PUBLIEKE DATA INLEZEN

Inleiding

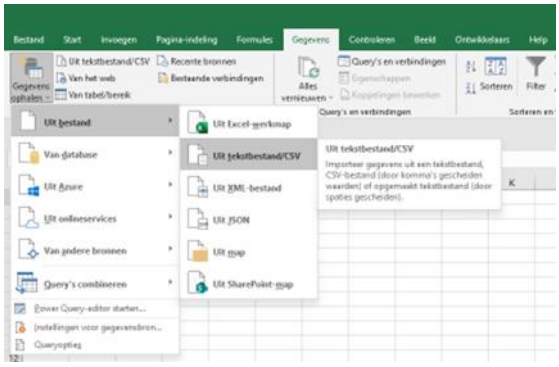
Veel organisaties stellen hun data beschikbaar voor het publiek via een API (*Application Programming Interface*) of download. Deze data heeft niet altijd het door jou gewenste formaat (denk aan eenheden, verschil in notatie), is niet altijd compleet, bevat soms overtollige informatie of veel meer gegevens of records dan waar je eigenlijk naar op zoek was.
In dit computerpracticum leer je aan de hand van een voorbeeld hoe je controle krijgt over de data. Deze opdracht is geen *kookboek*: niet alle stappen staan volledig beschreven. Je moet het deels zelf ontdekken!

Een dataset downloaden en voorbewerken in Excel

De manier waarop data wordt aangeboden verschilt erg per site. Het kan zijn dat je gegevens moet kopiëren van een pagina om ze vervolgens in een programma te plakken, maar het komt ook voor dat de data in één of meerdere formaten (xml, csv, xlsx, txt, json, etc.) kan worden gedownload.

Uit PISA-onderzoek (*Programme for International Student Assessment*; gericht op wiskunde, natuurwetenschap en lezen) blijkt dat de vaardigheid van Nederlandse leerlingen achteruit gaat (ook t.o.v. andere landen). Ook blijven jongens qua prestaties achter bij meiden. Hoe zit dat precies? Om dat uit te zoeken nemen we PISA als casus. Klik voor de databron op deze link.

1. Bekijk de pagina behorende bij de link. Klik rechtsonder bij de grafiek op het download-icoon en selecteer onder *Data* **Download full data**. sla dit zip-bestand op je OneDrive en pak het zip-bestand uit. We werken met *pisa-mean-performance-on-the-reading-scale-by-sex.csv*.
2. Open het bestand in *Kladblok* of *Notepad++* (dus niet in Excel). Welk scheidingsteken is gebruikt?
3. Open Excel en navigeer naar *Gegevens* en kies links voor *Gegevens ophalen* ➔ *Uit bestand* ➔ *Uit tekstbestand/CSV* en kies het door jou gedownloade csv-bestand. Zie eventueel de figuur hiernaast. Na de selectie volgt een previewscherm. Kies voor *Gegevens transformeren*.



Je hebt nu de *Power Query-editor* bereikt. Hier kunnen we de data bewerken en selecteren. Voor getalwaarden is een algemeen aandachtspunt dat wij kommagetallen hebben, waar men internationaal een punt gebruikt. Dit kan problemen opleveren als je ze in Excel wil bewerken. In sommige datasets is de komma het scheidingsteken!

In deze dataset zien we verder dat er voor de meeste kalenderjaren helemaal geen PISA-gegevens zijn. Er zijn hier twee datasets gecombineerd, waarbij er voor heel veel jaren wel een populatie bekend is.

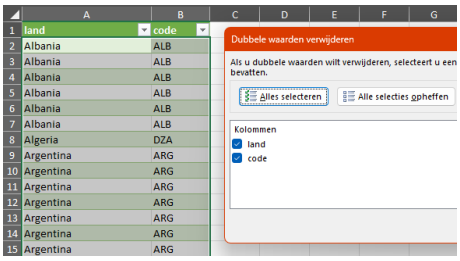
4. Klik in de *Power Query-editor* op ▼ achter de kolomnaam met resultaten voor *Female* en verwijder alle lege kolommen uit het resultaat.
5. Selecteer zowel de kolom met *Female* als met *Male* en navigeer naar *Waarden vervangen*. Vervang nu alle punten door komma's, om de Nederlandse notatie voor kommagetallen te krijgen. Klik vervolgens met je rechtermuisknop op de geselecteerde kolomnamen en kies voor *Type wijzigen* en daarna voor *Decimaal getal*.
6. Sluit de editor linksboven met *Sluiten en laden* om de dataset in Excel-dataformaat om te zetten.
7. Vervang de attribuutnamen door de korte Nederlandse versies *land*, *code*, *jaar*, *vrouw*, *man*, *inwoners* en *continent*.
8. Probeer het filter: selecteer bij *code* alleen *NLD* en sorteer de data aflopend op *jaar*. Als het goed is krijg je nevenstaande tabel.
9. Sla het bestand op met de naam *PISA_data.xlsx*.

land	code	jaar	vrouw	man	inwoners	continent
Netherlands	NLD	2015	514,7034556	491,1367106	16938492	Europe
Netherlands	NLD	2012	524,7607757	498,3174	16791850	
Netherlands	NLD	2009	520,5085628	496,1682434	16626379	
Netherlands	NLD	2006	519,0455834	494,8946254	16440091	
Netherlands	NLD	2003	523,7781508	502,8717855	16200948	

Data splitsen in tabellen

De data die we hebben gedownload is een mix van datasets. Op dit moment is het niet een correcte database. Er is sprake van redundante data (de combinatie *Netherlands* en *NLD* of algemeen *land* + *code*) en van lege velden voor het attribuut *continent*. Dat is niet onlogisch: in een goede database is het continent eenmalig gekoppeld aan het land (en de unieke landcode). Daarom voeren we correcties door.

10. Je hebt filters aangezet in je document. Zorg dat weer alle landcodes worden weergegeven.
11. Klik met je rechtermuisknop op het tabblad en maak een kopie. Geef deze de naam *scores*.
12. Verwijder op het tabblad *scores* de kolommen *land* & *continent*.
13. Ga terug naar het eerste tabblad en geef deze de naam *landen*. Verwijder alle kolommen m.u.v. *land* en *code*.
14. Ga naar *Gegevens* ➔ *Hulpmiddelen voor gegevens* en kies  (*Dubbele waarden verwijderen*). Zorg dat beide kolommen geselecteerd zijn en verwijder vervolgens alle duplicaten.
15. Verwijder tenslotte rijen zonder een waarde in de kolom *code*. Voor **zowel** het tabblad *landen* als het tabblad *scores*.
16. Sla het bestand op (als xlsx-bestand).



Data inladen in PHPMyAdmin en bewerken met SQL

Bij het computerpracticum over PHPMyAdmin heb je geleerd hoe je een csv-bestand kunt importeren. Deze kun je verkrijgen door (per tabblad) via *Bestand* ➔ *Opslaan als* te kiezen voor de extensie *csv*. Bij deze opdracht kiezen we een manier om in één handeling beide tabellen in te lezen via het ODS-formaat (*Open Document Spreadheet*).

17. Sla een kopie van jouw xlsx-bestand via *Bestand* ➔ *Opslaan als* op als *OpenDocument-werkblad*.
18. Maak in PHPMyAdmin een nieuwe lege database aan met de naam *Pisa*.
19. Klik daarna op *Importeren*, selecteer jouw ods-bestand en vink aan dat de eerste regel van de tabel de kolomnamen bevat. Laat de overige *settings* zoals ze zijn. Klik als laatste op *Starten*.
20. Bekijk de (inhoud van de) tabellen. Ze hebben een afwijkende opmaak (t.o.v. ons vorige computerpracticum).
21. Voeg primaire sleutels toe: *code* voor de landentabel en de combinatie *code+jaar* voor de scores.
22. Voeg tot slot een referentiesleutel toe tussen de tabellen *landen* en *scores*.

Query's gebruiken voor je onderzoeksvraag

We hebben nu een nette database met PISA-gegevens. Een aantal gegevens zijn hier rechtstreeks uit te halen door records te lezen, maar gebruikmakend van SQL kun je ook nieuwe informatie genereren. Met:

```
SELECT land,jaar,man,vrouw, round(vrouw-man) AS verschil
FROM landen,scores WHERE landen.code = scores.code AND landen.code = 'NLD' ORDER BY jaar DESC
```

filter je snel op de data van Nederland, maar kun je ook iets zeggen over (de ontwikkeling van) het verschil in prestaties tussen jongens en meisjes.

23. Probeer deze query uit. Wat gaat er mis?

24. Bedenk een onderzoeksvraag die beantwoord kan worden met deze database.
25. Maak een query (of meerdere query's) om jouw eigen onderzoeksvraag te beantwoorden.