Logboek Groep 31
Data Analysis and Visualization

Door:
Chris Al Gerges (11727845),
Diederik Salimans (11913894),
Dion Custers (11804122),
Geert Rien Bakker (10548602).

Week 1:

We hebben na overleg met Nora besloten dat de Global Food Prices Database de primaire dataset voor ons project zal zijn. We nemen hiernaast de dataset van de United Nations over vluchtelingenstromen als secundaire dataset en gaan deze met elkaar gaan vergelijken.
Voor het cleanen van beide datasets hebben we de groep opgesplitst in twee groepjes van twee. Het groepje bestaande uit Chris en Geert is aan de slag gegaan met het opschonen van de GFPD en het groepje van Diederik en Dion met het opschonen van de vluchtelingen-dataset. Ondanks dat we ons gesplitst hadden in twee groepen werden de grote keuzes voor het cleanen van de dataset gezamenlijk gemaakt.

Voor het cleanen van de Global Food Prices Database besloten we als eerste om de maandelijkse data om te zetten naar jaarlijkse data. Dit hebben we gedaan omdat de data in de dataset over vluchtelingen in jaren gegeven is en het anders lastig zou worden om de twee datasets met elkaar te kunnen vergelijken. We besloten hierbij om de jaren waar minder dan 4 maanden data over was te verwijderen uit de dataset, omdat dit geen goede representatie voor het gehele jaar zou kunnen zijn. Hierna besloten we om alle regio's binnen landen bij elkaar te voegen om op die manier data te hebben per land in plaats van per regio. Ook dit kwam vooral door de vluchtelingen-dataset die met landen werkte en niet met regio's binnen landen. Na deze maatregelen was de dataset al een stuk overzichtelijker, maar door het gebrek aan veel data voor het jaar 2000 en meerdere producten waar maar een enkele jaren data van was, was de GFPD nog steeds niet echt bruikbaar. Daarom dat we toen besloten hebben om alleen de data tussen 2001 en 2016 te behouden. Daarnaast zijn de producten waar minder dan vier jaar data over is ook verwijderd, om op deze manier significantie te behouden.

Ook bij de vluchtelingen-dataset hebben we de data tussen 2001 en 2016 behouden en de rest verwijderd. We hebben hiernaast twee kolommen toegevoegd die aangeven hoeveel vluchtelingen er per jaar een land in en uit zijn gegaan. We hebben daarna alle landen die niet in beide datasets aanwezig zijn uit de datasets verwijderd, op deze manier wisten we zeker dat we van elk land dat overbleef zowel data uit de GFPD als uit de vluchtelingen-dataset zouden hebben. Na deze maatregelen vonden wij de datasets clean genoeg voor analysatie.

We hebben de datasets vooral gecleand met behulp van pandas, maar hebben daarnaast soms ook gebruik gemaakt van LibreOffice. Dit hebben wij gedaan omdat het voor sommige simpele kolommen veel simpeler en sneller was om die via Libre te maken dan via pandas.

Week 2:

We zijn deze week begonnen met het creëren van een enorme hoeveelheid grafieken. We hebben ons hiervoor weer opgesplitst in twee groepen. De groep van Diederik en Chris heeft zich beziggehouden met het schrijven van een python programma dat voor elk land een grafiek maakt met daarin alle producten die zich in dat bevindt. Er wordt hierbij gekeken naar de prijs van dat product per jaar. De groep van Dion en Geert heeft zich bezig gehouden met een programma dat voor elk product kijkt welke landen dit product hebben, dan voor elk van deze landen een grafiek creëert en daarna checkt of andere landen die in de regio zitten van dat land ook dit product hebben. Is dit het geval, dan wordt deze ook toegevoegd aan de grafiek van het oorspronkelijke land. We hebben dit allebei via bokeh en pandas in python gedaan.

Om er voor te zorgen dat het tweede programma kan werken, moesten we eerst besluiten wat we verstaan onder regio's. We hebben er uiteindelijk voor gekozen om elk land zijn eigen regio te geven. Hierbij werd de regio van een land gezien als het land zelf plus zijn buurlanden. We kwamen erachter dat het voor ons vrijwel onmogelijk was om op een simpele manier de buurlanden op te vragen van een land via Pandas en csv. We kregen het niet voor elkaar om een nieuwe kolom te creëren die bestond uit meerdere lijsten en om daarna deze lijsten bruikbaar te krijgen. Daarom dat we er uiteindelijk voor gekozen hebben om het handmatig toe te voegen aan het python programma. Dit zorgde er daarentegen voor dat er soms lege grafieken ontstonden vanwege kleine typfouten, waardoor het uiteindelijk veel tijd kostte om dit te perfectioneren. We kwamen er uiteindelijk achter dat het te veel tijd zou kosten om voor elk product deze grafieken te creëren, waardoor we een simpele versie hebben gemaakt. Hierbij konden we zelf aangeven voor welk individueel product we deze grafieken wouden zien.

Nadat deze twee programma's beide werkten, kwamen we erachter dat er veel dezelfde producten niet precies hetzelfde zijn. Zo bleken er heel veel verschillende soorten rijst en graan te zijn, maar door de soms toch grote prijsverschillen konden we deze niet samenvoegen tot één product. Hierdoor kregen we heel veel grafieken met maar een enkele lijn, en waar we dus nauwelijks iets aan hadden. Ook bleken de grafieken die we met meerdere lijnen kregen niet betrouwbaar, vanwege het gebruik van verschillende valuta in verschillende landen.

We hebben toen donderdag besloten met Nora om een extra Currency Exchange dataset toe te voegen. Deze dataset zorgt ervoor dat we jaarlijks elk valuta om kunnen zetten naar USD. Vanwege de vervelende opmaak van deze dataset daarentegen kostte het veel tijd om deze op een handige manier toe te kunnen voegen aan de GFPD. We hebben de data ook in deze dataset weer van maandelijks naar jaarlijks omgezet. Voor sommige valuta was er onduidelijke data, waardoor er sommige landen uit de GFPD zijn gehaald. Uiteindelijk heeft dit wel tot betere grafieken geleid en konden we juiste conclusies trekken uit de gekregen resultaten. We hebben daarna ook nog besloten om gelijk alle soorten meeteenheden om te zetten naar een hoofdmeeteenheid om nog perfectere resultaten te krijgen. We besloten om de kilogram te nemen als de hoofdmeeteenheid.

Week 3:

We begonnen deze week met een kleine achterstand door de aanpassingen die we aan het einde van de vorige week nog wouden toepassen aan de GFPD. We hebben deze week besloten ons uiteindelijk volledig te gaan focussen op de landen uit het continent Afrika. Voor deze landen hadden we de meeste data in alle datasets en waren regio's het beste gedefinieerd, waardoor de gekregen resultaten en grafieken het meest duidelijk waren.

Ook is Chris veel bezig geweest met het perfectioneren van de grafieken die wij als groep het interessants vonden. Zo heeft hij ze vaak interactief gemaakt en ze zo overzichtelijk mogelijk gemaakt, zodat deze grafieken meteen gebruikt konden worden voor de uiteindelijke site. De rest van de groep heeft deze week vooral heel gericht gekeken naar verschillende producten die op elkaar leken en naar bepaalde regio's die een hoge correlatie bevatten.

Aan het einde van de week zijn Diederik en Geert begonnen met het schrijven van het technisch rapport en hebben zich daarin vooral gefocust op de methode.

Week 4:

We hebben ons deze week vooral gefocust op het afronden en perfectioneren van het project. Maandag hebben Dion, Diederik en Geert vooral besteed aan het schrijven van de inleiding en aan het kiezen van de uiteindelijke resultaten en conclusies voor de deelvragen. Dit zodat er dinsdag aan de resultaten en discussie gewerkt kon worden. Chris heeft zich volledig gericht op het perfectioneren van de grafieken.

Diederik en Geert hebben dinsdag de methode geperfectioneerd en afgemaakt en hebben de resultaten en discussie volledig geschreven, zodat de eerste versie van het rapport die dag nog naar Nora gestuurd kon worden. Dion en Chris hebben gekeken naar hoe Github Pages precies werken en hebben enkele dingen aangepast zodat de site in het verloop van de week makkelijk gemaakt kan worden.

We zijn woensdag begonnen met een ruwe schets van de site, maar zijn vooral bezig geweest met het maken en afmaken van de presentatie. Hiernaast hebben we het logboek en Github opgeknapt, zodat deze beide voor donderdag klaar waren.

Donderdag heeft Diederik de eerste versie van het rapport nog deels aangepast nadat we feedback hadden gekregen van Nora en heeft Geert het rapport daarna overgezet naar LateX. Chris heeft zich vooral gefocust op het afmaken van de site. Dion heeft het logboek afgemaakt en geperfectioneerd.

Conclusie:

Het Data Analysis and Visualization project is uiteindelijk goed verdeeld geweest over ons alle vier. We hebben ons in het begin alle vier gefocust op het programmeren en verzinnen en hebben in de laatste twee weken ons individueel wat meer gefocust op onze sterke punten. Zo heeft Chris zich in de laatste weken vooral gefocust op het maken van grafieken, omdat het duidelijk bleek dat hij hier goed en efficiënt in was. De rest heeft in deze tijd ook wel enkele simpele grafieken gemaakt, maar heeft zich vooral bezig gehouden met de resultaten die we uit de grafieken van Chris konden halen en uiteindelijk ook met het schrijven van het rapport en het logboek. Ook werden alle beslissingen eerst

met iedereen van de groep besproken, zodat iedereen ervan op de hoogte was en niemand werd buitengesloten.

We zijn hiernaast bijna elke werkdag bij elkaar gekomen op het Sciencepark van 11.00 tot 15.00 ~ 16.00. Als iemand niet aanwezig kon zijn op een bepaalde dag of vanwege een afspraak eerder weg moest, werd dit van te voren door iedereen goed aangegeven en werd dit door de rest van de groep ook altijd geaccepteerd. Dit zorgde ervoor dat er voortdurend een relaxte en fijne sfeer was in de groep.