**Final Report**  
Jaap Nieuwenhuizen, 10470611  
Programmeerproject  
Data Processing  
Gosia Migut  
21 juni 2016

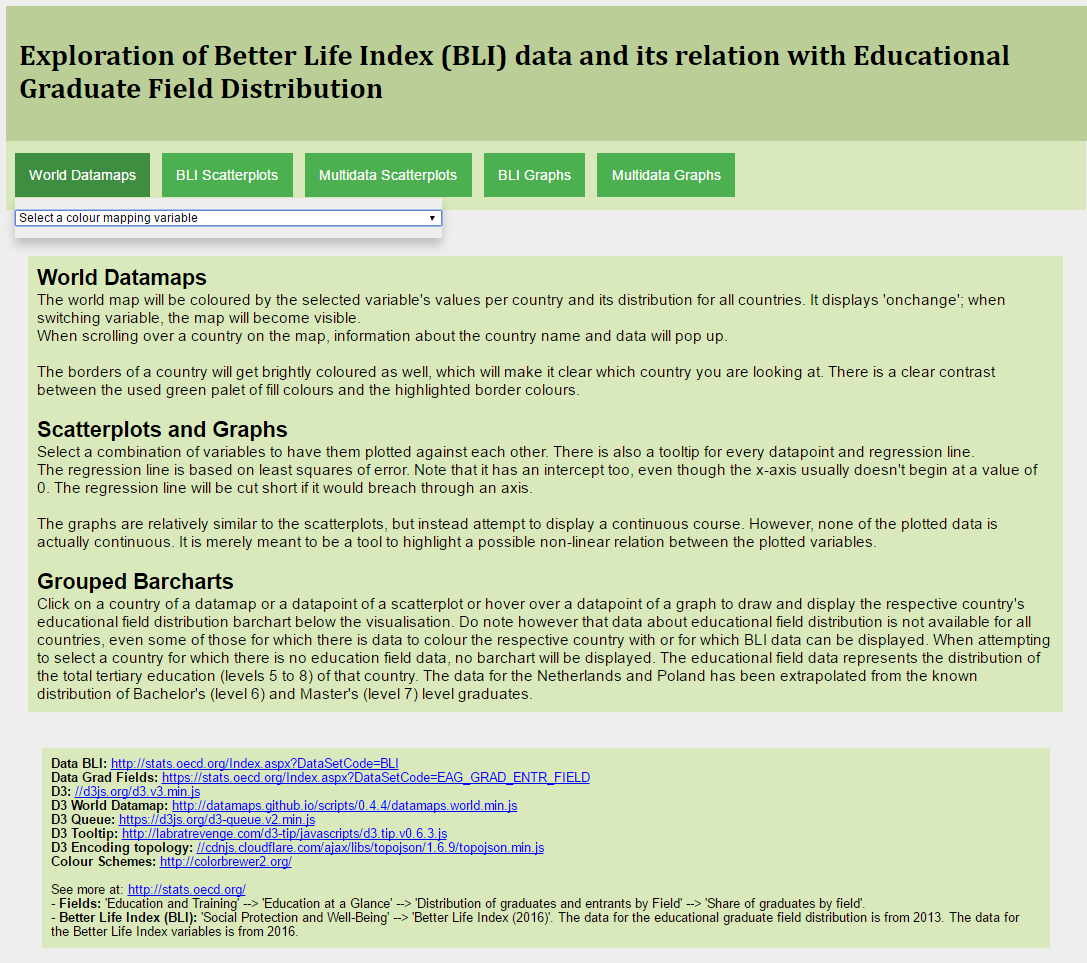
**Inleiding** Een land bestaat uit meer dan enkel koude cijfers zoals GDP en economische groei. Een Een land wordt gedefinieerd door haar bevolking, en deze bevolking bestaat uit individuen met behoeften, ervaringen, stemmingen, voorkeuren, wensen en ideeën. Hoe gaat het met de bevolking? Is een bevolking tevreden? Beschikken ze over voldoende voorzieningen; weinig luchtverontreiniging en zuiver drinkwater; goede gezondheid? Over hoeveel kamers beschikken ze thuis? Hoe degelijk bereidt de educatie van het land haar studenten voor op een gestandaardiseerde test? Is er veel werkgelegenheid in het land?

Dergelijke gegevens en meer worden bijgehouden in het kader van de ‘*Better Life Initiative*’. Dit initiatief poogt te meten wat beleidsmakers soms vergeten: wordt ons leven werkelijk beter? Meten we wel de juiste zaken om de vooruitgang van beleidsbeslissingen te peilen? Volgens de ‘*Organisation for Economics Cooperation and Development*’ (OECD) vangen macro-economische statistieken als het GDP niet voldoende in detail de leefomstandigheden en ervaringen van de gemiddelde burger. Dit was al evident in de vorige eeuw bij tijden van hoge economische groei, en is enkel duidelijker geworden met de financiële en economische crisis van tegenwoordig. Het *Better Life Initiative* dient om variabelen te meten – tezamen de zogeheten *Better Life Index* – die wel dit detail pogen te tonen; die wel de ervaringen van een bevolking pogen te illustreren. Maatschappelijke vooruitgang gaat namelijk over het verbeteren van de welzijn van burgers. Het beoordelen van dergelijke progressie vereist dus dat niet enkel naar economische variabelen en factoren gekeken wordt, maar ook naar diverse ervaringen en leefomstandigheden. Het bijhouden van de *Better Life Index* is daarmee van belang voor zowel de geloofwaardigheid van beleidsbeslissingen als voor het functioneren van de democratie. Het OECD poogt daarmee dus de brug tussen beleidsbeslissingen en democratie te versterken.

Het educationele systeem neemt een belangrijke positie in voor niet alleen economische vooruitgang, maar ook welzijnsfactoren. De opleidingsrichtingskeuze van studenten is van belang voor zowel het land als de persoon zelf. Het heeft namelijk de potentie een belangrijke levensbeslissing te zijn. Alhoewel voor de totstandkoming van complexe data als de Better Life Index en educationele keuzes meer dan enkel economische en culturele invloeden van belang zijn, zou het onderzoeken van het verband tussen deze datasets mogelijk alsnog onverwachte resultaten kunnen opleveren. Het OECD bezit over een zodanig veelomvattende databank, dat zelfs de opslag van gegevens over afstudeerrichtingen in percenten voor allerlei landen publiekelijk beschikbaar is.

Dergelijke data is mogelijk ook voor culturele interesse van meerwaarde. Zijn de vooroordelen werkelijk waar dat Fransen en Spanjaarden veelal kunstenaars worden, en Duitsers voornamelijk mekaniekers en ingenieurs zijn? Is er überhaupt een groot verschil in de educationele profiel van landen?  
  
**Het product** De gebouwde applicatie staat een gebruiker toe om twee gekozen datasets volledig te exploreren, en om te zoeken naar eventuele onderlinge verbanden tussen variabelen van deze datasets. Het vertelt geen verhaal, en behoort dientengevolge een objectieve, neutrale uitstraling te hebben, met voldoende duidelijke gebruikersaanwijzingen en mogelijkheden. De gekozen kleuren illustreren deze doelstelling: rustig en neutraal, maar ook duidelijk en intuïtief.

De applicatie beschikt over vier (verbonden) visualisaties die geselecteerd kunnen worden met dropdown menu’s: datakaarten, scatterplots, lijngrafieken en barcharts. Met deze vier visualisaties is het mogelijk om de data van een enkele dataset afzonderlijk te exploreren (m.n. met de datakaarten en barcharts), maar ook om verbanden tussen twee verschillende datasets in volledigheid te ontdekken (m.b.v. scatterplots en lijngrafieken). Een screenshot van het product zoals het wordt weergegeven als het geladen wordt is als volgt:



De datakaarten geven in één oogopslag voor alle landen de waarde van een *Better Life Index* variabele weer. De waarde op de geselecteerde BLI variabele wordt per land gebruikt om het respectievelijke land mee in te kleuren. Het weergeven van een datakaart gebeurt ‘*onchange*’: het veranderen van de waarde in het select menu van het dropdown menu van de datakaarten ertoe leidt dat de gekozen datakaart direct weergegeven wordt. Door gebruik van een intuïtief kleurenpalet kan de gebruiker zien welke landen extreem hoog, zeer hoog, hoog, gemiddeld, laag, zeer laag en extreem laag scoren op de geselecteerde variabele. Stel bijvoorbeeld dat de gebruiker wil weten welke landen een hoge mate van werkgelegenheid kennen. Deze datakaarten zijn dan wellicht het beste middel om zulke vragen te beantwoorden. De gebruiker kan vervolgens over een land van interesse scrollen om de tooltip te laten weergeven die het geselecteerde land en de exacte waarde voor de geselecteerde variabele tonen. Door te klikken op een land wordt een barchart weergegeven met data van de andere dataset. Op deze manier kan dus ook naar verbanden gezocht worden, alleen is het hier niet geschikt voor. Deze datakaarten dienen met name om alle objectieve informatie zo volledig en duidelijk mogelijk weer te geven.

De barcharts tonen net zoals de datakaarten op neutrale wijze objectieve informatie. Deze barcharts dienen om de data over het educationele profiel van een land weer te geven. Voor een geselecteerd land wordt een gegroepeerde barchart gecreëerd van het land met percentuele waarden op de y-as voor elke afstudeerrichting op de x-as. Bovendien staat naast elke percentuele balkwaarde van het land ook een gemiddelde balkwaarde als referentiepunt. Ook heeft elke balk een tooltip met het exacte percentage.

De scatterplots geven niet enkel en alleen een indicatie van de datapunten op een x- en y-as weer, maar beschikken over tooltips met de landnaam en exacte coördinaten van het geselecteerde datapunt. Als vervolgens geklikt wordt op een datapunt verschijnt de bovenstaand beschreven gegroepeerde barchart van het land. Maar wellicht nog interessanter is dat elke scatterplot ook over een regressielijn beschikt met bijbehorende regressievergelijking en verklaarde variantiewaarde. Deze regressielijn is zeer illustratief voor het onderzochte verband tussen de door de gebruiker geselecteerde variabelen. Zowel variabelen van een enkele dataset (de BLI) als van meerdere datasets (BLI en ‘*Educational Field Data’*) kunnen tegen elkaar geplot worden.

De lijngrafieken zijn vergelijkbaar met de scatterplots, maar hebben de potentie om een niet-lineair verband te openbaren. De datapunten van deze lijngrafieken worden ‘on hover’ geselecteerd: als gescrold wordt over een datapunt, dan wordt het respectievelijke datapunt automatisch geselecteerd en in tekst beschreven, maar ook de barchart wordt direct gemaakt. Wat betreft functionaliteit zijn er dus wel degelijk verschillen tussen de lijngrafieken en de scatterplots. Dit biedt de gebruiker meer mogelijkheden om verbanden naar eigen voorkeur te onderzoeken. Voor een product met een exploratieve doelstelling is dit mijns inziens een meerwaarde.  
  
**Technische beschrijving** Doormiddel van het extern oproepen van functies met de dropdown / select menu’s, kan de gebruiker een visualisatie selecteren. Aan deze functies wordt door selectie van een variabele in het menu een indicator/trefwoord meegegeven aan het menu.js script, waarmee met de ingeladen data het te visualiseren array van dataobjecten gecreëerd kan worden, en waarin ook de titel, assentekst, en elementen op identificatiewaarde zichtbaar dan wel onzichtbaar gemaakt worden. Vervolgens wordt in menu.js de functie opgeroepen waarmee de gekozen visualisatie daadwerkelijk gemaakt wordt van het doorgegeven array van dataobjecten en assentekst; menu.js is dus een soort regel- en doorgeefscript.

Voor het visualiseren van data uit twee datasets met d3.queue is het proces vergelijkbaar, maar worden twee indicatoren van de HTML-pagina opgehaald door een menu.js functie, in tegenstelling tot het doorgeven van de enkele nodige indicator aan de functie zoals dat bij visualisaties voor een enkele dataset plaatsvindt (World Datamaps, BLI Scatterplots, BLI Graphs en Grouped Barcharts). De datasets en variabelen worden pas ingeladen en opgeslagen als een visualisatie opgevraagd wordt waarvoor dit noodzakelijk is.

Met graphs.js en scatterplot.js wordt respectievelijk een lijngrafiek en een scatterplot gebouwd aan de hand van de variabelen die op het moment van oproepen geselecteerd zijn. De functie wordt opgeroepen als geklikt wordt op de ‘*submit*’-knop van de dropdown van de gekozen visualisatie.

In graphs.js is een interactief element verwerkt, waarbij twee verschillende labels getoond wordt bij een automatisch geselecteerd datapunt. Een datapunt wordt namelijk elke keer geselecteerd dat de muis beweegt over de zogeheten ‘overlay’, en is altijd het punt dichtstbijzijnd aan de coördinaten van de muis. De crosshairs zullen naar dit punt wijzen, het datapunt zal dankzij de css-file van kleur veranderen en een barchart zal gemaakt worden van het land waarvoor dit datapunt is.

In scatterplot.js wordt een regressielijn aangemaakt die afgekapt wordt indien deze een as dreigt te snijden. De lijn wordt eerder getekend dan de datapunten om de lijn op de achtergrond te houden. Elk datapunt heeft een eigen tooltip gemaakt met de library die het land en de coördinaten weergeeft. Ook verandert het datapunt van kleur als er over gescrold wordt. Met een onclick-functie kan een barchart aangemaakt worden.

In maps.js is een standaard model om wereldkaarten aan te maken, inclusief een functie die alle landen/objecten kleurt aan de hand van de waarde van het respectievelijke land relatief aan de verdeling van alle waarden van de gekozen variabele. De klassen die gegenereerd worden met die kleurenfunctie worden in de legenda weergegeven. Bij het creëren van de map worden ook de border- en fill-colours meegegeven.

De kleurenfunctie in maps.js neemt twee argumenten op: een enkele waarde (van het gekozen land) en de gehele serie aan waarden (voor alle landen) van de gekozen variabele. Er wordt vervolgens gezocht naar de waarde van het gekozen land binnen de gehele serie. Als bijvoorbeeld blijkt dat de variabele hoger is dan de eerste drempelwaarde, maar lager dan de tweede drempelwaarde, dan krijgt het land met die variabele een ‘veryLow’-kleurenlabel, en wordt dit land dus als zeer laag scorend geacht op de geselecteerde variabele. De doorgegeven serie is altijd gesorteerd van laag naar hoog, en de drempelwaarden zijn gebaseerd op indices relatief aan de totale lengte van de serie. De eerste drempelwaarde is bijvoorbeeld een waarde hoger dan de afgeronde index 1/10de van de totale lengte van de serie. Deze drempelwaarde nemen vervolgens steeds toe. Deze kleurenfunctie van maps.js wordt opgeroepen in menu.js op het moment dat het nodige array van dataobjecten gecreëerd wordt.

De functie in menu.js voor de grouped barcharts worden in zowel maps.js als scatterplot.js opgeroepen met een zogeheten ‘*onclick event*’: de barchart van een aangeklikt land wordt gemaakt doormiddel van het doorgeven van de naam van een land als deze in een andere visualisatie aangeklikt wordt. In graphs.js wordt het land gevonden van de datapunt waar de ‘focus’ op ligt. Elke keer als de muis beweegt wordt de *focus circle* bepaalt en wordt een barchart gemaakt. In menu.js wordt vervolgens de vorige barchart (als er überhaupt één is) verwijderd, en wordt de data ingeladen waarmee het array van barchart dataobjecten gemaakt kan worden. Op basis van het doorgekregen land kan in de dataset gezocht worden naar een object met de afstudeerrichting (field) en de bijbehorende percentuele waarde (Value) die als zodanig opgeslagen worden in een nieuw object, en toegevoegd wordt aan de array. Aan ditzelfde array worden ook de gemiddelden per afstudeerrichting toegevoegd. Vervolgens worden de groepsnamen (altijd Country en Average) gedefinieerd in een nieuwe variabele en wordt deze variabele toegevoegd aan elk object. Hierna wordt de functie aangeroepen waarmee met het land, de array en de groepsnamen de barchart getekend kan worden.

Overigens worden de gemiddelden per afstudeerrichting in een afzonderlijk script berekend: field\_averages.js. Dit omdat het nogal omslachtige en relatief ingewikkelde code is. In dat script wordt namelijk met twee dictionaries/objecten en één array informatie verzameld van de som aan percentages voor een field en de lengte van de som voor datzelfde field, en de namen van alle afstudeerrichtingen om als index te gebruiken. Met als doel om van het tweede dictionary een field-gemiddelde pair als key-value pair te creëren en op te slaan.

In field\_barchart\_group.js is het script om aan de hand van drie argumenten een barchart met tekst aan te maken. In dit script zit namelijk de createBarchart functie. De barchart wordt alleen weergegeven als het eerste object van de array ook werkelijk data bevat. Indien dit niet het geval is, dan zal de barchart niet verschijnen. Voor landen zonder data wordt dus geen barchart weergegeven.  
  
**Proces en uitdagingen** De belangrijkste implementatie was het verbinden van een geselecteerde indicator op de menuknoppen aan het script waarmee een visualisatie gemaakt moest worden. Dit is niet alleen eleganter, maar ook noodzakelijk, omdat beide datasets veel variabelen bevatten en dus nauwelijks alle combinaties hard-coded geïmplementeerd kunnen worden. Dit werd wel in eerste instantie geprobeerd. In de tweede week heb ik dan ook mijn visualisatie hard-coded weten te implementeren, met enige bugs/lelijkheden en zonder dropdown menu’s.

De eerste poging tot automatisering leidde ertoe dat ik zeer veel knoppen aanmaakte die elk dezelfde functie opriepen met een andere hard-coded indicator en tekst. Ook nadat ik dropdown menu’s geïmplementeerd had, had ik een dergelijke button/knop om een functie aan te roepen. Dit heb ik nog steeds voor de datakaarten, maar was niet mogelijk voor het plotten van twee variabelen. Alhoewel er aan een functie wel een string-input mee gegeven kon worden, was het niet mogelijk deze string als twee afzonderlijke variabelen te herkennen, dus moesten deze in menu.js gesplit worden en opgeslagen worden in afzonderlijke variabelen. Het huidige concept is echter zeer geschikt voor dataexploratie.

Een nieuwe datakaart was noodzakelijk om een legenda met d3 toe te voegen, en bovendien was mijn oorspronkelijke kaart tamelijk lelijk. Ook het extrapoleren van data voor de landen waarvoor dit mogelijk was (Nederland en Polen) was niet eenvoudig.

Om alle visualisaties tot één geheel te verwerken zonder dat ze met elkaar interfereren, en met zo efficiënt mogelijke code, is natuurlijk altijd een uitdaging. Specifiekere uitdagingen waarmee ik te maken kreeg waren bijvoorbeeld kleine bugs zoals de menuopties achter de datakaarten en tooltips die niet verschenen of met elkaar interfereerden, regressielijnen die de as sneden, maar ook en voornamelijk data-analytische problemen zoals het berekenen en verwerken van gemiddelden per studierichting in de barchart.

**Verdediging** Naar mijn mening is het huidige concept zeer geschikt voor dataexploratie. Helaas leveren de gebruikte datasets echter weinig interessante resultaten op. Ik denk echter dat dit product naar tevredenheid tegemoet komt aan de doelstellingen.

Als ik opnieuw een project zou doen, dan had ik graag gepoogd een beter en nuttiger product te leveren door geheel andere data te gebruiken. Ik kende in de eerste week een behoorlijke stroeve start, en twijfelde behoorlijk aan mijn ideeën. Want alhoewel veel mogelijk is met datavisualisatie, had ik relatief weinig inspiratie, en kon ik ook niet goed inschatten wat realistisch mogelijk is met mijn programmeercapaciteiten. Achteraf denk ik echter mijn progressie en capaciteiten enigszins te hebben onderschat.

Verder had ik weinig idee van alle visualisatiemogelijkheden. Want ook had ik graag nog meer nieuwere visualisaties willen implementeren. Andere soorten visualisaties zouden echter weinig praktisch nut hebben gehad voor het doel dat ik voor ogen had met dit project.

Als ik meer tijd had, dan zal ik die voornamelijk hebben besteed aan de opmaak: mooiere scatterplots en grafieken, en vooral tooltips/interactieve selectie van datapunten. Een ander probleem dat ik niet heb weten op te lossen is het feit dat datapunten soms exact hetzelfde zijn, waardoor sommige datapunten niet meer geselecteerd kunnen worden. Voor zover ik weet is dit echter het enige probleem dat niet goed opgelost kon worden.

Over het algemeen ben ik tevreden met een geheel afgerond en nagenoeg foutloos – alhoewel eenvoudig – product, dat met name door mijn inzet en progressie in de tweede en derde week van dit project tot stand is gekomen.