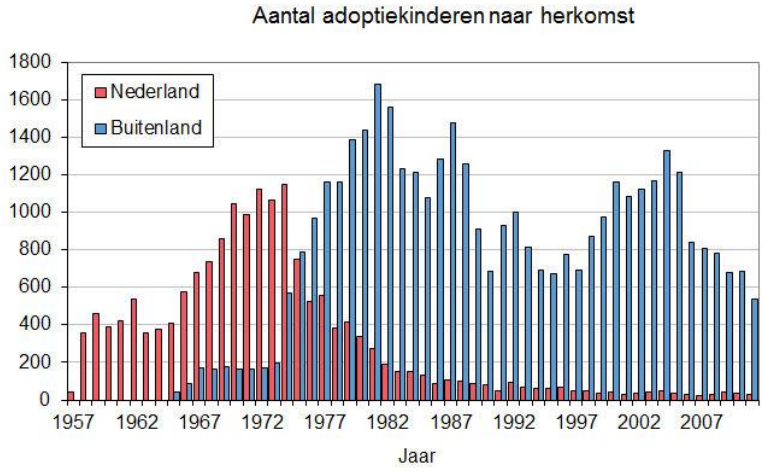
Figuur 10.3. Een lijndiagram met twee variabelen



Duidelijk is te zien hoe omstreeks 1974 het aantal adoptiekinderen uit Nederland sterk afneemt en het aantal adoptiekinderen uit het buitenland sterk toeneemt. Uit de grafiek valt niet goed af te lezen hoe het verloop van het totaal aantal adoptiekinderen in de tijd is.

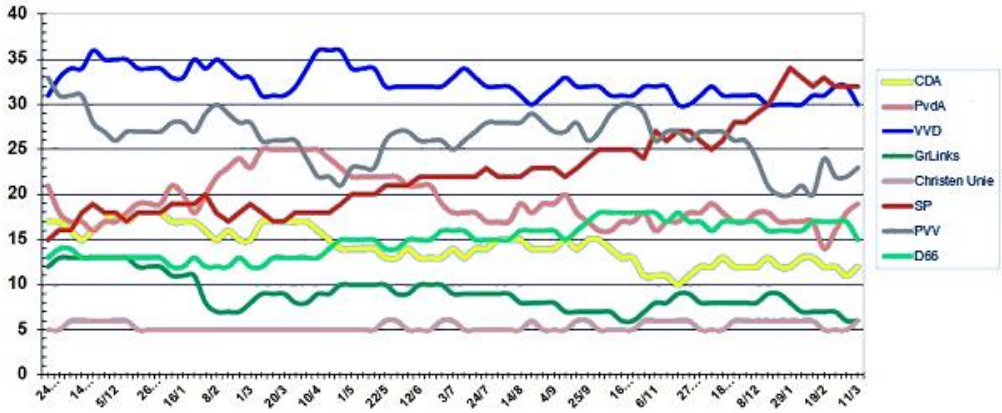
Figuur 10.4 laat zien wat er gebeurt als je de grafiek in figuur 10.3 vervangt door een staafdiagram. Het grote aantal staven in combinatie met meer dan één variabele maakt het lastiger de grafiek te interpreteren. Het lijkt zelfs alsof er nog een derde variabele in de grafiek staat (tussen 1967 en 1987).

Figuur 10.4. Een staafdiagram met twee variabelen



Het weergeven van twee variabelen in een lijndiagram, zoals in figuur 10.3, kan goed werken. Het is echter niet zo dat je willekeurig veel variabelen in een dergelijke grafiek kunt opnemen. Op een gegeven moment wordt de grafiek onleesbaar. Figuur 10.5 is afkomstig van de website *peil.nl* van Maurice de Hond. De grafiek toont de zetelvoorspellingen door de tijd heen van een flink aantal politieke partijen. Vooral als lijnen op ongeveer hetzelfde niveau liggen en elkaar kruisen, wordt de interpretatie steeds moeilijker. Het wordt ook steeds lastiger om voldoende onderscheidbare kleuren te vinden om de lijnen aan te duiden. Wat nog wat kan helpen is om in de legenda dezelfde volgorde aan te houden als bij de eindpunten van de lijnen aan de rechterkant van de grafiek.

Figuur 10.5. Een lijndiagram met veel variabelen

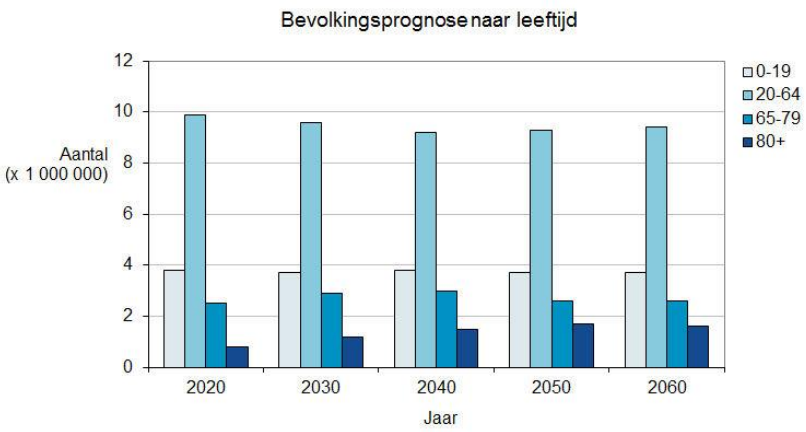


Een tijdreeks van een kwantitatieve variabele

Ook bij het volgen in de tijd van een kwalitatieve variabele kun je kiezen voor een staafdiagram of een lijndiagram. Als je kiest voor een staafdiagram, dan zijn er nog twee varianten: een samengesteld staafdiagram en een stapeldiagram.

Figuur 10.6 bevat een bevolkingsprognose. De kwalitatieve variabele is leeftijd in vier klassen. In deze grafiek is gebruik gemaakt van een samengesteld staafdiagram.

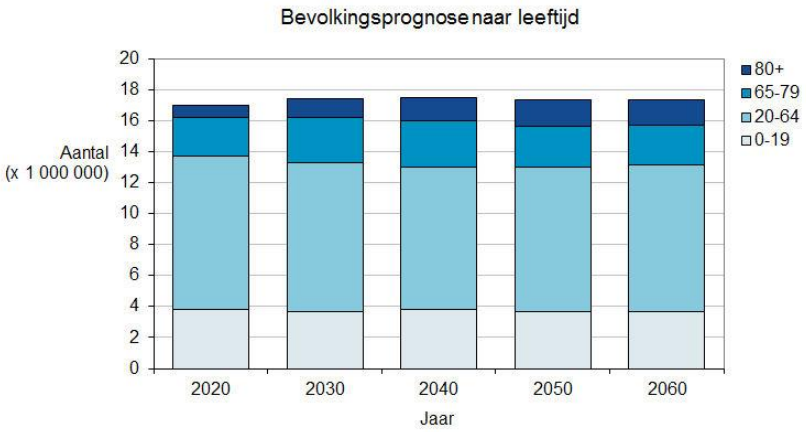
Figuur 10.6. Gebruik van een samengesteld staafdiagram



Bij dit type grafiek kun je de verandering in de tijd van de verschillende leeftijdscategorieën goed volgen. Zo is een toename te zien van het aantal personen van 80 jaar en ouder. En het aantal mensen tussen de 20 en 64 jaar lijkt af te nemen.

Uit figuur 10.6 valt niet af te lezen wat er gebeurt met de bevolking als geheel. Daarvoor is het beter om een stapeldiagram te gebruiken. Figuur 10.7 bevat de bevolkingsprognose als een stapeldiagram. Nu is te zien dat de bevolking eerst nog licht toeneemt en daarna afneemt. Het is echter lastiger om het verloop van de omvang van de verschillende leeftijdscategorieën in de tijd te volgen.

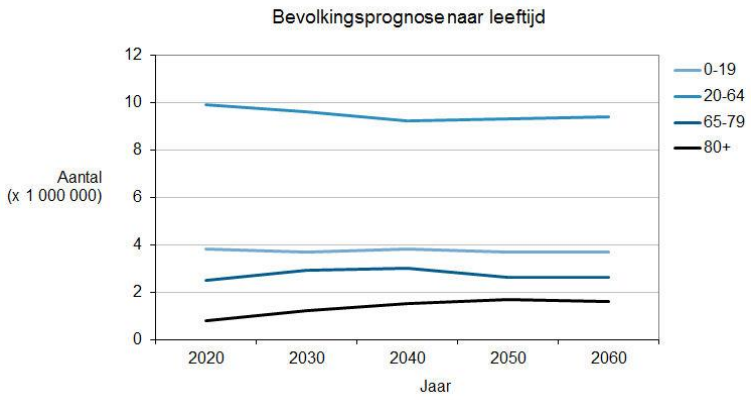
Figuur 10.7. Gebruik van een stapeldiagram



Merk op dat in figuren 10.6 en 10.7 de kwalitatieve variabele een ordinale variabele is. Er is sprake van een natuurlijke ordening. Daarom zijn verschillende tinten van één kleur gebruikt om de categorieën aan te duiden.

In figuur 10.8 is de ontwikkeling in de tijd van een kwalitatieve variabele weergegeven in de vorm van een lijndiagram. De ontwikkeling in de tijd is voor elke categorie apart goed te volgen. Lastiger is het om vast te stellen of de samenstelling van de leeftijdscategorieën verandert.

Figuur 10.8. Gebruik van een lijndiagram



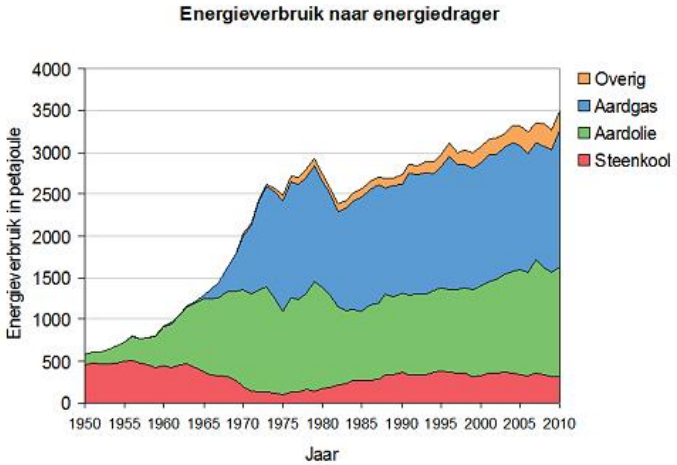
Voor een kwalitatieve variabele is het ook nog mogelijk een zogenaamd *oppervlaktediagram* te maken. Dit lijkt op een lijndiagram met meer variabelen, maar zit toch iets anders in elkaar.

Figuur 10.9 bevat een voorbeeld van een oppervlaktediagram. De kwalitatieve variabele is het energieverbruik naar energiedrager. Op een bepaald tijdstip wordt de omvang van een categorie niet bepaald door de afstand tot de horizontale as, maar door de 'dikte' van de laag. Zo is energieverbruik voor overige energiedragers erg klein.

De bovenkant van de bovenste laag geeft het totale energieverbruik weer. In die zin lijkt het oppervlaktediagram wel wat op het stapeldiagram. Dit type grafiek heeft toch wel wat interpretatieproblemen. Zo is eigenlijk alleen het energieverbruik van steenkool goed af te lezen. Voor de andere categorieën is dat een stuk lastiger, omdat zowel de onderkant als de bovenkant van de laag varieert.

Je moet erg oppassen bij het interpreteren van de pieken en dalen. Zo lijkt er een piek te zitten in het gebruik van aardgas tussen 1975 en 1980. Die piek wordt echter veroorzaakt door een piek in het aardoliegebruik in de laag eronder.

Figuur 10.9. Gebruik van een oppervlaktediagram



Een variant op figuur 10.9 is het gepercenteerde oppervlaktediagram. Daarbij wordt het energieverbruik in elk jaar op 100% en worden in de lagen de aandelen (in procenten) van de verschillende energiedragers weergegeven. Deze grafiek leidt aan dezelfde euvels als de grafiek figuur 10.8, zij het dat nu de trends in de omvang van de eerste en de laatste laag goed kunnen worden afgelezen. Voor de tussenliggende lagen blijft het lastig.

Samenvatting

- Teken de tijd-as horizontaal. Het nulpunt voor de tijd kun je weglaten.
- Als de variabele een omvang, gewicht, duur of waarde meet, laat de verticale as van het lijndiagram of staafdiagram dan bij 0 beginnen. Vermijd een knip in de as.
- Gebruik verschillende kleuren voor een nominale kwalitatieve variabele en verschillende tinten van dezelfde kleur voor een ordinale kwalitatieve variabele.
- Beperkt het aantal lijnen in een lijndiagram.
- Roosterlijnen kunnen helpen bij het aflezen van waarden in de grafiek. Teken ze niet te opvallend.
- Vermijdt allerlei onnodige toeters en bellen (‘chart junk’).
- Vermijd driedimensionaal perspectief.

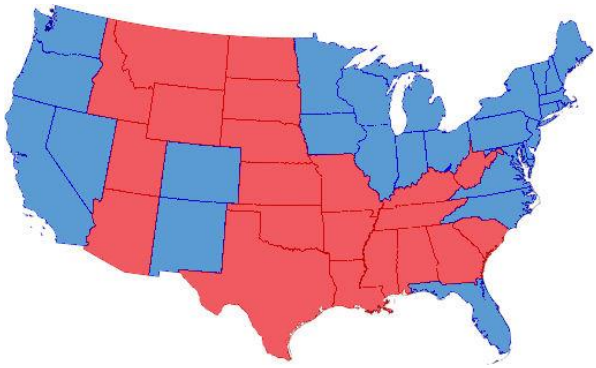
11. Thematische kaarten

Een thematische kaart is een speciaal type grafiek waarbij je de verdeling van een variabele over een gebied weergeeft. Daarmee breng je in feite een mogelijk verband tussen een variabele en de geografische locatie in beeld.

Thematische kaarten kunnen heel informatief zijn. Het in beeld brengen van de verbeelding moet echter wel zorgvuldig gebeuren, want het gevaar van verkeerde interpretatie ligt op de loer. Een specifiek probleem is de vertekening die wordt veroorzaakt door het feit dat niet alle gebieden dezelfde oppervlakte hebben. In het Engels wordt dit aangeduid met *area bias*.

Figuur 11.1 toont een voorbeeld van area bias. Deze thematische kaart brengt de uitslag van de presidentsverkiezingen in Amerika in 2008 in beeld. In de rode staten heeft de Republikein John McCain gewonnen en in de blauwe staten ging de winst naar de Democraat Barack Obama.

Figuur 11.1. Een voorbeeld van area bias (bron: PoliticalMaps.org)



Zo op het oog lijkt John McCain de winnaar van de verkiezingen. Er is immers veel meer rood dan blauw gebied. Obama was echter de echte winnaar, met 365 tegen 173 kiesmannen, en met 53% tegen 46% van de stemmen. De verklaring van de vertekening is dat het bij het rode gebied in het midden vooral om grote, dunbevolkte staten gaat. Dit grote oppervlak wekt onterecht de indruk dat er daar dan ook wel veel mensen gestemd zullen hebben. Dat is echter niet het geval. Vooral in het noordoosten wonen veel mensen dicht op elkaar in kleine staten (qua oppervlakte).

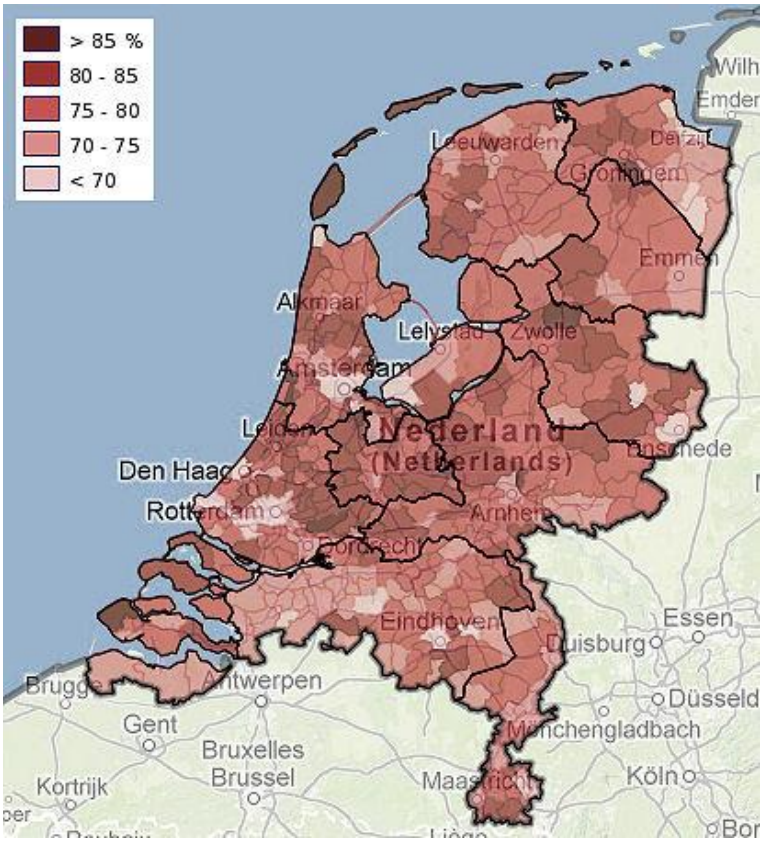
Kaarten waarbij je de gebieden inkleurt met kleuren of tinten die corresponderen met de categorieën van een variabele, noemen we *choropletische kaarten* (of kortweg: *choropleten*) genoemd. Het zijn vooral deze kaarten waarbij area bias kan optreden.

Kaarten met een kwantitatieve variabele

Voor het weergeven van een kwantitatieve variabele in een thematische kaart, zet je de variabele meestal om in een beperkt aantal categorieën. Omdat er dan sprake is van een natuurlijke ordening, gebruik je verschillende tinten van één kleur om de gebieden in te kleuren.

Als de kwantitatieve variabele een verschijnsel weergeeft in termen van relatieve cijfers (bijvoorbeeld een dichtheid of percentage) dat kan je hiervoor zonder problemen een choropleet maken. Figuur 11.2 toont een voorbeeld. De variabele is het opkomstpercentage bij de verkiezingen voor de Tweede Kamer in 2010. De kaart is afkomstig uit de NRC (2010). Duidelijk is bijvoorbeeld te zien dat de opkomst in de grote steden laag was, en op de Waddeneilanden hoog.

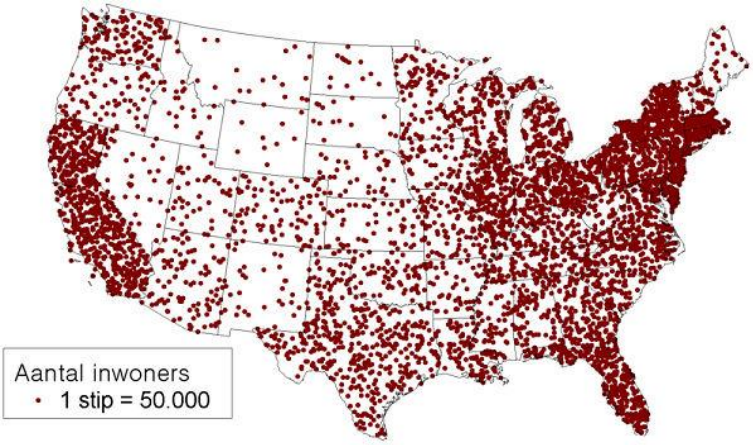
Figuur 11.2. Choropletische kaart met kwantitatieve variabele



Als het in de thematische kaart om absolute cijfers gaat (bijvoorbeeld aantal, grootte, omvang of waarde), dan kun je daarvoor beter niet een choropleet gebruiken, omdat dit al gauw aanleiding geeft tot area bias, vooral als de gebieden nogal in oppervlakte verschillen. Alternatief wordt geboden door de *dichtheidskaart* (*density map*) en de *symbolenkaart* (*symbol map*).

Essentieel voor een *dichtheidskaart* is dat de gebieden niet worden ingekleurd. In elk gebied wordt een aantal symbolen geplaatst. Het aantal is evenredig met de waarde van de variabele in het gebied. In figuur 11.3 zijn daarvoor stippen gebruikt.

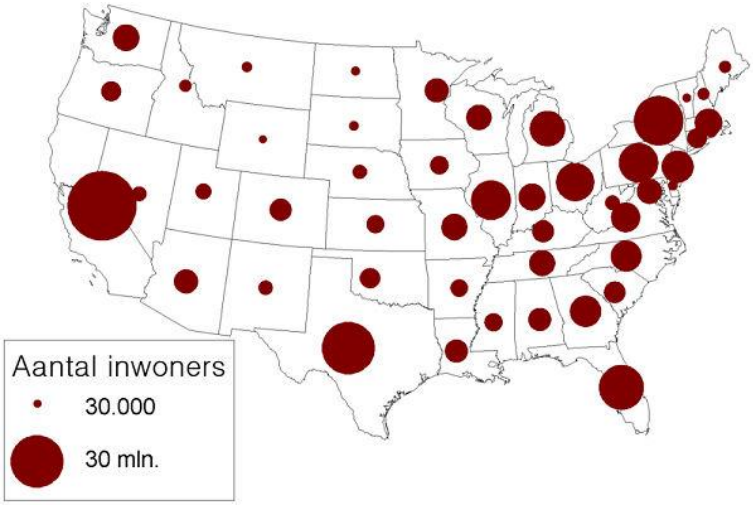
Figuur 11.3. Dichtheidskaart van de bevolking van Amerika
(bron: nl.wikipedia.org/wiki/Thematische_kaart)



Figuur 11.3 geeft de spreiding van de inwoners over Amerika weer. Elke stip staat voor 50.000 inwoners. Er vallen duidelijke patronen te onderscheiden. De staten in het midwesten zijn maar dunbevolkt. Er wonen echter heel veel mensen in het noordoosten, en in iets minder mate in Florida en Californië. De grenzen en oppervlaktes spelen nog maar een amper een rol bij de interpretatie van de grafiek. Ze zijn meer bedoeld ter oriëntatie.

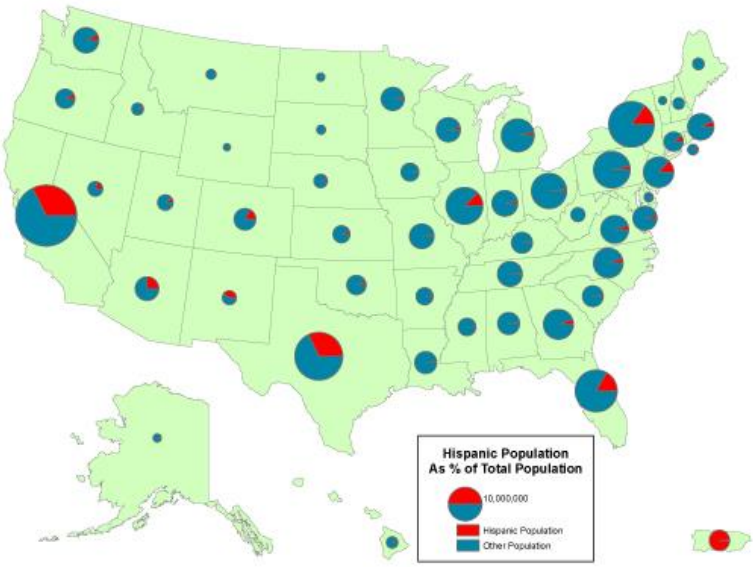
Een andere manier om de informatie weer te geven, is de *symbolenkaart* (*symbol map*). Per gebied teken je dan één symbool. Die symbolen variëren in grootte. Die grootte moet evenredig zijn met de waarde van de variabele.

Figuur 10.4. Symbolenkaart van de bevolking van Amerika
(bron: nl.wikipedia.org/wiki/Thematische_kaart)



Ook deze versie van de thematische kaart geeft de boodschap op juiste wijze weer: er wonen veel mensen in het noordoosten en het westen, en maar weinig mensen in het midwesten.

Figuur 11.5. Symbolenkaart met cirkeldiagrammen
(bron: Nature of Geographic Information, PennState, PA)



Uiteraard kun je ook andere symbolen dan stippen of cirkels gebruiken. Door in figuur 10.4 de cirkels te vervangen door cirkeldiagrammen van een andere variabele, kan die variabele worden meegenomen in de grafiek. Figuur 11.5 geeft daarvan een voorbeeld. De cirkels geven de omvang van de bevolking in de staten aan. De cirkels zijn cirkeldiagrammen voor de variabele

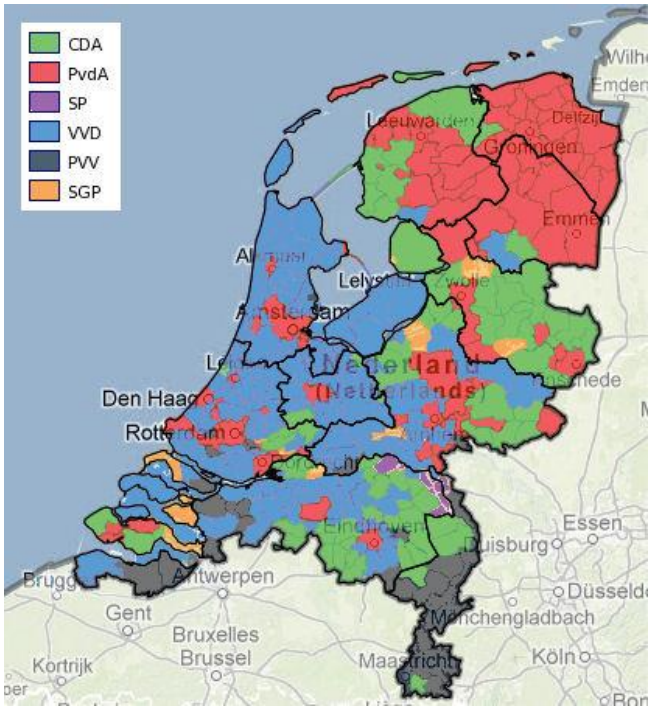
die weergeeft welke aandeel van de bevolking 'Hispanic' is, dus Spaans als moedertaal heeft. Nu kun je ook nog vaststellen dat het aandeel Hispanics vooral groot is in Californië, Nieuw-Mexico en Texas.

Kaarten met een kwantitatieve variabele

Het is ook mogelijk om een kwalitatieve variabele op een kaart weer te geven. Je vult de gebieden dan met kleuren die corresponderen met de categorieën. Bij voorkeur is er een beperkt aantal categorieën en dus een beperkt aantal kleuren. Zijn er heel veel verschillende kleuren, dan kan je categorieën niet meer uit elkaar houden in de grafiek. De perceptie van de intensiteit van de kleuren moet voor alle categorieën hetzelfde zijn.

Figuur 11.5 toont een voorbeeld. De variabele die in beeld is gebracht, is geeft aan welke partij bij de verkiezingen in 2010 de grootste is geworden. De kaart is afkomstig uit de NRC (2010).

Figuur 11.5. Choropletische kaart met kwantitatieve variabele



Dit voorbeeld toont nog eens de kracht van een thematische kaart. Vrijwel in één oogopslag wordt duidelijk dat de PvdA de grootste partij is in het noordoosten en het CDA in Overijssel. De PVV is vooral sterk in Limburg. De VVD is de grootste partij in het westen en midden van het land.

12. Literatuur

Bethlehem, J.G. (1987), *Het spreidingsdiagram opnieuw bekeken*. CBS-rapport, Centraal Bureau voor de Statistiek, Voorburg.

Chambers, J.M., Cleveland, W.S., Tukey, P.A. & Kleiner, B. (1983), *Graphical Methods for Data Analysis*. Chapman & Hall, New York.

Escher, B.G. (1924), De methodes der grafische voorstelling. Wereldbibliotheek, Amsterdam

Few, S. (2004), *Show Me the Numbers: Designing Tables and Graphs to Enlighten*. Analytics Press, Burlingame, CA.

Funkhouser, H.G. (1936), A Note on a Tenth Century Graph. *Osiris* 1, blz 260-262.

Holmes, N. (1984), *Designer's Guide to Creating Charts and Diagrams*, Watson-Guptill Publications, New York.

Mercator, G. & Hondius, J. (1607), Designatio Orbis Christiani. Amsterdam.

NRC (2010), *De Verkiezingskaart, uitslagen in een oogopslag*. www.nrc.nl/nrclab/2010/06/11/de-verkiezingskaart-uitslagen-in-een-oogopslag.

Playfair, J. (1786), *The Commercial and Political Atlas*. London.

Schmid, C.F. (1983), *Statistical Graphics, Design Principles and Practices*. John Wiley & Sons, New York.

Scott, D.W. (1979), On Optimal and Data-based Intervals. *Biometrika* 66, pp. 605-610.

Snijkers, G.J.M.E. (1985), *De vormgeving van grafieken met grafisch-statistische programmatuur*. Rapport 12345-88-M1-1, Centraal Bureau voor de Statistiek, Voorburg.

Sturges, H.A. (1926), The Choice of a Class Interval. *Journal of the American Statistical Association* 21, pp.65-66.

Talbot, J., Lin, S. & Hanrahan, P. (2010), An Extension of Wilkinson's Algorithm for Positioning Tick Labels on Axes. *IEEE Transactions on Visualization and Computer Graphics* 16, pp. 1036-1043.

Tukey, J.W. (1977), *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

Tufte, E. (1983): *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, Connecticut.

United Nations (2009), *Making Data Meaningful, Part 2: A Guide to Presenting Statistics*. United Nations Economic Commission for Europe, Geneva.

Wainer, H. (1997): Visual revelations: graphical tales of Fate and Deception from Napoleon Bonaparte to Ross Perot. Copernicus / Springer Verlag, New York.

Wallgren, A., Wallgrens, B., Persson, R., Jorner, U. and Haaland, J-A. (1996), *Graphing Statistics and Data, Creating Better Charts*. Sage Publications, London.

Zeileis, A., Hornik, K. & Murrell, P. (2009), *Escaping RGBland: Selecting Colors for Statistical Graphics*. Computational Statistics & Data Analysis 53, pp. 3259-3270.