OpenRefine

Handleiding

Afbeelding met logo, Graphics, Lettertype, tekst

Automatisch gegenereerde beschrijving

*Studenten:*

Petra Forgó - 1798544

Marjolein Wildeboer - 1854070

Pleuni van Dijk - 1854678

Ruben Overbeek - 1705834

Inhoud

[Inleiding 2](#_Toc149322338)

[Leerdoelen 3](#_Toc149322339)

[Tutorialdag 3](#_Toc149322340)

[Introductie tot de tool 4](#_Toc149322341)

[Omschrijving en totstandkoming 4](#_Toc149322342)

[Kernfuncties OpenRefine 5](#_Toc149322343)

[Excel vs OpenRefine 8](#_Toc149322344)

[Toepasbaarheid in het werkveld 9](#_Toc149322345)

[Handleiding 11](#_Toc149322346)

[Facets 11](#_Toc149322347)

[Edit cells 13](#_Toc149322348)

[Edit columns 14](#_Toc149322349)

[Reconcile 15](#_Toc149322350)

[Script 17](#_Toc149322351)

[Verdieping/verbreding 18](#_Toc149322352)

[Voorbeeld: Gebruik van GREL-codeertaal 18](#_Toc149322353)

[Bronnenlijst 19](#_Toc149322354)

[Bijlagen 19](#_Toc149322355)

[Bijlage 1: Mail voorafgaand aan de tutorialdag 19](#_Toc149322356)

[Bijlage 2: Dataset bij handleiding 19](#_Toc149322357)

[Bijlage 3: Afbeeldingen bij handleiding 20](#_Toc149322358)

[Bijlage 4: Afbeelding bij verdieping/verbreding 28](#_Toc149322359)

Inleiding

OpenRefine is een softwaretool die te gebruiken is op een browser. Dit maakt het erg makkelijk om te downloaden. OpenRefine wordt voornamelijk gebruikt om ruwe onbruikbare data op te schonen en te transformen.

In deze handleiding worden de leerdoelen uiteengezet. Daarna wordt kort verteld over het ontstaan van OpenRefine. Verder krijgt u uitleg over de belangrijkste (basis)functies van OpenRefine.

# Leerdoelen

Hieronder vindt u de leerdoelen van deze handleiding. Dit is een bondige opsomming van kennis en vaardigheden die u als gebruiker na het doorlopen van deze handleiding redelijkerwijs zal bezitten.

* Na het doorlopen van deze handleiding weet de gebruiker in hoofdlijnen wat OpenRefine inhoudt.
* Na het doorlopen van deze handleiding weet de gebruiker wat het verschil is tussen Excel en OpenRefine. Er kan met inhoudelijke kennis van de tool beslist worden in welke situatie de voorkeur uitgaat naar het gebruik van OpenRefine ten opzichte van Excel en andersom.
* Na het doorlopen van deze handleiding weet de gebruiker hoe hij/zij om moet gaan met de basisfuncties van OpenRefine en weet de gebruiker hoe hij/zij de interface kan besturen.
* De gebruiker weet na het doorlopen van deze handleiding wat de kernfuncties van OpenRefine zijn en wat deze uitzonderlijk maakt.
* De gebruiker weet na het doorlopen van deze handleiding wat messy data is, en hoe deze opgeschoond kan worden tot een werkbaar bestand, door middel van OpenRefine.

# Tutorialdag

Om kennis te maken met OpenRefine, organiseren wij een workshop waarbij geïnteresseerden de eerste stappen van het programma zullen ontdekken. Er wordt een presentatie en een demo gegeven waarna er eventueel ruimte is voor een opdracht.

Voorafgaand aan de workshop is een e-mail verzonden naar alle deelnemers met instructies hoe het programma gedownload kan worden (zie bijlage). Daarnaast zijn de datasets voor zowel de demo als de opdracht aan een openbare map toegevoegd in Teams (waar alle deelnemers toegang tot hebben).

Eerst zal er een presentatie plaatsvinden van maximaal 10 minuten. Hierin wordt verteld over het ontstaan en de belangrijkste kenmerken van OpenRefine. Verschillen en voor- en nadelen tussen Excel en OpenRefine worden ook toegelicht. Naderhand is er ruimte voor vragen om eventuele onduidelijkheden op te helderen.

Na de introductie word je als deelnemer meegenomen in een demo van het programma. Hier zal je eerste kennismaking zijn met de interface en krijg je stap voor stap wat handige tips mee, inclusief de werking van de kernfuncties.

Na de demo start het workshopdeel. Hier krijgen de deelnemers van de workshop de kans om zelf een eerste kennismaking te hebben met het programma; je krijgt vragen waar je een antwoord op moet geven door middel van het gebruik van verschillende functies (hier was echter bij deze workshop geen tijd meer voor).

Introductie tot de tool

Dit hoofdstuk biedt een inleiding tot OpenRefine, inclusief zijn ontstaansgeschiedenis, kernfuncties en het belang van deze tool in gegevensbeheer en -analyse. Van de oorsprong als een intern project bij Metaweb (bedrijf) tot de open-source gemeenschap die zich eromheen heeft gevormd. OpenRefine heeft een lange weg afgelegd om uit te groeien tot een essentiële hulpbron voor professionals die met gegevens werken. We zullen ook de belangrijkste functies van OpenRefine nader toelichten.

## Omschrijving en totstandkoming

OpenRefine, voorheen bekend als Google Refine, is een krachtige open-source softwaretool die wordt gebruikt voor het schoonmaken en transformeren van ruwe, ongestructureerde gegevens in een bruikbare en gestructureerde vorm. Het is ontworpen om te helpen bij het opruimen, normaliseren en verrijken van gegevens. Het is vooral handig voor data cleaning, data preprocessing en datawranglingtaken (OpenRefine History, 2013).

OpenRefine is ontstaan als een Google-project en werd voor het eerst uitgebracht door Google in 2010. Het werd oorspronkelijk ontwikkeld door David Huynh als een intern project bij Metaweb; een bedrijf dat later werd overgenomen door Google. Het werd aanvankelijk uitgebracht onder de naam Google Refine. De ontwikkelaars wilden een tool creëren waarmee gebruikers eenvoudig grote hoeveelheden ongestructureerde gegevens konden analyseren en transformeren. Het project had als doel om data-cleaning en data-preprocessing toegankelijker te maken voor een breder publiek; van datawetenschappers tot journalisten en andere professionals die met gegevens werken.

In 2012 werd het project opengesteld als open-source software onder de naam OpenRefine, waardoor de bredere gemeenschap van ontwikkelaars kon bijdragen aan de verdere ontwikkeling en verbetering van de tool. Sindsdien heeft OpenRefine een actieve gemeenschap van gebruikers en ontwikkelaars aangetrokken en is het uitgegroeid tot een belangrijk instrument voor gegevensbeheer en -analyse.

## Kernfuncties OpenRefine

OpenRefine bevat veel verschillende functies die gebruikt kunnen worden om data op te schonen en te verkennen. De meest gebruikte functies van OpenRefine zijn gemakkelijk te begrijpen en aan te leren, wanneer deze functies uitgelegd worden door middel van een handleiding of tutorial.

Het programma is snel en gratis te downloaden en werkt op een websurfer van de computer. Hierdoor is het erg toegankelijk. Om alle functies van OpenRefine te kunnen begrijpen en toe te kunnen passen op een dataset, is er veel verdieping in de sofware nodig (OpenRefine, 2022).

OpenRefine heeft een aantal kernfuncties die belangrijk zijn voor het gebruik van de sofware. Om te begrijpen wat er met de verschillende functies uitgevoerd kan worden, zijn deze onderstaand per functie beschreven.

**Facets**

Met de functie ‘facetten’ kan er gemakkelijk gekeken worden naar patronen en trends in datasets. Door middel van de ‘tekst facet’ kan gekeken worden naar de hoeveelheid overeenkomstige woorden. Deze kunnen ook gemakkelijk gesorteerd worden op hoeveelheid of op alfabetische naam. Door middel van de functie ‘numeriek facet’ worden de getallen gesorteerd op het bereik wat gemakkelijk aangepast kan worden. Dit wordt afgebeeld met een histogram.

**Tekst filter**

Door middel van een ‘tekst filter’ kan er gezocht worden naar een bepaalde tekst in de gekozen kolom. Bij de functie tekst filter is tevens de optie ‘omkeren’. Hiermee worden alle waarden/woorden getoond behalve het gezochte woord.

**Clustering (clusteren)**

De definitie van clustering luidt: "Het vinden van groepen met verschillende waarden die alternatieve representaties van hetzelfde kunnen zijn."

Door gebruik te maken van de functie ‘clusteren’ kunnen (typ)fouten en inconsistenties gemakkelijk opgespoord en verbeterd worden. Dit is erg gunstig voor het opschonen van een dataset.

Bij het clusteren heb je twee verschillende methodes; ‘belangrijke botsing’ en ‘dichtstbijzijnde buurman’ (deze worden in onderstaande alinea’s nader toegelicht). Verder zijn er verschillende key-functies die gekozen kunnen worden om de verschillende soorten ‘algoritmische zoekfuncties’ aan te duiden.

De clustermethode belangrijke botsing is een snelle cluster methode; deze is token-gebasseerd.

De meest voorkomende key-functie die daarbij gebruikt wordt is ‘vingerafdruk’. Dit algoritme genereert ‘matches’ op basis van tekenreekswaarde. Het herstelt overige witruimte en spaties, hoofdletters en kleine letters, het negeert interpunctie, het splitst alle woorden op en sorteert ze op alfabetisch volgorde.

Een volgende key-functie is ‘N-gram vingerafdruk’. Deze key-functie kan helpen bij het matchen van cellen met typefouten of onjuiste spaties, bijvoorbeeld bij het matchen van ‘uitkijk’ en 'kijk uit’. De key-functie werkt door verschillende n-waardes in te stellen. Als een n-waarde van één gram wordt ingesteld, zullen alle letters in de cel in alfabetische volgorde geordend worden door segmenten van één letter/teken lang te maken. Als er een n-waarde van twee gram wordt ingesteld, dan bevatten alle segmenten twee tekens. Het algoritme zet deze segmenten op alfabetische volgorde, verwijdert duplicaten en voegt deze daarna weer samen. Hoe hoger dus de n-waarde is, hoe minder clusters er kunnen worden geïdentificeerd.

De volgende vier key-functies zijn gebaseerd op verschillende fonetische algoritmes. Dit betekent dat letters worden geïdentificeerd die hetzelfde klinken als ze hardop worden uitgesproken. Ook wordt er gekeken naar foutieve tekstwaarden; bijvoorbeeld een 's' die foutief is getypt in plaats van een 'z’.

De key-functie ‘metafoon 3’ is gebaseerd op een Engelstalig fonetisch algoritme. De ‘Keulen fonetisch’ is gebaseerd op de Duitse uitspraak van woorden. Key-functie ‘Daitch-Mokotoff’ is een algoritme voor Slavische en Jiddische woorden. Als laatste is de key-functie ‘Baider-Morse’. Deze bevat een algoritme dat goed namen kan matchen die overeenkomen met de verwachte naam.

De tweede clustermethode, waarvoor gekozen kan worden, is de dichtstbijzijnde buurman. Deze clustermethode is langzamer dan de belangrijke botsing en is teken-gebasseerd. Hiermee wordt bedoeld dat er gesorteerd wordt op een mate van gelijkenis van een teken. De gekozen waarde geeft de gelijkenis in verschillende woorden aan, zoals bij “New York” en “newyork”

De eerste key-functie die hierbij gebruikt kan worden is ‘Levenshtein’. Deze functie telt het aantal bewerkingen dat nodig is om de waardes precies overeen te laten komen.

Een andere key-functie die hierbij gebruikt kan worden is ‘PPM’ (prediciton by partial matching). Deze functie gebruikt compressie om te zien of waarden verschillend of vergelijkbaar zijn.

**Undo/redo (ongedaan maken/opnieuw)**

OpenRefine slaat de gemaakte wijzigingen van gegevens op. Deze zijn terug te vinden onder het tabblad ‘ongedaan maken/opnieuw’. De geschiedenis van de gemaakte stappen zijn hier zichtbaar. Ook kan er gemakkelijk teruggegaan worden naar een vorige stap. Als er dan een nieuwe bewerking gemaakt wordt, worden de latere acties (alles wat grijs is) gewist.

**Sorting (sorteren)**

Met de functie ‘sorting’ kunnen kolommen gesorteerd worden op basis van verschillende gegevenstypes zoals, alfabetische volgorde of omgekeerd, getallen naar grootste of kleinste, datums op vroegste of laatste datum of booleaanse waarden met false of true.

Daarnaast kan bij het sorteren de volgorde van de waarden, lege cellen en fouten in de sortingen bepaald worden. Als het sorteren is toegepast, kan deze permanent gemaakt of weer aangepast worden.

**Edit cells (cellen bewerken)**

Er zijn in OpenRefine verschillende manieren om cellen te bewerken. Een van deze manieren is het handmatig per cel aanpassen van de waarde. Een ander manier is het gebruiken van de ‘transform’ functie. Hiermee kan je onder andere tekst omzetten naar cijfers.

**Edit columns (kolom bewerken)**

Door middel van de optie ‘kolom bewerken’ kunnen er gemakkelijk nieuwe kolommen toegevoegd, samengevoegd, verplaatst, opgesplitst en verwijderd worden.

Bij het opsplitsen van kolommen kan er gekozen worden om te splitsen door middel van een scheidingsteken of op basis van de lengte.

**Reconcile (verzoening)**

‘Reconcile’ is een functie van OpenRefine waarbij de dataset vergeleken wordt met een externe bron. Dit kan erg handig zijn om snel spelling te corrigeren en gegevens te koppelen aan een bestaande dataset. De dataset die standaard in OpenRefine zit is Wikidata. Het is ook mogelijk om een eigen dataset te koppelen/toe te voegen.

**Script**

In OpenRefine kan de bewerkingsgeschiedenis geëxtraheerd worden. Dit betekent dat alle gemaakte stappen die iets hebben veranderd aan de dataset, genoteerd worden. Deze opgeschreven code kan opgeslagen en toegepast worden op een nieuwe dataset. Hierdoor wordt in de toekomst tijd bespaard bij het opschonen van soortgelijke datasets.

## Excel vs OpenRefine

Excel en OpenRefine hebben vergelijkbare functies en wegen om hetzelfde doel te bereiken. In het onderstaande deel zal duidelijk worden wat de voordelen en vergelijkingen zijn van OpenRefine ten opzichte van Excel.

**Stappen zijn makkelijk terug te vinden en te bewerken**

OpenRefine biedt een gedetailleerde geschiedenis van alle bewerkingen die zijn uitgevoerd op gegevens en kunnen deze tevens bewerken. In Excel is het ook mogelijk om stappen terug te nemen, maar de geschiedenis van bewerkingen zijn in Excel niet zichtbaar in tegenstelling tot OpenRefine.

**Bewerkingen zijn te exporteren en te importeren**  
OpenRefine maakt het mogelijk om de geschiedenis van bewerkingen te exporteren en te importeren als een JSON-(project)bestand. Hierdoor kunnen gebruikers hun werk en bewerkingen delen met anderen of op verschillende momenten doorgaan met het bewerken van gegevens. Naast het exporteren, kan het script (JSON-bestand) eventueel ook toegevoegd worden aan een nieuwe dataset die dezelfde verwerkingsstappen nodig heeft.  
Een Excel-bestand kan worden opgeslagen en worden gedeeld, maar de bewerkingsgeschiedenis wordt niet opgeslagen. Als het bestand een volgende keer wordt geopend, is het niet meer mogelijk om terug te gaan naar vorige bewerkingen in tegenstelling tot OpenRefine.

**Externe wikibase**

OpenRefine kan worden geïntegreerd met externe gegevensbronnen, zoals Wikibase. Dit maakt het mogelijk om externe gegevens rechtstreeks in OpenRefine te raadplegen en te bewerken. Je kan bijvoorbeeld gegevens uit een Wikibase-database importeren om ze te combineren met je lokale gegevens in OpenRefine.

Excel kan ook gegevens importeren en koppelen aan externe gegevensbronnen, maar dit vereist het handmatig zoeken en invoeren van een passende dataset om de integratie tot stand te brengen. Daarnaast vereist Excel een eigen dataset die geen spellingsfouten bevat om te koppelen aan een nieuwe dataset en meerdere ‘verticaal zoeken’ functies. OpenRefine heeft daarentegen een bepaalde ‘waarschijnlijkheidsformule’ in het systeem om in te zien dat bijvoorbeeld “Portuugal” waarschijnlijk “Portugal” inhoudt.

**Witte regels verwijderen**

OpenRefine biedt ingebouwde functies om witte regels (rijen met lege gegevens) te identificeren en te kunnen verwijderen. Gebruikers kunnen dit doen door filters toe te passen op lege cellen of lege regels om deze zichtbaar te maken en vervolgens te verwijderen als daar de wens voor is.

In Excel kunnen gebruikers ook witte regels verwijderen door filters of sorteerfuncties te gebruiken om lege rijen te identificeren en handmatig te verwijderen.

Uit bovenstaande punten kan er geconcludeerd worden dat over het algemeen OpenRefine vooral gespecialiseerd is in het transformeren en schoonmaken van ongestructureerde gegevens, terwijl Excel een veel breder scala aan functionaliteiten biedt. Denk aan gegevensanalyse, berekeningen en het maken van grafieken. De keuze tussen de twee hangt af van de specifieke behoeften van de gebruiker en de complexiteit van de taak.

## Toepasbaarheid in het werkveld

OpenRefine is een krachtige tool die in verschillende sectoren en toepassingsgebieden kan worden gebruikt. Met name voor het opschonen en transformeren van messy data. Messy data verwijst naar gegevens die ongestructureerd, ongeorganiseerd of inconsistent zijn, waardoor ze moeilijk te analyseren of te gebruiken zijn. Messy data kunnen uit verschillende bronnen afkomstig zijn, waaronder:

* *Handmatige invoer:* Dit verwijst naar gegevens die met de hand zijn ingevoerd door mensen. Het kan leiden tot fouten en inconsistenties, omdat menselijke invoer gevoelig is voor typefouten en variabiliteit.
* *Verschillende bronnen:* Dit betekent dat gegevens afkomstig zijn uit diverse bronnen, zoals databases, spreadsheets, of websites. De gegevens kunnen variëren in structuur, formaat en notaties, waardoor integratie en analyse uitdagend kan zijn.
* *Tekstvelden:* Tekstvelden zijn gegevensvelden die vrije tekst bevatten, zoals opmerkingen of reacties. Deze velden kunnen ongestructureerde informatie bevatten, waardoor ze moeilijker te analyseren zijn dan gestructureerde gegevens zoals getallen.
* *Spelfouten en synoniemen:* Spelfouten verwijzen naar foutief gespelde woorden in de gegevens, wat tot verwarring kan leiden. Synoniemen zijn verschillende woorden met vergelijkbare betekenissen. Ze kunnen inconsistenties veroorzaken als ze in gegevenssets worden gebruikt.
* *Onvolledige gegevens:* Dit verwijst naar ontbrekende gegevens, zoals ontbrekende waarden in een dataset, ontbrekende labels of onvolledige informatie. Het kan de analyse bemoeilijken en onzekerheid geven in de resultaten.

Hier zijn enkele beschreven toepassingsgebieden waarin de software OpenRefine effectief kan worden toegepast. Dit illustreert hoe het programma in de praktijk wordt gebruikt binnen diverse professionele contexten.

**Journalistiek**

Journalisten kunnen OpenRefine gebruiken om ruwe gegevens van verschillende bronnen te verwerken en te transformeren in een formaat dat geschikt is voor journalistiek onderzoek. Dit kan het analyseren van tekst, het extraheren van belangrijke informatie uit documenten of het opschonen van datasets omvatten.

**Financieel**

In de financiële sector kan OpenRefine worden gebruikt om financiële gegevens te analyseren en te transformeren. Het kan helpen bij het opschonen van financiële rapporten, het harmoniseren van gegevens uit verschillende bronnen en het identificeren van trends of patronen in financiële gegevens.

**Inschrijvingen**

Organisaties kunnen OpenRefine gebruiken om inschrijvingsformulieren te verwerken en de ingevoerde gegevens te standaardiseren. Dit kan nuttig zijn voor het beheer van evenementen, cursussen of andere inschrijvingen.

**Feedbackformulieren**

Feedback van klanten of gebruikers komt vaak in verschillende formaten en vormen binnen. OpenRefine kan worden gebruikt om feedback te analyseren en te categoriseren, zodat organisaties inzicht kunnen krijgen in klantopmerkingen en -suggesties.

**Enquêtes**

Bij het analyseren van enquêteresultaten kunnen OpenRefine-technieken worden toegepast om open tekstreacties te categoriseren, trends te identificeren en gegevens te standaardiseren voor verdere analyse.

OpenRefine biedt tools en functies om deze problemen aan te pakken door gegevens op te schonen, te standaardiseren, te transformeren en te verrijken, waardoor de gegevens geschikt worden voor analyse en rapportage. Het is bij uitstek geschikt voor het verwerken van zowel gestructureerde als ongestructureerde gegevens voor verscheidene sectoren. Dit maakt het daarmee een erg toepasbare tool in het werkveld.

# Handleiding

Bij de onderstaande stappenplannen wordt een [Excel-document](https://hogeschoolutrecht-my.sharepoint.com/:x:/g/personal/marjolein_wildeboer_student_hu_nl/EYYuBbNn7LlCv8yc8AXplVgB_6UWB9uGdGdIwC-tHu7yTA?e=lcHeB6) gebruikt over de boekingen voor het ‘City Hotel’ voor de maanden januari t/m mei van de jaren 2015 t/m 2016. Door middel van onderstaande handleiding zal duidelijk worden hoe je OpenRefine kan gebruiken om vragen te beantwoorden en het document op te schonen naar eigen wens. De mogelijkheden die hieronder worden laten zien, zijn op meerdere manieren te bereiken binnen OpenRefine; dit zijn enkel voorbeelden.

## **Facets**

**Afbeelding met tekst, schermopname, diagram, ontwerp

Automatisch gegenereerde beschrijvingFacet by blank**

***Hoe verwijder je rijen zonder informatie?***

**Het huidige bestand bevat 79330 rijen. Wanneer je bij een willekeurige kolom (behalve de kolom ‘all’) op het pijltje klikt, facet, customized facets en facet by blank, kan je kijken of hier ook rijen bij zitten die geen informatie bevatten en daarmee opgeschoond/verwijderd moeten worden.**

**Aan de linkerkant is te zien dat er 49915 rijen zijn die overeenkomen (true) met deze vraag. Door vervolgens bij all naar edit rows en remove matching rows te gaan (matchend met voorgaande geselecteerde true die oranje is geworden), worden de overbodige rijen verwijderd.**

**Tekst facet**

*Hoe zie ik welke jaren in het document staan?*

Afbeelding met tekst, schermopname, nummer, software

Automatisch gegenereerde beschrijvingDoor middel van tekst facet zijn de unieke waarden in een kolom te zien. Door bij de benodigde kolom naar facet en tekst facet te gaan, is er te zien dat dit gaat om 2015 t/m 2017, waarbij 2016 blijkbaar de meeste boekingen had. Dit kun je zien door bovenin bij sort by voor count te kiezen in plaats van name.

**Numeric facet**

**Afbeelding met tekst, schermopname, diagram, software

Automatisch gegenereerde beschrijving***Hoe kan ik in één oogopslag zien voor hoeveel volwassenen er gemiddeld wordt geboekt per boeking?*

**Wanneer je bij de kolom ‘adults’ bij facet de numeric facet selecteert, krijg je links een bepaald interval (histogram) te zien die loopt van 0-4.02. In het midden is een grote staaf te zien met het aantal volwassenen die blijkbaar het meest voorkomt bij boekingen. Door de schuifknoppen aan de zijkanten van het histogram te gebruiken om de grootste staaf te selecteren, is er te zien dat dit zal gaan om voornamelijk ongeveer 2 volwassenen per boeking.**

**Tekst filter**

**Afbeelding met tekst, schermopname, software, nummer

Automatisch gegenereerde beschrijving*Ik wil alleen de data van alle maanden behalve mei inzien, hoe doe ik dat?***

**Wanneer je op het pijltje klikt van de kolom ‘arrival\_date\_month’ en vervolgens kiest voor tekst filter, krijg je een zoekfunctie aan de linkerkant in beeld. Ook wel bekend als de gebruikelijke ctrl + f sneltoets. Hier kan je ‘may’ invoeren zodat je enkel alle data van mei te zien krijgt. Je wilt echter alles behalve de maand mei zien. Dit kan door boven in de balk van de zoek functie te klikken op invert.**

**Afbeelding met tekst, schermopname, software, diagram

Automatisch gegenereerde beschrijvingClustering**

***Zijn alle statussen bij ‘reservation\_status’ goed gespeld?***

**Door middel van de cluster and edit functie kan onder andere gekeken worden of er spelfouten aanwezig zijn om deze vervolgens te corrigeren. Wanneer je bij de kolom ‘reservation\_status’ kiest voor edit cells en cluster and edit, krijg je een nieuw scherm in beeld. Door de keying function te veranderen in metaphone3, komen er spellingsfouten naar voren voor de status ‘Canceled’. De new cell value aan de rechterkant betreft de goede spelling die OpenRefine voorstelt. Deze klopt en kan door middel van een geselecteerd vinkje bij ‘merge?’ en vervolgens merge selected & close, gecorrigeerd worden om vervolgens dit venster te sluiten.**

**Sorting**

**Afbeelding met tekst, schermopname, software, Computerpictogram

Automatisch gegenereerde beschrijving*Wat is het hoogste aantal kinderen voor wie een keer/meerdere keren een boeking is geplaatst?***

**Door bij de kolom ‘children’ te kiezen voor de optie sort, krijg je een nieuw venster in beeld om te bepalen wat je wilt sorteren en hoe. De default setting geeft aan de rechterkant aan dat eerst de valid values worden getoond; dit wil je behouden. Aan de linkerkant kies je voor numbers en vervolgens largest first om de vraag te kunnen beantwoorden. Boven in de balk (boven de dataset) verschijnt nu het woordje sort. Het antwoord op de bovenstaande vraag   
- 3 - is bekend door de verandering in de dataset, maar als je de dataset permanent wilt veranderen naar deze sortering, kies je voor reorder rows permanently. Als je de sortering wilt verwijderen (wanneer je het nog niet permanent hebt toegepast), kan dit door op remove sort te klikken.**

## **Edit cells**

**Trim leading and trailing - & collapse consecutive whitespace**

Afbeelding met tekst, Lettertype, nummer, lijn

Automatisch gegenereerde beschrijving*Staan er onnodige spaties in de kolom ‘country’?*

Door op de kolom ‘country’ te klikken en vervolgens bij edit cells te kiezen voor common transformations, zijn er bovenin twee opties te zien die gaan om ‘whitespaces’, oftewel, spaties. Overbodige spaties kunnen plaatsvinden voor en na een woord (\_woord\_) of als twee opeenvolgende spaties (een\_\_woord). Door op bovenstaande twee opties te klikken (bovengenoemde stappen twee keer nalopen), wordt dit probleem verholpen, als hier sprake van is in de dataset. Boven in het scherm komt een gele balk in beeld waarin voor beide opties te zien is dat dit niet ter sprake was.

**Celtype veranderen**

Afbeelding met tekst, schermopname, nummer, diagram

Automatisch gegenereerde beschrijving*Hoe zorg ik ervoor dat de getallen in een kolom ook als getal gezien worden door het programma (celtype getal)?*

Wanneer er gekeken wordt naar de kolom met het aantal ‘special requests’, is er te zien dat deze als een van de enige getallen bevat die niet groen en rechts uitgelijnd zijn, maar zwart en links uitgelijnd. Dit betekent dat het momenteel in het bestand staat als een ‘cel die tekst bevat’. Je wilt echter dat het programma deze gaat lezen als een getal met een waarde. Door op de desbetreffende kolom te klikken, te kiezen voor common transforms en to number, worden alle waarden in deze kolom veranderd naar het celtype ‘getal’. Boven in het scherm staan in de gele balk dat alle rijen (29415) zijn veranderd naar het celtype getal en de getallen in de kolom staan nu groen en rechts uitgelijnd.

## **Edit columns**

**Split into several columns**

**Afbeelding met tekst, software, nummer, schermopname

Automatisch gegenereerde beschrijving*Hoe zorg ik ervoor dat in de kolom met jaartallen niet steeds het jaar 2000 wordt herhaald, maar enkel ‘15, ‘16 en ‘17 worden aangegeven?***

**Wat je hiervoor nodig hebt is een splitsing tussen 20 en 15/16/17, zodat je enkel deze laatste cijfers te zien krijgt, maar dan voor iedere cel. Door te klikken op de kolom ‘arrival\_date\_year’ en vervolgens te kiezen voor edit column en split into several columns, komt er een nieuw venster in beeld. Hierin kun je de ‘seperator’ aangeven, oftewel, op basis van welk scheidingsteken wil je dat er een splitsing wordt gemaakt in de waardes? In dit geval is het cijfer 0; vanaf dit punt wil je namelijk dat er een splitsing wordt gemaakt. Nu is te zien dat er een nieuwe kolom is aangemaakt met enkel de cijfers 15, 16 of 17 die verwijzen naar 2015, 2016 of 2017.**

**Afbeelding met tekst, schermopname, nummer, Lettertype

Automatisch gegenereerde beschrijvingRemove this column**

*Hoe zorg ik ervoor dat de nieuwe kolom waarin enkel 2’en staan vermeld wordt verwijderd (zie uitkomst vorige vraag)?*

Door op de kolom ‘arrival\_date\_year 1’ te klikken en vervolgens te kiezen voor edit column en remove this column, wordt deze onnodige kolom verwijdert en daarmee de dataset verder opgeschoond.

## **Reconcile**

**Reconciliation** **proces**

**Afbeelding met tekst, software, diagram, schermopname

Automatisch gegenereerde beschrijving*Hoe zorg ik ervoor dat de afkortingen van de landen volledig uitgeschreven worden?***

**In de kolom ‘country’ staan enkel de afkortingen van de landen. Om deze te veranderen naar gehele namen, heb je een andere dataset nodig die zowel de afkortingen van de landen weet als de volledige naam. Dit kan door middel van de functie reconcile. Door op de kolom ‘country’ te klikken en vervolgens te kiezen voor reconcile en start reconciling, komt er een nieuw venster in beeld. Hierin moet je kiezen welke service je wilt gebruiken. De default setting geeft de Wikidata-service aan die tal van datasets bevat. Wanneer je deze selecteert, krijg je wederom een nieuw venster in beeld. Dit keer staan aan de linkerkant datasets vanuit wikibase die mogelijk hiervoor gebruikt kunnen worden. Hierin kies je voor country. Aan de rechterkant staan de namen van de kolommen die in de dataset aanwezig zijn. Stel dat hier nog kolommen bij zaten die bijvoorbeeld extra context geven over de landen, kan je deze als extra ‘hulpmiddel aan het programma’ toevoegen, maar dit is hier niet het geval.**

**Afbeelding met tekst, schermopname, nummer, Lettertype

Automatisch gegenereerde beschrijvingUiteindelijk zorgt het programma ervoor dat alle afkortingen gekoppeld worden aan de namen van de landen. Aangezien de voorkeur niet uitgaat om dit handmatig te doen/checken, staat het vakje auto-match candidates with high confidence geselecteerd. Nu kan je kiezen voor start reconciling.**

**Boven in het scherm is nu een geel balkje te zien waarin de voortgang van het ‘reconcile-proces’ bij te houden is. Wanneer dit proces voltooid is, zijn de namen van de landen die overeenkomen met de afkortingen automatisch gematched en weergegeven in de dataset. Aan de linkerkant staat dat alle rijen gematched zijn en dit is ook te checken aan de hand van het groen gevulde balkje onder de kolomnaam ‘country’.**

**Reconciliation permanent maken**

**Afbeelding met tekst, schermopname, diagram, lijn

Automatisch gegenereerde beschrijving*Hoe zorg ik ervoor dat de tekst facet functie (zie tekst facet) de hele namen van de landen aangeeft in plaats van de afkortingen van voorheen*** *(zie uitkomst vorige vraag)?*

***Afbeelding met tekst, diagram, Perceel, lijn

Automatisch gegenereerde beschrijving*Wanneer je bij de nieuwe kolom met de volledig uitgeschreven namen van de landen een tekst facet toevoegt (zie tekst facet), krijg je weer de oude afkortingen te zien in plaats van de nieuwe volledige namen. Dit wordt opgelost door middel van de functie add column from reconciled values: Je klikt op de kolom ‘country’ en vervolgens op edit column om op de desbetreffende functie uit te komen. Er komt een nieuw venster in beeld, waarin je links bovenin wederom aangeeft op basis van welke dataset de afkortingen van de landen zijn ‘reconciled’. Na het invoeren van ‘country’ klik je op oké. Nu is er een nieuwe kolom bijgekomen die er weliswaar hetzelfde uitziet (met een andere kolomnaam), maar wanneer je hier een tekst facet toevoegt, is er dit keer te zien dat de nieuwe waardes zijn overgenomen.**

## **Script**

**Afbeelding met tekst, schermopname, Website, Webpagina

Automatisch gegenereerde beschrijving*Is er een manier waardoor ik niet alle stappen hoef te herhalen voor de data van de maanden juni t/m december?***

**OpenRefine heeft de optie om alle stappen waarin je veranderingen hebt aangebracht in de dataset als zijnde een soort script op te slaan en toe te kunnen passen op een nieuw document. Bij de undo/redo tab zijn al deze permanente veranderingen in te zien. Door vervolgens bovenin te klikken op extract, krijg je een nieuw venster in beeld. Hierin kan je aan de linkerkant aangeven welke stappen je wilt overnemen om uit te voeren in een nieuw document. Vervolgens staat er aan de rechterkant een zogeheten ‘JSON’ type script die je kan kopiëren naar je klembord. Wanneer je het nieuwe document opent voor de maanden juni t/m december, ga je wederom naar de undo/redo tab, maar klik je dit keer op apply. Hier zet je het gekopieerde script in, waarna je klikt op perform operations.**

# Verdieping/verbreding

Wanneer deze handleiding met succes is doorlopen, bent u als gebruiker in het bezit van de basiskennis en vaardigheden voor het gebruik van OpenRefine. U kunt met het programma aan de slag. Echter bevat deze handleiding slechts een klein deel (de basis) van de functies van het programma. Als u zich verder wilt verdiepen in het programma, zijn er diverse mogelijkheden om uw kennis en vaardigheden verder uit te breiden, waaronder de codeertaal GREL. Hoe dit toegepast kan worden bij OpenRefine en welke codeerfuncties gebruikt worden, is te zien op de website van [OpenRefine GREL-functie](https://openrefine.org/docs/manual/grelfunctions).

Op de website van OpenRefine kunt u alle functies van het programma vinden om u verder te verdiepen in de tool [OpenRefine Website](https://openrefine.org/docs).

Voorbeeld: Gebruik van GREL-codeertaal  
Afbeelding met tekst, schermopname, software, nummer

Automatisch gegenereerde beschrijvingDoor middel van de GREL-codeertaal, kan bijvoorbeeld vanuit het [‘hotel boekingen document’](https://hogeschoolutrecht-my.sharepoint.com/:x:/g/personal/marjolein_wildeboer_student_hu_nl/EYYuBbNn7LlCv8yc8AXplVgB_6UWB9uGdGdIwC-tHu7yTA?e=lcHeB6) (gebruikt bij de handleiding) de kolom met data, geformuleerd als zijnde ‘2015-07-02T22:00:00Z’, omgezet worden naar bijvoorbeeld een kolom met de datum op deze wijze: 02/07/2015.

Nadat de kolom is gesplitst op basis van de seperator ‘T’, wordt er op de desbetreffende kolom geklikt en kies je voor edit cells en vervolgens transform. De huidige opdracht vanuit de programmeertaal die wordt gegeven is ‘value’; oftewel als de value 5 is, wil ik ook de value 5 zien. In dit geval wil je dat de value (als datum en string zijnde) wordt veranderd naar de formulering ‘dag/maand/jaar’. Door middel van de volgende code: value.toDate.toString(“dd/MM/yyyy”) wordt dit gerealiseerd.

# Bronnenlijst

*OpenRefine user manual*. (2022, December 29). OpenRefine. Geraadpleegd op 16 oktober 2023,   
van <https://openrefine.org/docs>

*OpenRefine History*. (2013, 12 oktober). OpenRefine. Geraadpleegd op 16 oktober 2023,   
van <https://openrefine.org/blog/2013/10/12/openrefine-history>

# Bijlagen

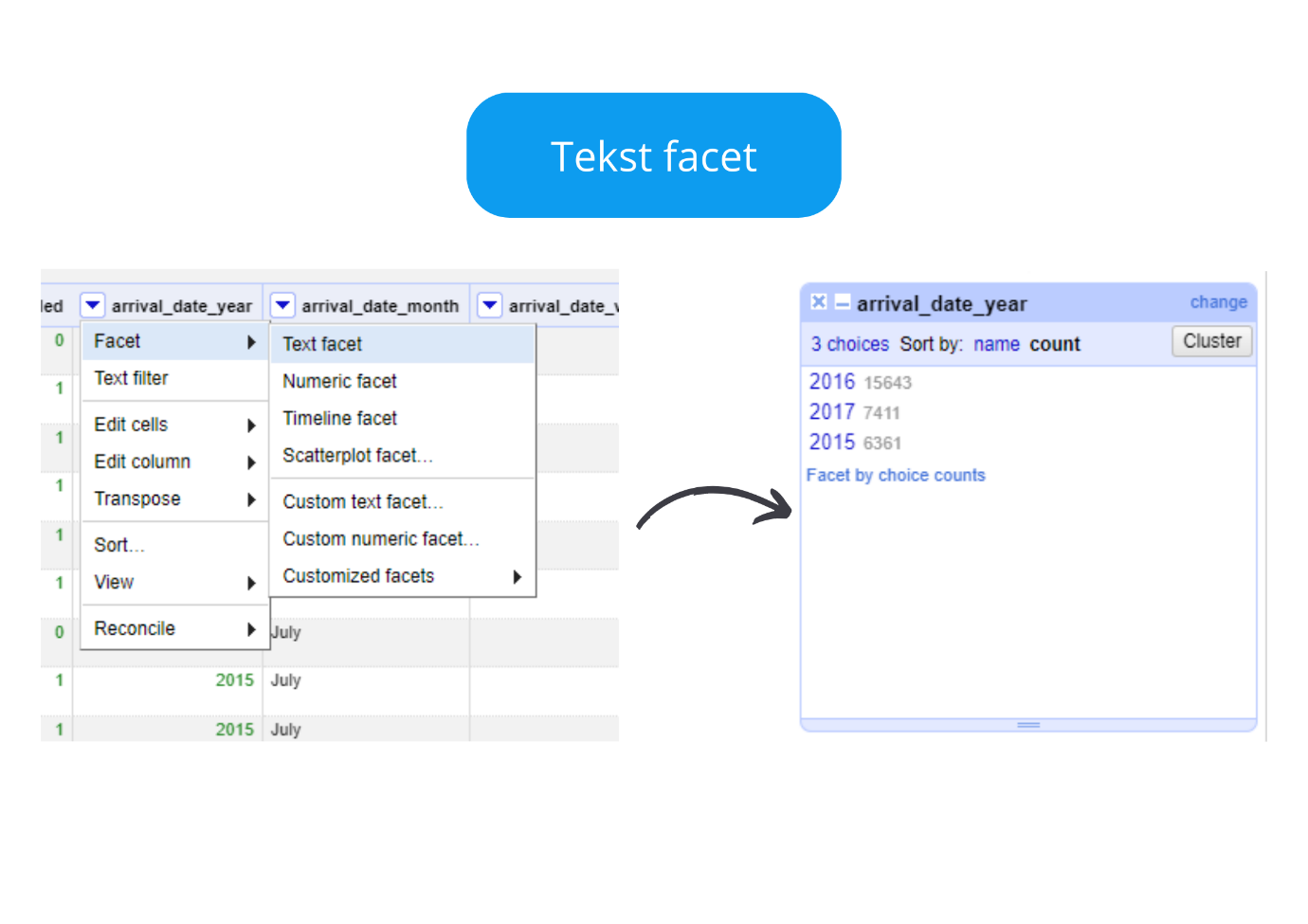
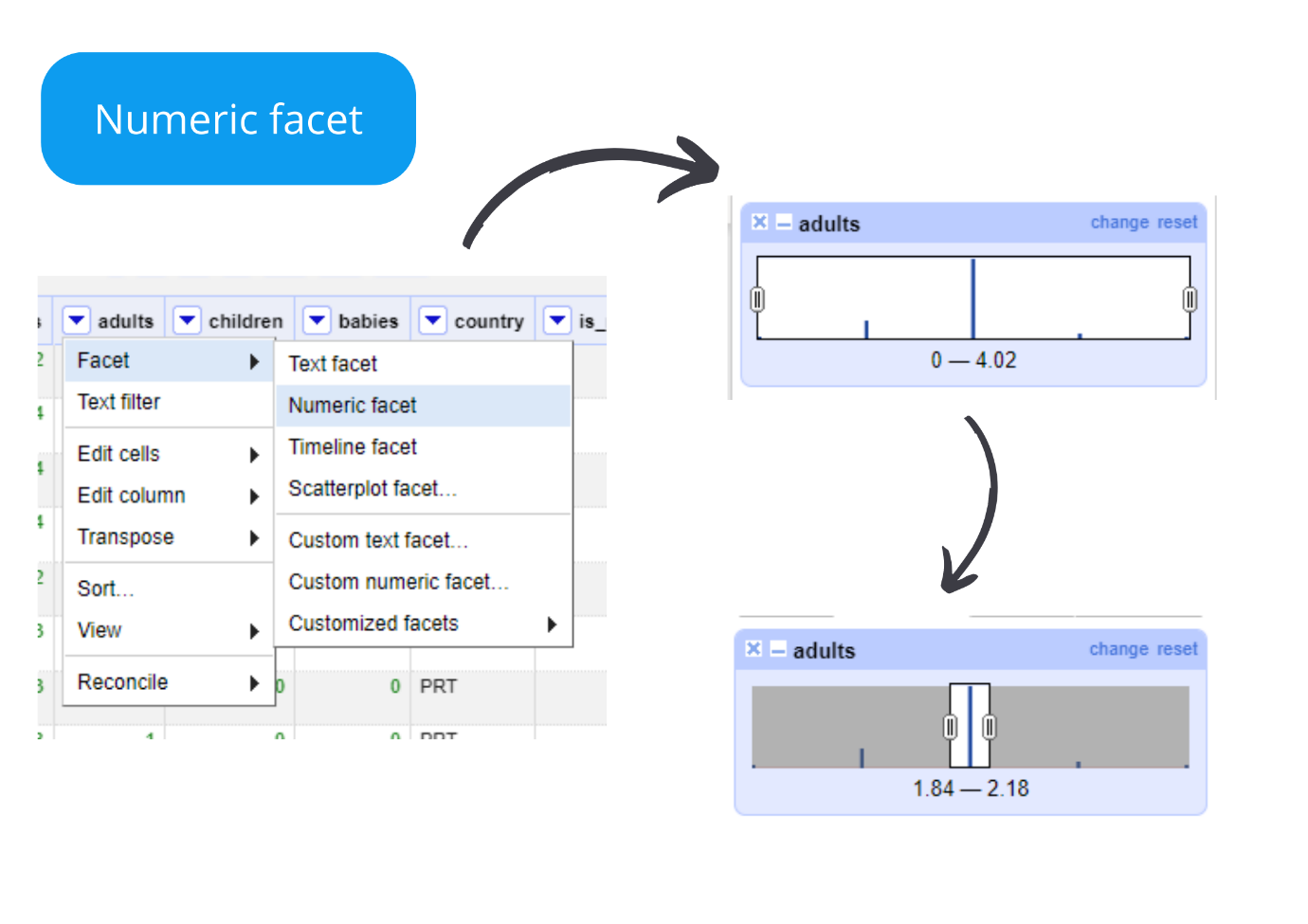
## Bijlage 1: Mail voorafgaand aan de tutorialdag

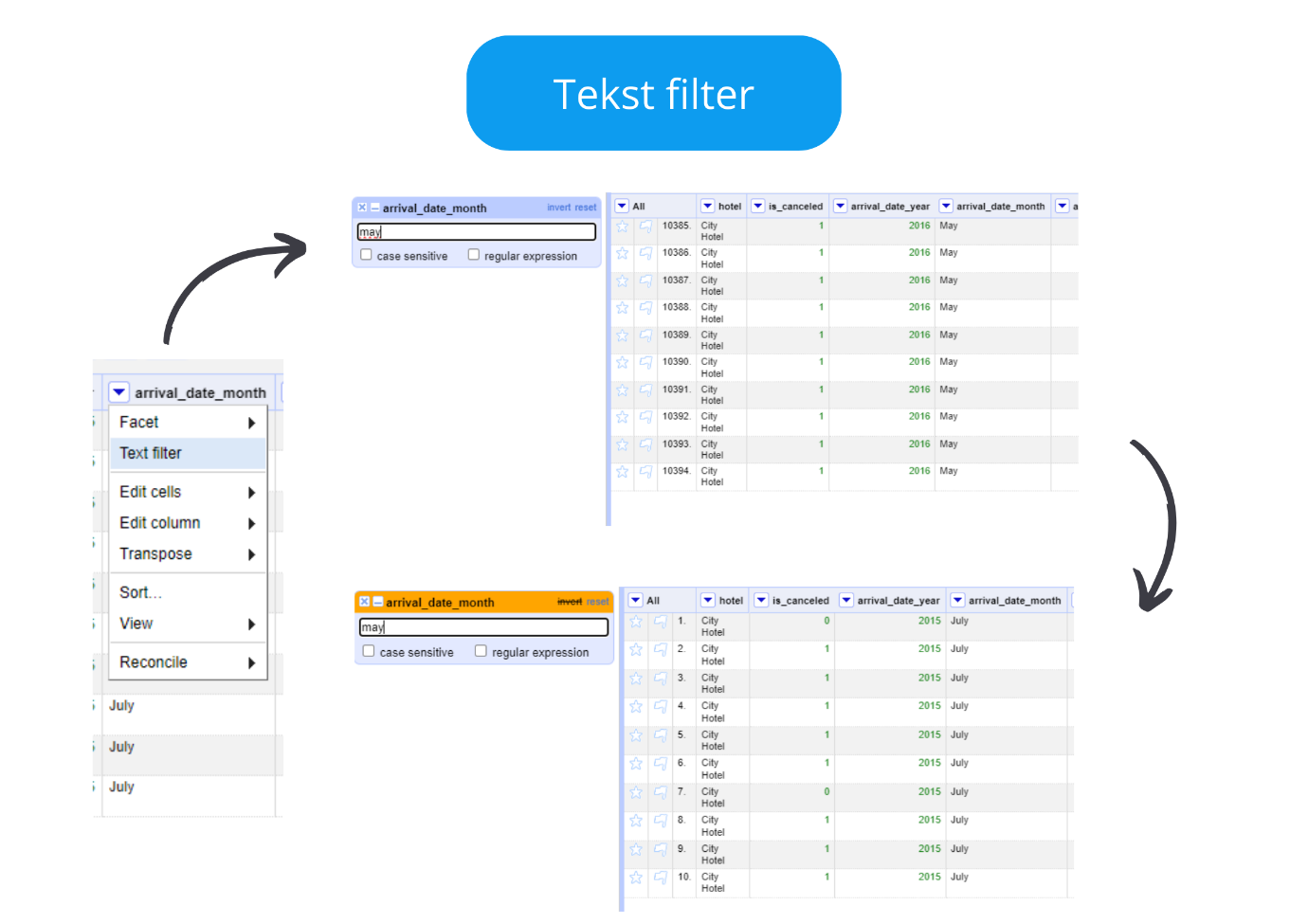
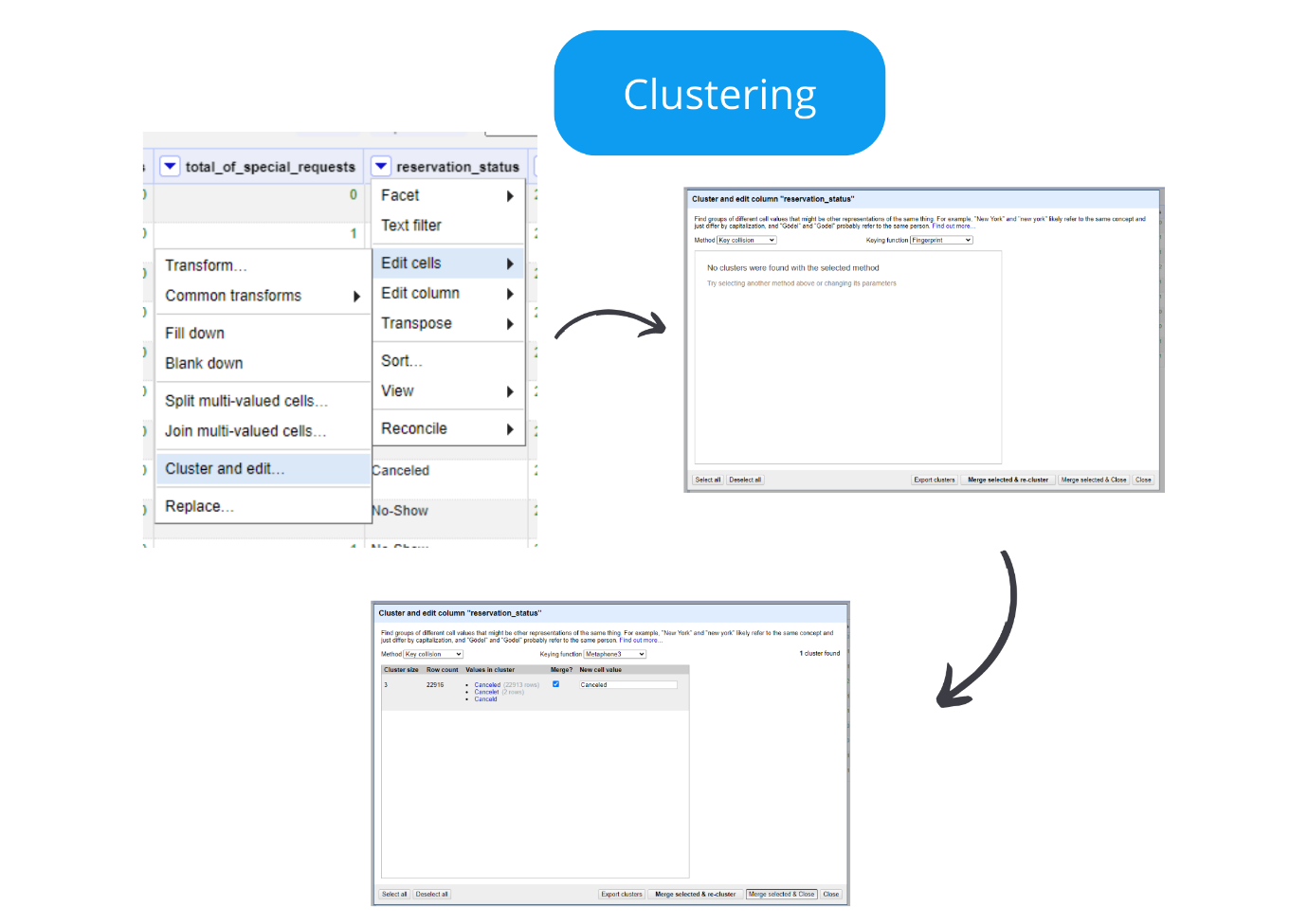
Bijlage 1: Mail downloaden OpenRefine

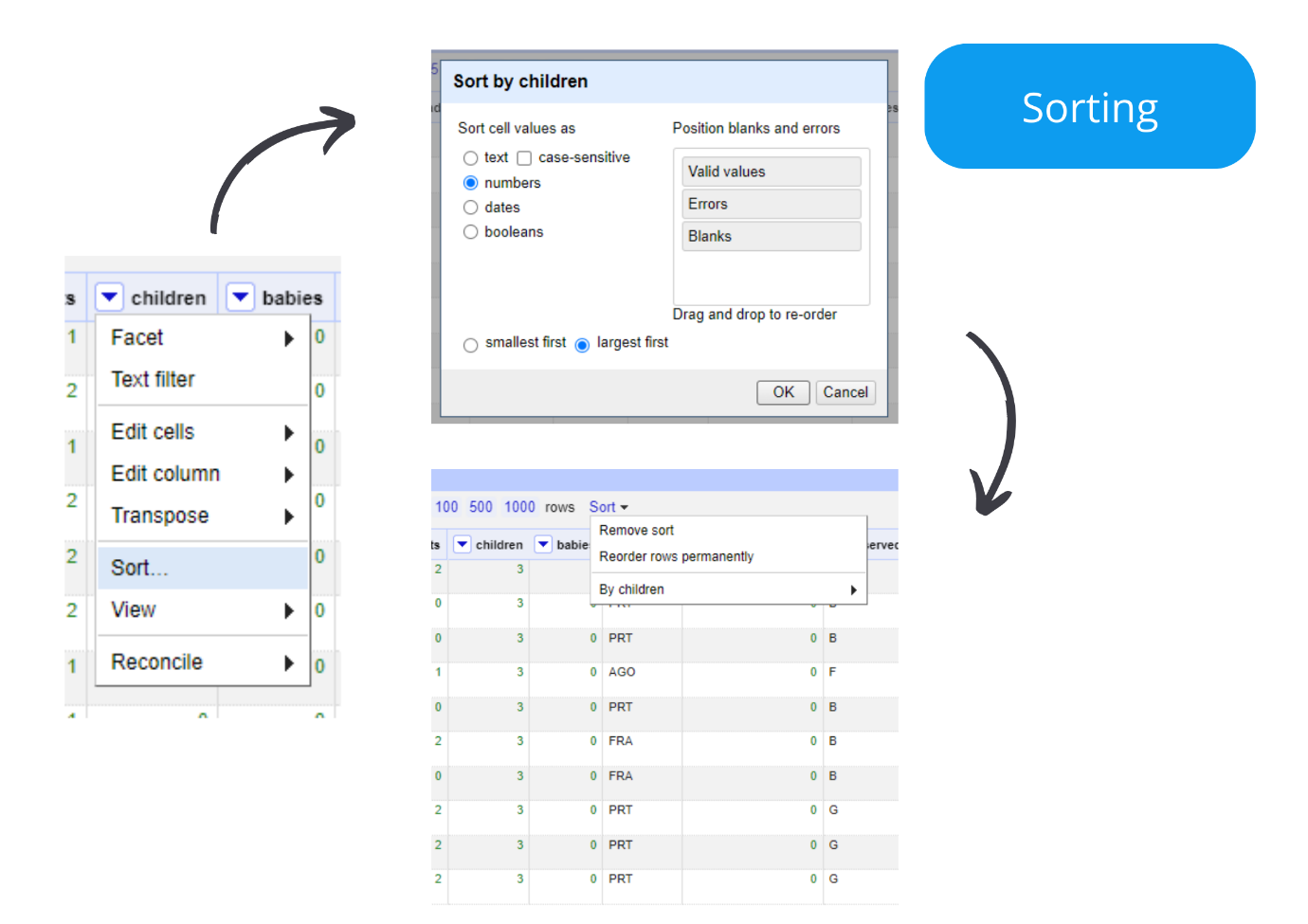
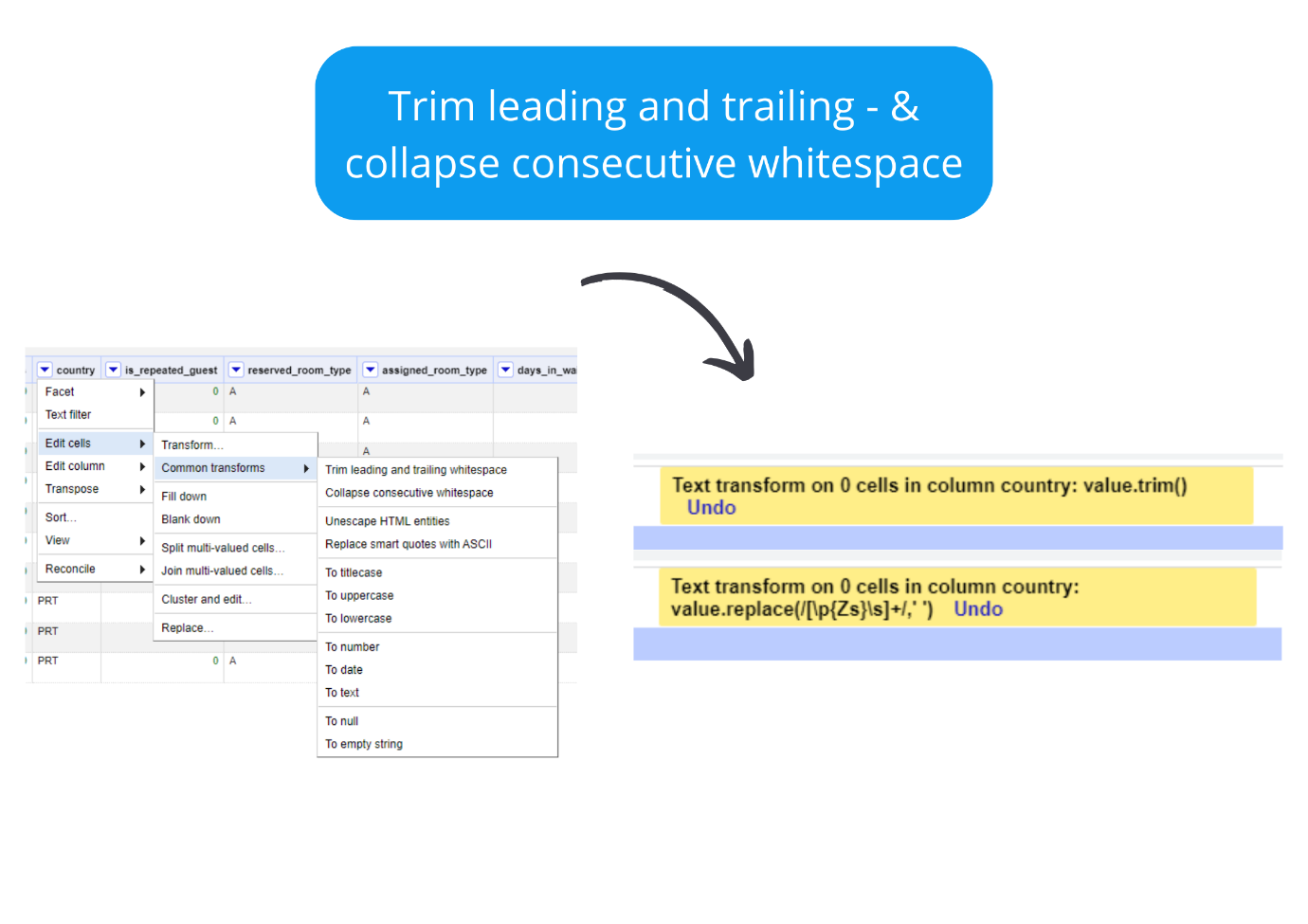

## Bijlage 2: Dataset bij handleiding

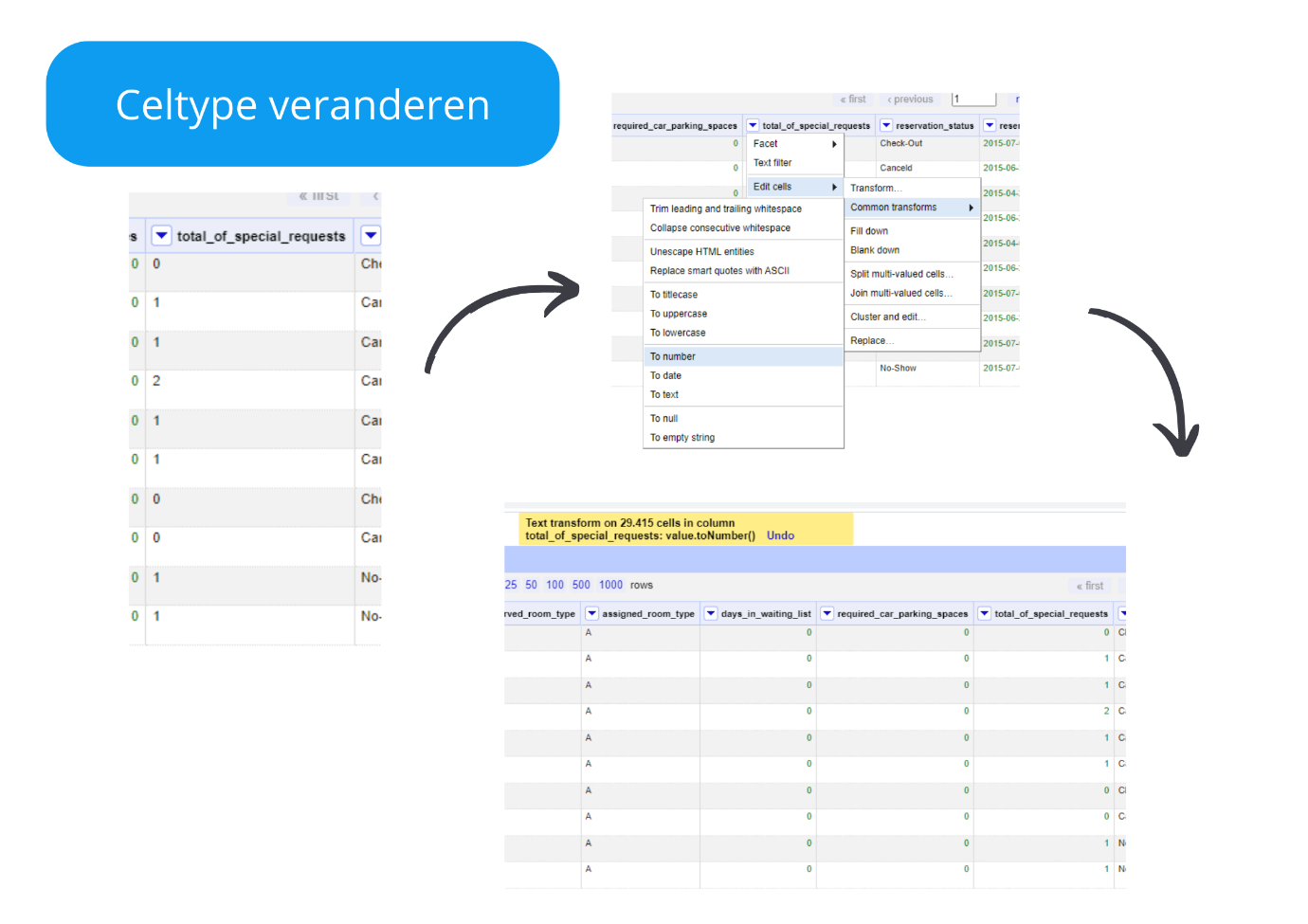
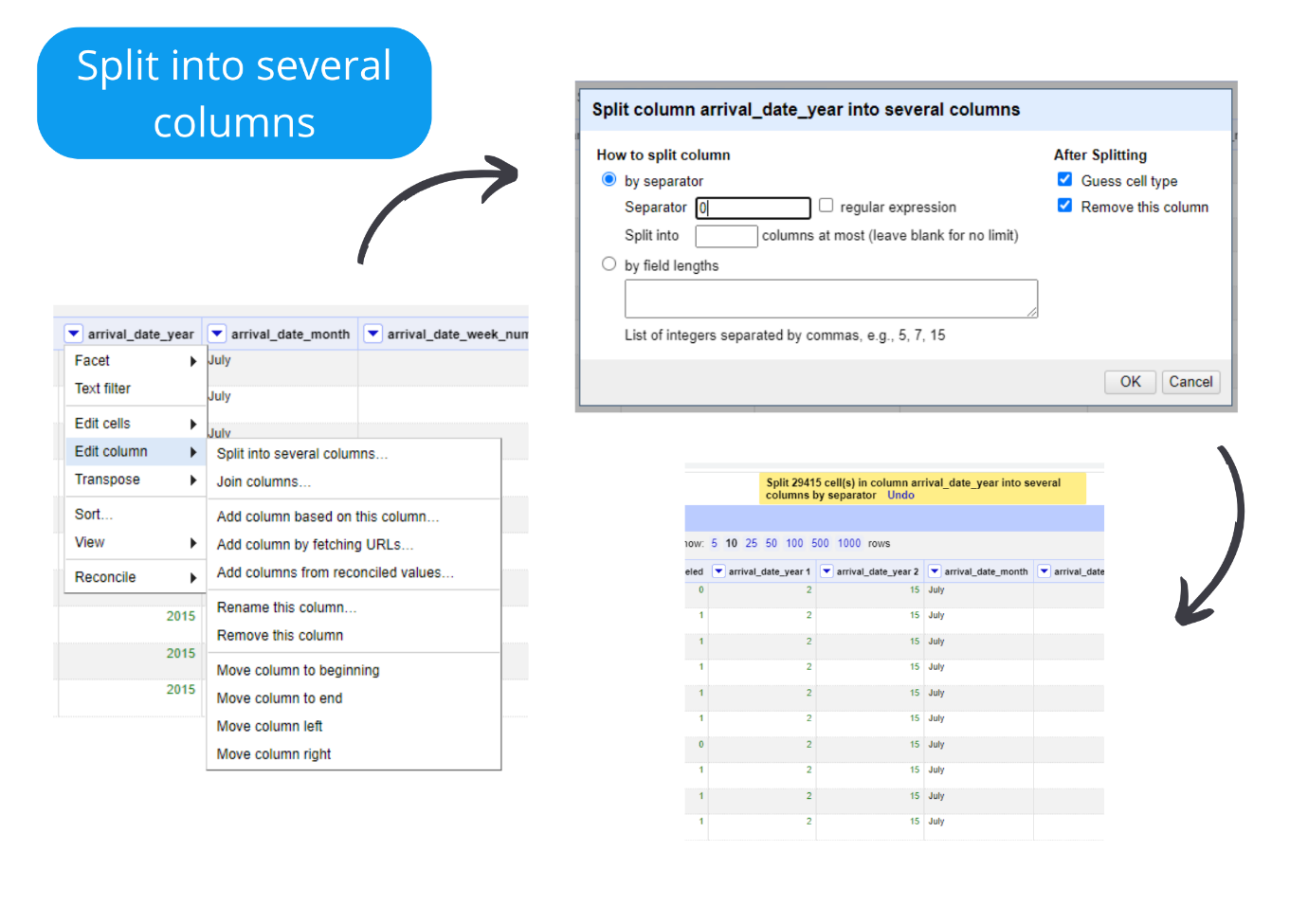
[Excel-document](https://hogeschoolutrecht-my.sharepoint.com/:x:/g/personal/marjolein_wildeboer_student_hu_nl/EYYuBbNn7LlCv8yc8AXplVgB_6UWB9uGdGdIwC-tHu7yTA?e=lcHeB6); hotelboekingen ‘City Hotel’ van januari t/m mei 2015, 2016 en 2017

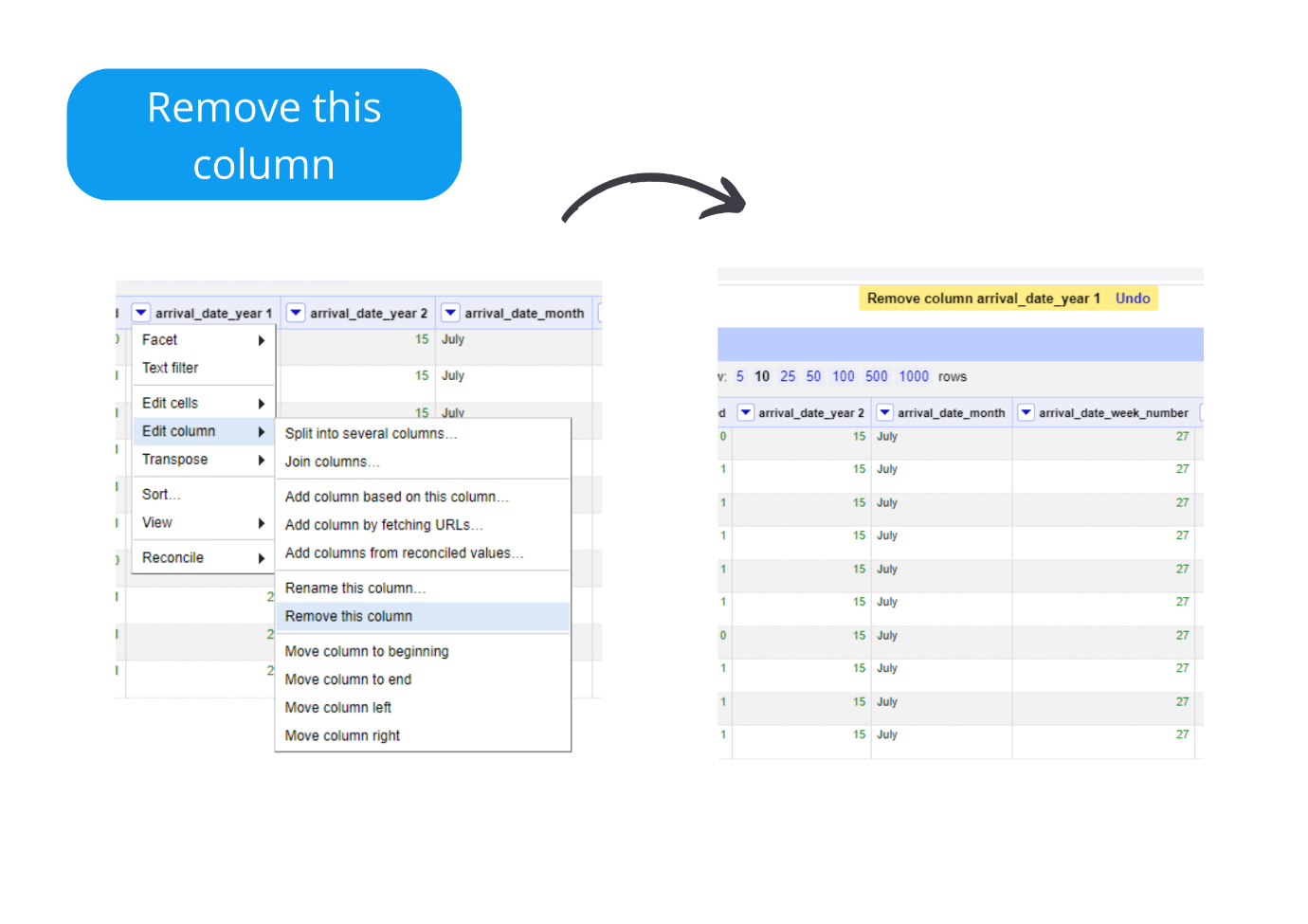
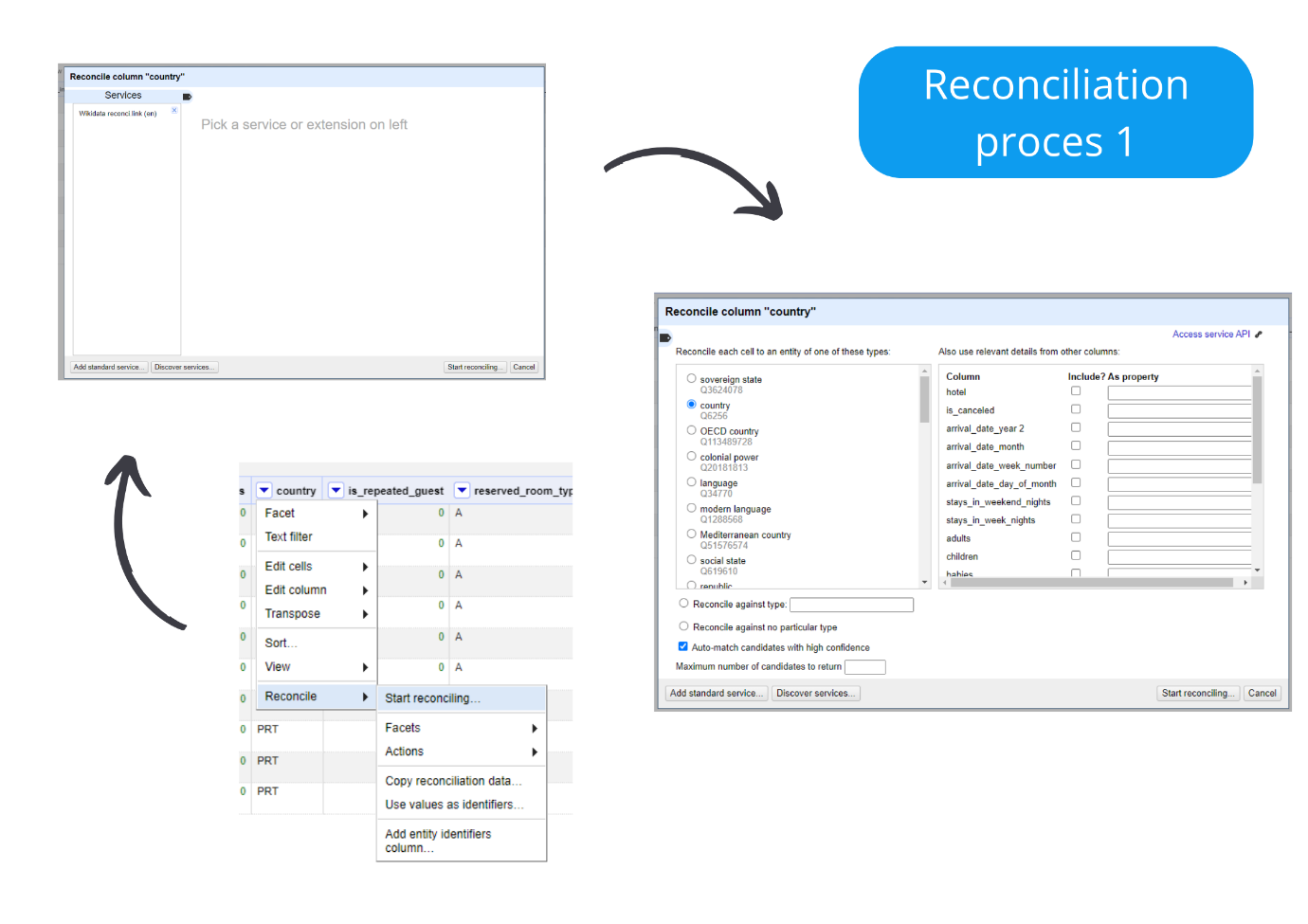
## Bijlage 3: Afbeeldingen bij handleiding

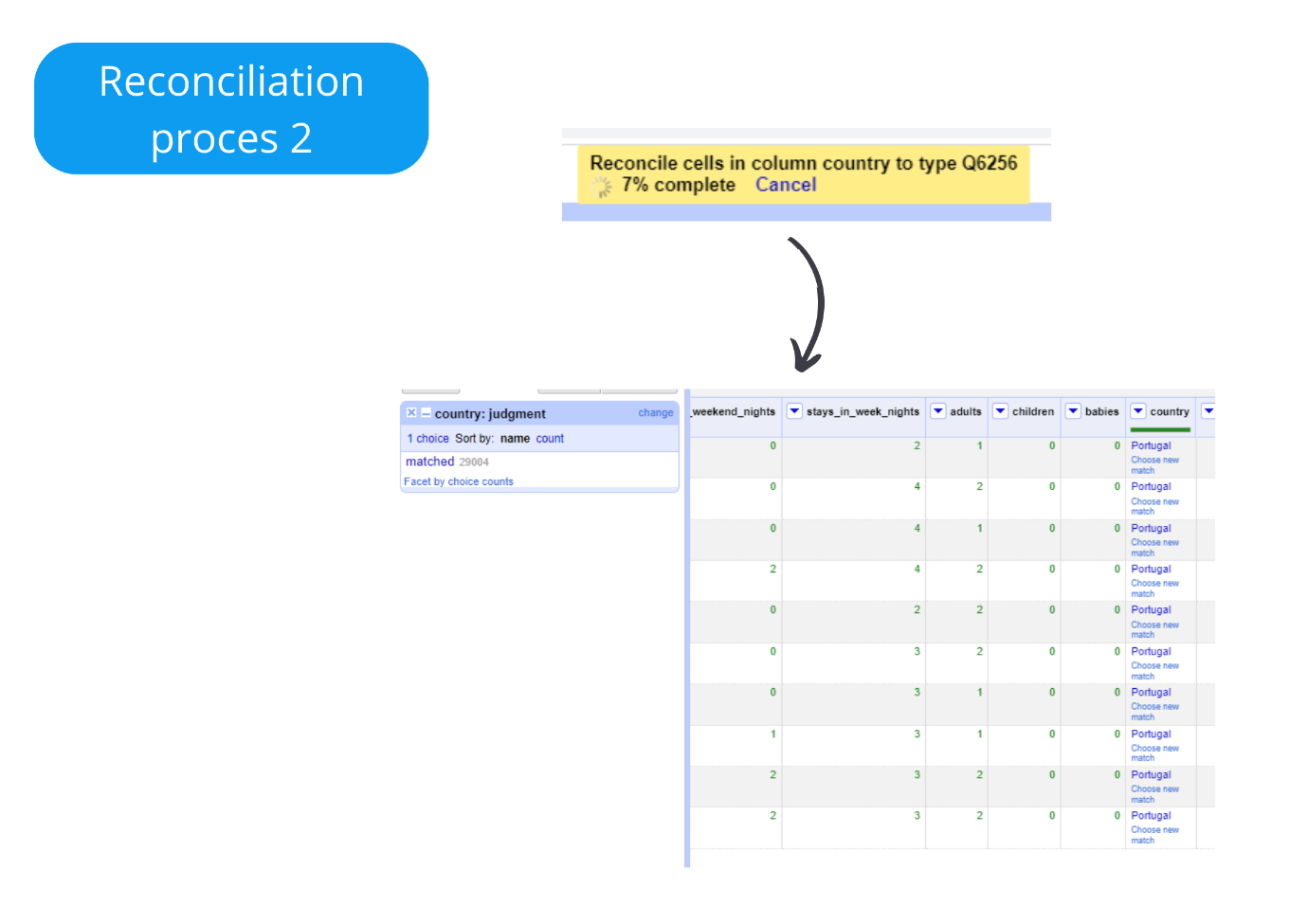
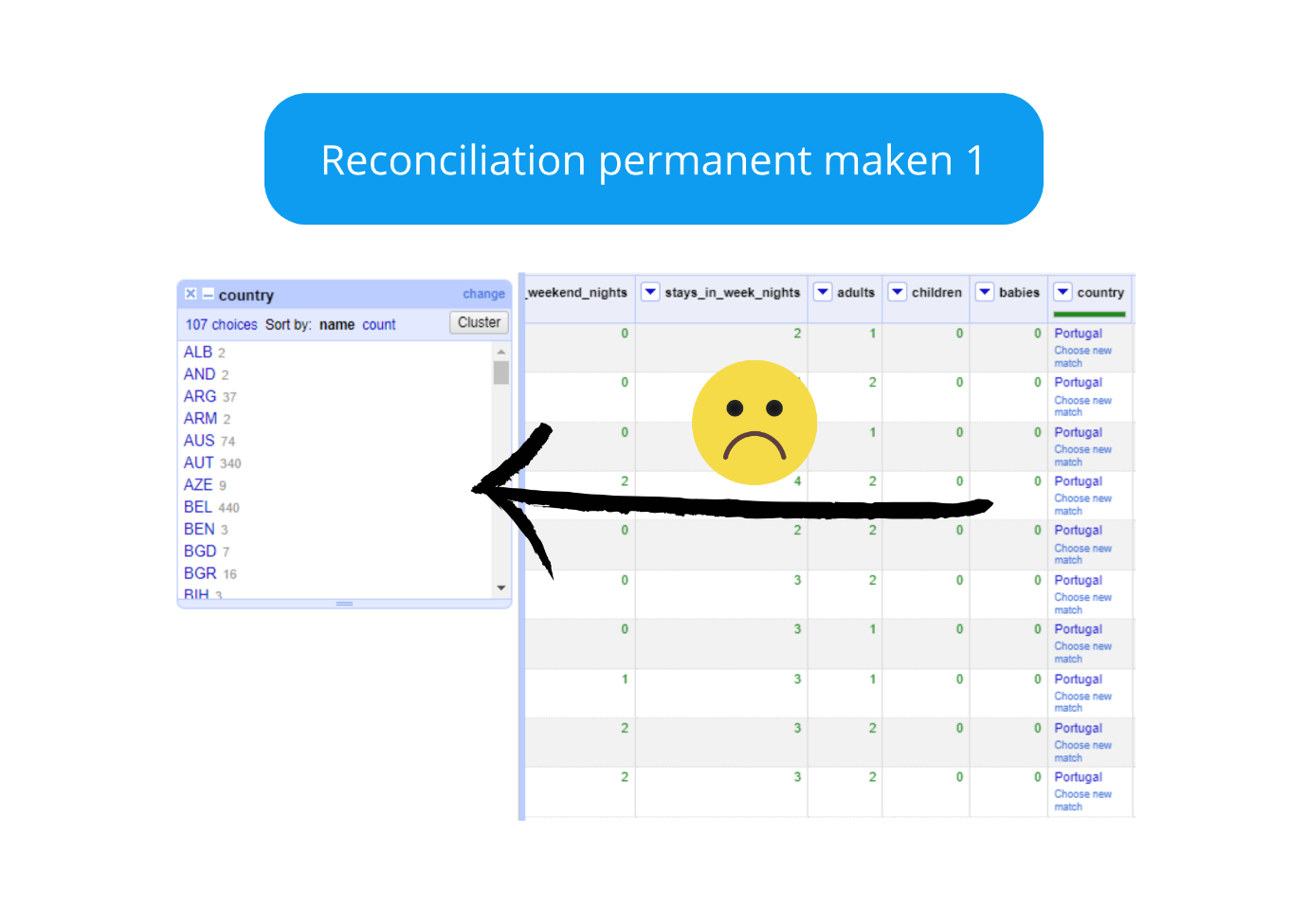


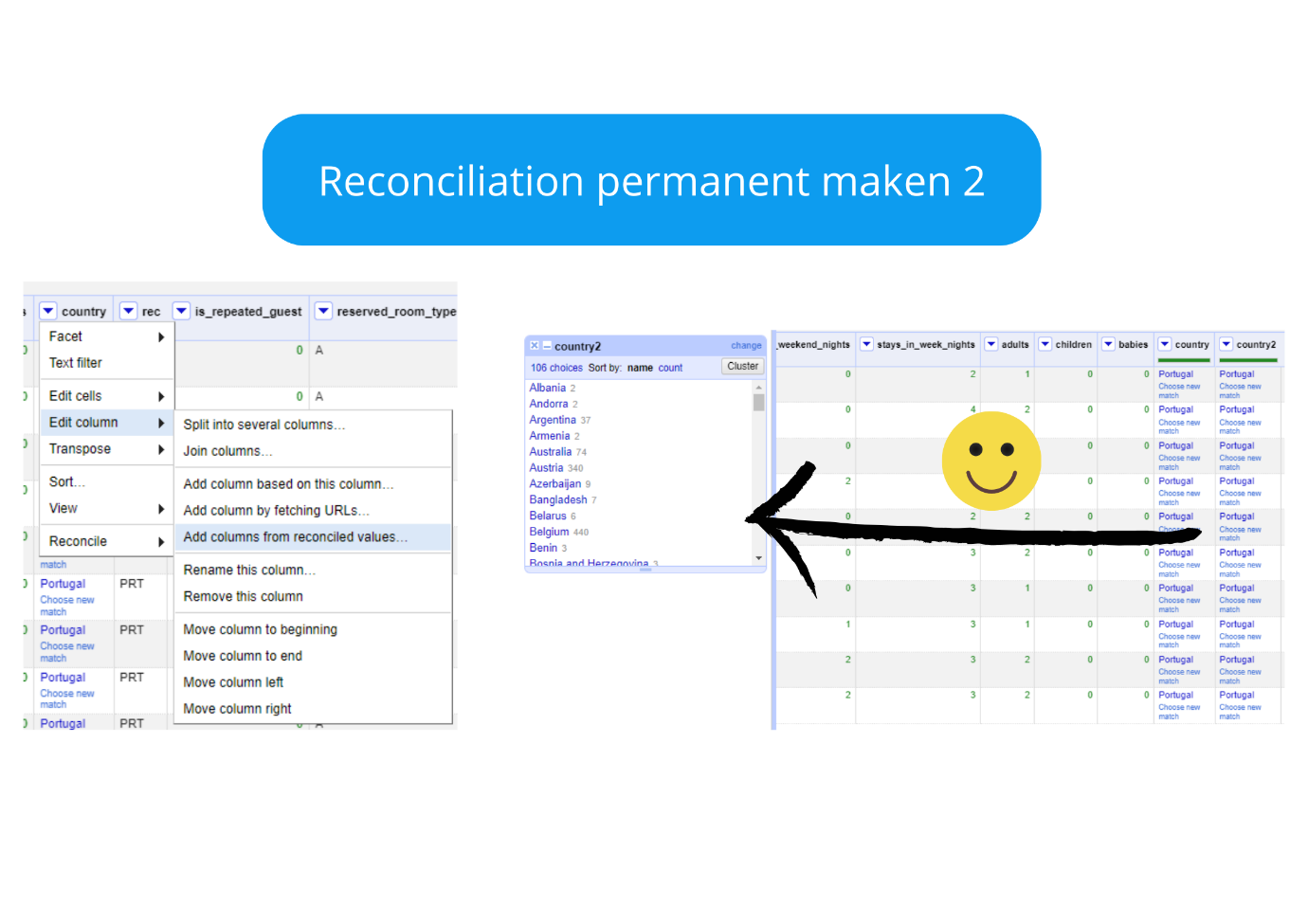
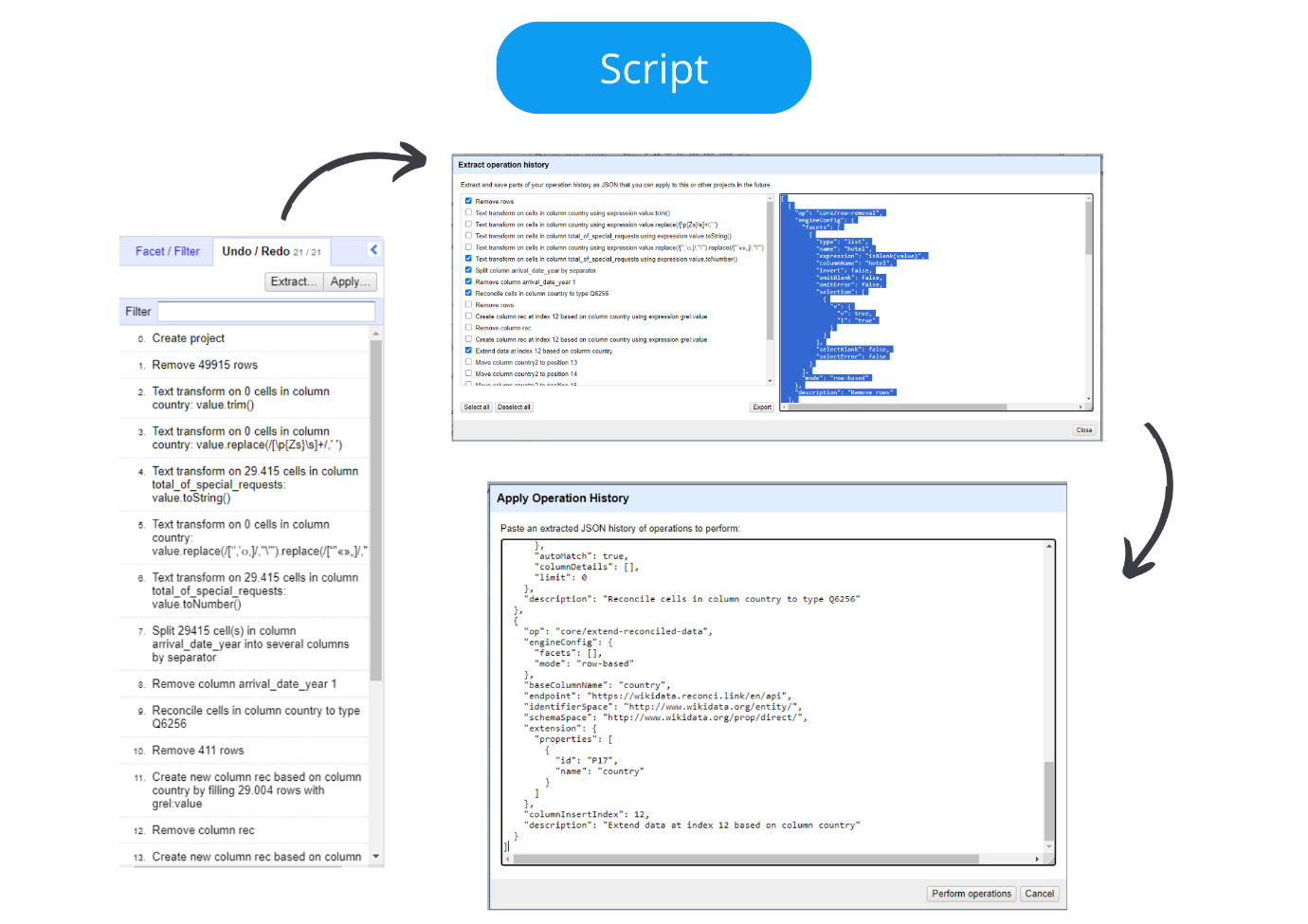












## Bijlage 4: Afbeelding bij verdieping/verbreding

