Process Book

Jardenna, Janne, Jonne, Julius

June 2018

Week 1

Maandag 04/06/18

Preprocessing

De kolom met "country ID's" is uit de dataset verwijderd. deze kolom was overbodig omdat de landnamen ook al uniek waren dit was handig omdat het de dataset zou comprimeren. Alle valuta, die geen stabiele wisselkoers hebben met de US Dollar zijn uit de database verwijderd. We wisten dat we later landen met elkaar zouden vergelijken, en dat de prijzen dus genormaliseerd zouden moeten worden.

Dinsdag 05/06/18

Steden met dezelfde naam zijn gedisambigueerd door de afkorting van het land achter de stadsnaam te zetten. Dit zou het opvragen van data per stad mogelijk maken en ervoor zorgen dat we niet per ongeluk dubbele data zouden terugkrijgen bij een query op een stad. Jaar en maand kolom is gecombineerd tot één kolom dit zou de dataset verder comprimeren en bovendien later de omzetting naar daytime (voor het plotten van grafieken) makkelijker maken.

Woensdag 06/06/18

Kolom met regio's zijn verwijderd deze kolom gaf nauwelijks extra informatie: de meeste regio's bevatten maar informatie van één of twee steden, waardoor de kolom met steden al voldoende zou zijn. Bovendien wisten we dat we later toch geen provincies met elkaar zouden willen vergelijken.

Donderdag 07/06/18

De rijen van producten en steden waarvan meer dan twee opeenvolgende maanden aan data bij misten zijn verwijderd. We wilden later lijnen met elkaar vergelijken, en daarvoor hadden we onafgebroken lijnen nodig. Gaten van twee maanden vonden we nog verantwoord om met lineaire regressie in te vullen.

Week 2

Maandag 11/06/18

Gaps van één of twee maanden zijn ingevuld met lineaire regressie. Zodat er meer nuttige vergelijkingen kunnen worden gedaan. Met behulp van een andere dataset zijn landen samengevoegd tot sub-regio' zodat het ook makkelijker wordt om landen in dezelfde regio te vergelijken.

Dinsdag 12/06/18

Voor een van onze deelvragen onderzoeken we of er een verband bestaat tussen voedselprijzen binnen gelijksoortige regio's; het was dus wenselijk data te kunnen selecteren op sub-regio's. Alle eenheden zijn geconverteerd naar KG en Liters dit was wenselijk zodat prijzen beter met elkaar vergeleken konden worden. Rijen met Quartilla en (goat) head als eenheid zijn verwijderd.

Woensdag 13/06/18

We konden geen vaste conversion-rates vinden voor deze eenheden, en ze konden dus moeilijk op betrouwbare wijze vergeleken worden met andere eenheden. Labour Prices' zijn als producten verwijderd, omdat dit geen voedselproduct was. We wilden immers voedselprijzen met elkaar vergelijken, en hadden geen deelvragen met betrekking tot arbeidskosten.

Donderdag 14/06/18

Dataset is gesplitst op seller en national average omdat bij verschillende verkopers de prijzen op verschillende manieren tot stand komen (door grotere inkoop mogelijkheden bijvoorbeeld), zouden vergelijkingen tussen verschillende kopers ons doel vermoeilijken; door de dataset op te splitsen konden we nu verschillende verkopers apart analyseren. De dataset is genormaliseerd op koopkracht per land. Met behulp van een andere dataset is voor elk land per maand de GDP achterhaald. Door deze te delen door de productprijs ontstond een maat die aangaf 'betaalbaar' dat product in dat land is.

Vrijdag 15/06/18

Data-punten waarvoor geen GDP bekend was zijn verwijderd. dit zou de dataset verder comprimeren en daardoor het oproepen van data later versnellen.

Voor alle productprijzen over de tijd is de afgeleide berekend en toegevoegd in een nieuwe 'gradient' kolom door de afgeleide te pakken zouden we prijsveranderingen tussen producten beter kunnen vergelijken

Week 3

Maandag 18/06/18

Clustering

Er is een eigen versie van het K-means algoritme geïmplementeerd om het clusteren van data mogelijk te maken. De gegeven versie van SK-Learn is overwogen. Echter aangezien dit algoritme minder goed omgaat met missing values is er voor gekozen om de eigen versie van K-means te gebruiken. Namelijk bij het clusteren werden soms producten vergeleken over verschillende tijdsspannen waardoor er dus als het ware missing values ontstonden. In onze versie van K-means worden missing values simpelweg genegeerd. Dit betekend dat deze dus ook niet worden meegeteld als datapunten bij het gemiddelde waardoor het gemiddelde een betere representatie geeft van de beschikbare datapunten.

Als Exploratory Data Analysis hebben het K-means algoritme toegepast op de producten van India en Oekraïne zodat clusters ontstonden. We hebben deze landen gekozen omdat deze een grote variatie aan producten hebben waarvan we de classificaties zelf konden beoordelen.

Dinsdag 19/06/18

Bedenken van deelvragen

Naast de drie deelvragen gegeven op canvas hebben we nog een aantal andere vragen bedacht die ons interessant leken om te onderzoeken. Hiermee hebben we gelijk ook bepaald welke invalshoek we zullen hebben bij onder andere het geven van de presentatie. Twee van onze deelvragen gaan over de olieprijzen omdat wij verwachten dat er een correlatie bestaat tussen de voedselprijzen en olieprijzen. Namelijk voor het vervoeren van voedsel is brandstof nodig. Als de brandstof duurder wordt, wordt het ook duurder om een product te leveren. Vandaar dat wanneer de olieprijzen toenemen ook de voedselprijzen toenemen.

De andere deelvragen gaan over het verband tussen de voedselprijzen en het sterftecijfer in een land en de vluchtelingen-stromen. Dit leek ons interessant omdat we verwachten dat als de voedselprijzen stijgen dat in arme landen dan ook het sterftecijfer toeneemt. Bovendien kan het stijgen van de voedselprijzen ook een oorzaak zijn voor het toenemen van de vluchtelingen-stromen.

Extra datasets

Naast de extra datasets van het bruto nationaal product en valuta die in week 1 zijn uitgekozen is er naar aanleiding van de bedachte deelvragen gezocht naar een paar andere datasets. Namelijk die van vluchtelingenstromen en de sterftecijfers binnen een land.

Woensdag 20/06/18

Correlatie

Voor het beantwoorden van de deelvragen is het handig om de mate van correlatie tussen twee factoren te kunnen bepalen. Dit is gedaan met behulp van regressie. De methodes die zijn overwogen zijn Pearson en de Spearman's rank. De Spearman correlatie coëfficiënt is gebaseerd op de geordende waardes voor elke variabele in plaats van de ruwe data. Vandaar dat de Spearman's rank vooral wordt gebruikt wanneer er sprake is van ordinale variabelen. Een variabele bij onze dataset is tijd, wat een continue variabele. Ook zijn we vooral geïntereseerd in de correlatie tussen de prijzen van twee producten. Vandaar dat Spearman voor ons niet de juiste keuze was. De Pearson correlatie evalueert de relatie tussen twee continue variabelen. Vandaar dat er is gekozen om Pearson te gebruiken.

Donderdag 21/06/18

Visuele analyse

Er is een begin gemaakt aan het beantwoorden van de deelvragen. We gebruiken hiervoor de bokeh plots, Pearson en k-means. Allereerst analyseren we de bokeh plots en kijken we waar veel data punten zijn zodat de resultaten representatief zijn. Daarna bereken we de correlatie coëfficiënt en kijken we met behulp van K-means welke clusters gevormd worden.

Vrijdag 22/06/18

Technisch rapport

Vandaag is vooral gewerkt aan een opzet van het technische rapport. Er wordt in LateX gewerkt omdat dit het verslag op een overzichtelijke universele manier weergeeft. Voor het technische rapport is er voor gekozen om voor alle deelvragen een aparte hypothese, opzet en verwachting te maken. Ook de methode resultaten en discussie zal geschreven worden aan de hand van de gescheiden deelvragen. Dit hebben we gedaan omdat er zo het best overzicht gehouden kan worden. Ook is zo duidelijk is welke stappen we hebben ondernomen om bepaalde resultaten te verkrijgen. Bovendien is verder gegaan met het beantwoorden van deelvragen dit is wederom gedaan met behulp van K-means en de correlatie coëfficiënt en bokeh-plots.

Week 4

Maandag 25/06/18

Webpagina

Er is gewerkt aan de visualisatie en webpagina. De belangrijkste beslissingen die zijn gemaakt zijn hoe we ons dashboard eruit willen laten zien. Er is gekozen voor een interactief dashboard zodat bij de presentatie elke deelvraag ondersteund kan worden met behulp van de website. Hierbij kan de data van elk land in een bepaalde periode worden opgevraagd. Ook kan er worden geselecteerd op regio en op product en wordt dit vervolgens op een overzichtelijke manier gevisualiseerd.

Dinsdag 26/06/18

Vandaag is voornamelijk gewerkt aan openstaande processen waaronder het beantwoorden van de deelvragen, het technisch rapport en de website. De belangrijkste beslissingen voor de website waren dat er een wereldkaart is gemaakt waar de landen oplichten wanneer de data ervan geselecteerd worden. Ook kan worden aangepast welke periode gevisualiseerd wordt. Vervolgens wordt hiervan de voedselprijzen gevisualiseerd in een grafiek en de gradiënt. Ook is de correlatie weergegeven in een tabel. De cellen waar hoge correlaties in staat zijn donkerder gekleurd zodat er snel een goed overzicht kan worden verkregen van de mate van correlatie tussen producten.

Woensdag 27/06/18

Vandaag zijn keuzes gemaakt voor het visualiseren van data. Er is overwogen om een tabel te gebruiken om clusters te visualiseren of een grafiek waarbij de lijnen die bij een cluster behoren dezelfde kleur hebben. Echter is elke datum bij ons een dimensie. T-SNE is een goede methode om *high dimensional data* te visualiseren. Vandaar dat we ervoor hebben gekozen om dit gebruiken.

Donderdag 28/06/18

Het sterftecijfer van het land is in een lijngrafiek weergegeven op de website. Dit is nodig voor het beantwoorden van een van onze deelvragen. Het zelfde geldt voor het aantal vluchtelingen per land. Ook kan op de kaart de vluchtelingen stromen worden weergegeven zodat er ook snel kan worden gezien hoe alles in elkaar zit.

Presentatie

Vandaag is een ook een groot deel van de presentatie voorbereid. Er is er voor gekozen om niet te diep in te gaan op alles omdat we maar 5 minuten hebben

voor de presentatie. We hebben besloten om eerst een kort overzicht te geven van de website en daarna onze deelvragen te beantwoorden aan de hand van de website en nog wat ruimte over te laten voor eventuele vragen.

Vrijdag 29/06/18