Process Book

Project: Analysis and Visualisation of World Food Prices
Groepsbegeleider: Willemijn Beks
Uitgevoerd door: Darius Barsony, Ellen Bogaards,
Joshua de Roos en Roos Vervelde
Datum: 28-06-2018

Week 1

Actiepunten

- pre-processing WFP dataset
- vinden van vluchtelingen data en preprocessing van deze data
- vinden van BBP data en preprocessing van deze data
- vinden van wisselkoers data en preprocessing van deze data
- vinden van populatie data en preprocessing van deze data
- vinden van neerslag data en preprocessing van deze data

Gemaakte beslissingen en belangrijke inzagen

- World Food Prices als dataset gebruiken
- We zijn begonnen door met elkaar afspraken te maken over wat we van het vak verwachten en wanneer iedereen tijd moet vrijhouden voor dit vak. Ook hebben we andere verplichtingen met elkaar gedeeld zodat iedereen van elkaar weet wanneer hij/zij bijvoorbeeld moet werken. Zo zien deze afspraken eruit:

  Algemeen:
  - Doel is een 8.5
  - Dinsdag en donderdag is iedereen vrij voor dit vak
  - Communicatie verloopt via de app

  Planning:
  - Joshua werkt maandag vanaf 15:00 en vrijdag vanaf 15:00
  - Ellen werkt woensdag vanaf 13.30
  - Roos werkt doordeweeks niet
  - Darius werkt donderdag vanaf 15 uur (tot 2 weken)

  Week 1:
  - Joshua heeft essay
  - Ellen moet vanaf vrijdag op festival werken
  - Roos is vanaf vrijdag op zeilweekend
  - Darius moet in het weekend werken.
  - Conclusie: voor vrijdag alles af!

- Welke extra data we erbij willen zoeken. Deze hebben we gekozen na een klein vooronderzoek in wetenschappelijke literatuur en krantenberichten:
  - de regenval van de landen uit WFP-bestand per jaar
  - de populatiegrootte van de landen uit WFP-bestand per jaar
  - wisselkoers van de landen uit WFP-bestand per jaar
  - BNP van de landen uit WFP-bestand per jaar

    - Opgevangen vluchtelingen van de landen uit WFP-bestand per jaar
- Alle eenheden van de goederen proberen zoveel mogelijk te converteren naar standaard eenheden zoals ‘KG’ en ‘L’. Dan kunnen producten beter vergeleken gaan worden.
- In het WFP-bestand stonden van bepaalde goederen veel verschillende soorten. Deze soorten van één goed willen we samenvoegen omdat we denken dat de soorten voor onze analyse niet interessant zijn. Dus hoge/lage kwaliteit rijst is bijvoorbeeld samengevoegd tot het goed ‘rijst’.
- Het CSV databestand over de voedselprijzen van de wereld (WFP) is in een Pandas tabel gezet, zodat met dit bestand makkelijk gewerkt kan worden
- Extra databestanden erbij gezocht als uitbreiding van het WFP-bestand. Uit deze extra databestanden zijn alleen de landen gekozen die ook in het WFP-bestand zijn.
- Voor het populatiebestand is de populatie van het jaar 2017 uit een andere bron erbij gekomen dan de jaren ervoor, dit kan inconsistenties geven aangezien het bestand nu afkomstig is uit twee verschillende bronnen.
- De namen van landen uit de extra bestanden zijn handmatig veranderd in dezelfde naam zoals het in het WFP-bestand is gegeven. Hierdoor kunnen de landen uit de verschillende bestanden met elkaar vergeleken worden.
- Sommige gebieden misten de prijs van goederen. Hiervoor is bij alle districten het nationaal gemiddelde ingevoerd.

Week 2

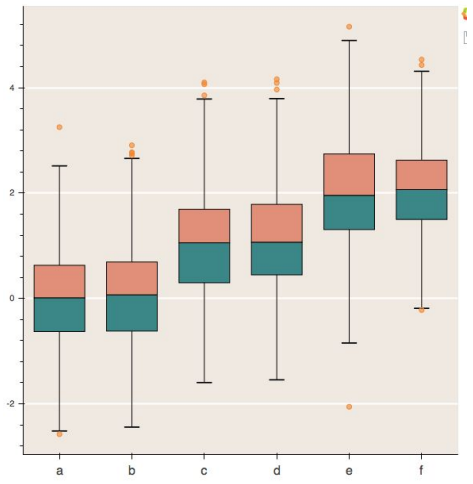
Actiepunten

- Berekenen van correlaties van alle goederen per land voor alle landen. Hieruit kunnen dan opvallende gevallen worden gefilterd voor diepere analyse.
- Maken van boxplots die de prijsverandering van rijst door de jaren heen van alle landen visualiseert.
- Maken van staafdiagram die BNP plot van alle landen in het WFP bestand om te kijken of hier interessante gegevens uitkomen.
- correlaties maken van alle goederen met BNP van bepaalde landen in WFP-bestand ten opzichte van de gemiddelde prijs.

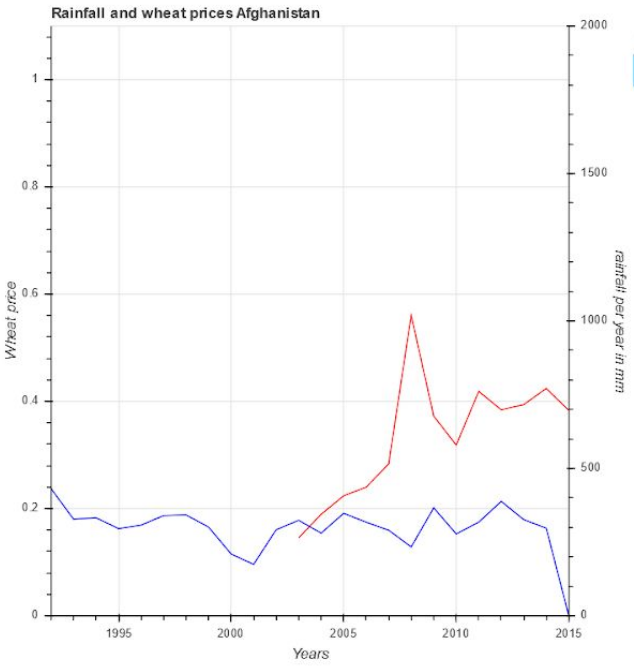
Gemaakte beslissingen en belangrijke inzagen

- Deze week al beginnen met een aantal visualisaties omdat we zo meer inzicht krijgen in de data. Nu hebben we nog weinig overzicht hierover.
- Vooral focussen op correlatiecoëfficiënten omdat we vooral numerieke data gebruiken.
- Vooral focussen op multivariate analyse omdat prijzen alleen in verhouding tot elkaar vergeleken kunnen worden.
- Preprocessing zo snel mogelijk afmaken. We merkten dat dit meer tijd kostte dan verwacht omdat we veel verschillende datasets hadden. Hierdoor hebben we in week 2 ook nog preprocessing gedaan. Vooral het omzetten van alle prijzen naar dollars was een grote klus. Dit wilden we per se doen zodat de producten allemaal in verhouding met elkaar vergeleken konden worden.

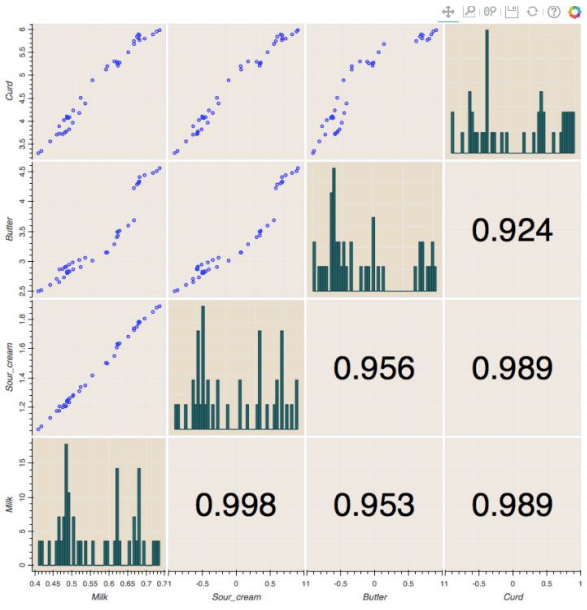
- Om de prijsverandering van rijst door de jaren heen te visualiseren is voor een boxplot gekozen omdat deze makkelijk weergeeft welke landen gelijke verdelingen hebben van de prijsverandering en welke landen sterk verschillen. Het product rijst is gebruikt bij deze analyse omdat rijst het meest gemeten product is en zo konden we dus wat inzicht krijgen in onze data. Onderstaande afbeelding is een voorbeeld-boxplot dat gebruikt is voor het maken van deze boxplots.



- Alle prijzen in het WFP-bestand zijn omgezet naar dollars, zodat de prijzen goed met elkaar vergeleken kunnen worden.
- In het regenval-bestand was de regenval gegeven in maanden, dit is omgezet in jaren. Dit komt omdat regenval in maanden niet relevant was voor het WFP-dataset.
- Regenval geeft geen correlatie met de prijs van voedsel maar dit is logisch omdat het effect van regenval waarschijnlijk niet in dezelfde maand zichtbaar is in de prijs.
- Als je echter per jaar kijkt, dan is er wel correlatie te vinden bij één product: Brood. Voor brood is er in veel landen een correlatie van boven 0,6 en onder -0,6. Wat heel vreemd is, want ook graan heeft een correlatie rond de nul. Hier hebben we nog geen verklaring voor.
- Er zijn plots gemaakt van regenval en graan prijzen voor verschillende landen, met dus twee y-assen voor deze waarden. De x-as gaf de jaren aan. Dit hebben we gedaan om te kijken of er toch een verband van regenval zichtbaar is, die niet in de correlaties te vinden waren.



- Scatter matrix gemaakt die de scatterplots van de prijs van boter, melk, zure room en wrongel, de verdelingen van de producten en de correlaties tussen de producten weergeeft. In onderstaande afbeelding is deze voorlopige scatter matrix te zien. Deze scatter matrix hebben we gemaakt om te kijken of de overeenkomende ingrediënten in deze producten hoge correlaties geven. Deze scatter matrix kan gebruikt worden voor de beantwoording van deelvraag 1.



Week 3

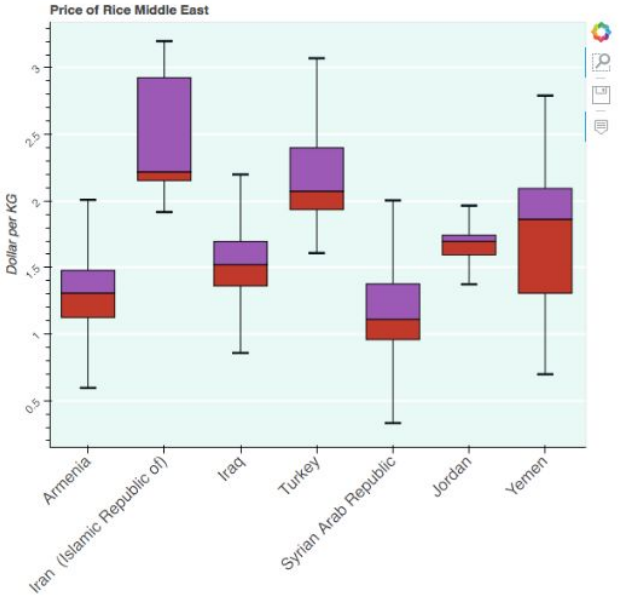
Actiepunten

- Clustering maken van alle databestanden bij elkaar door middel van K-means.

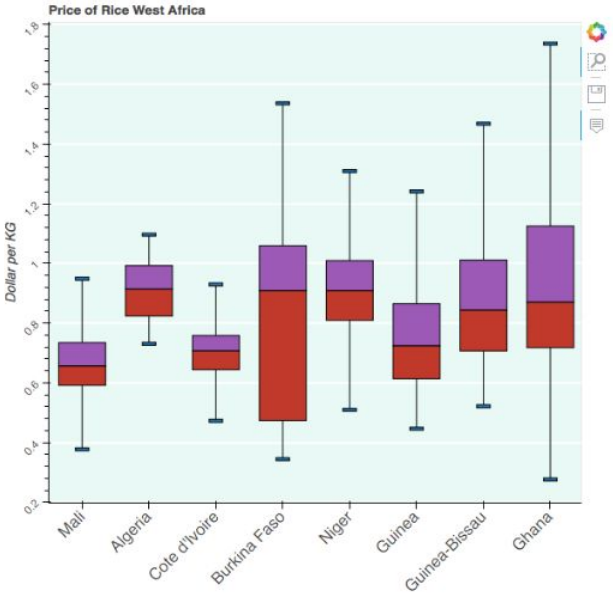
- Werkende boxplot maken die de prijs distributie van rijst door de jaren heen voor elk land kan weergeven, deze visualisatie kan gebruikt worden voor de beantwoording van deelvraag 2.
- lineaire regressielijnen maken om de data te kunnen analyseren
- Bubble plot maken om BNP met de gemiddelde rijstprijzen en populatie te combineren.

Gemaakte beslissingen en belangrijk inzagen

- Code geschreven die boxplot maakt van specifiek gevraagde landen. Hierdoor kunnen landen in hetzelfde gebied met elkaar en met landen uit een ander gebied vergeleken worden. Deze visualisatie kan gebruikt worden voor de beantwoording van deelvraag 2.
- Boxplots van de verandering van rijstprijzen in Midden Oosten en West Afrika:

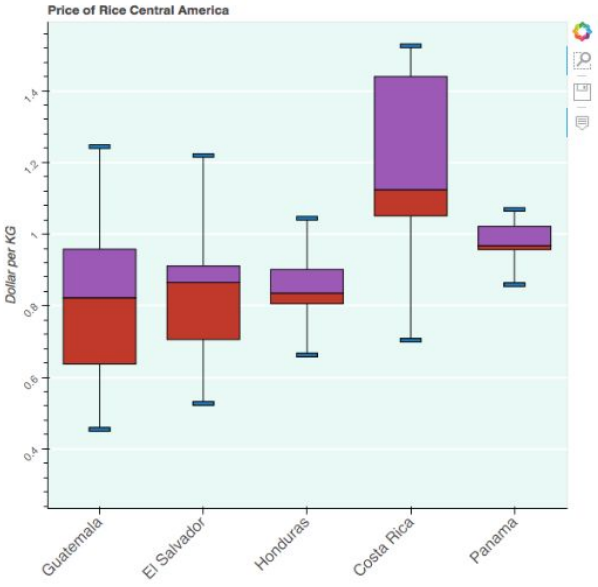


-



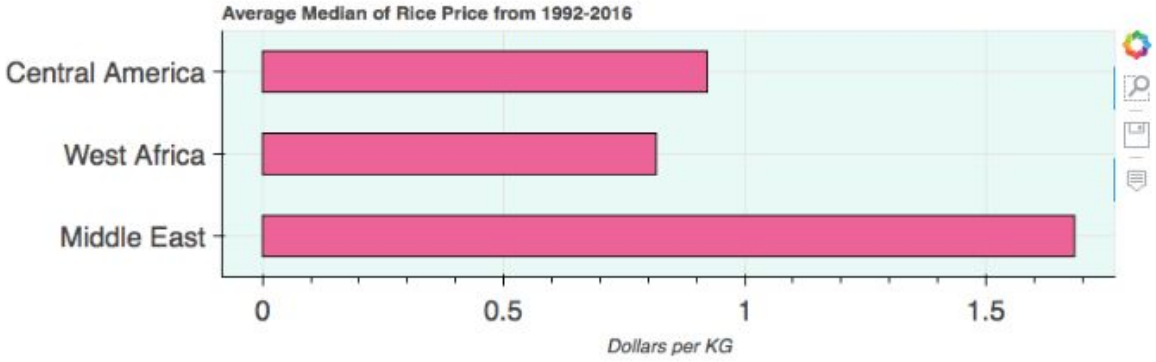
-

- Boxplot Central America is toegevoegd om een extra regio te kunnen tonen



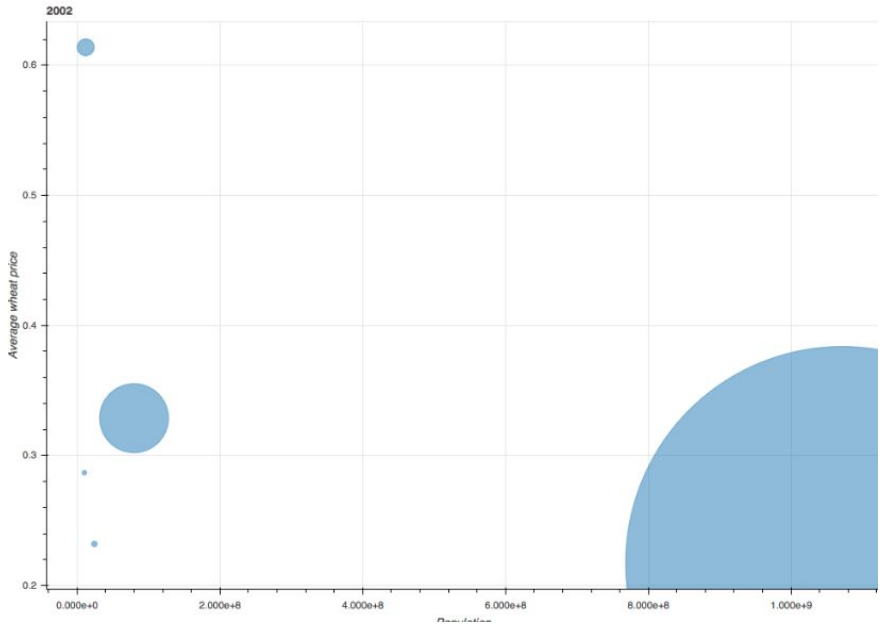
-

- Staafdiagram laat de gemiddelde medians zien van de rijstprijzen van de landen van Centraal Amerika, West Afrika en Midden Oosten die in de boxplots zijn getoond van jaar 1992 tot 2016, deze staafdiagram laat duidelijk zien dat de prijzen van rijst van de landen uit het midden Oosten een stuk hoger zijn dan de prijzen van rijst in West Afrika. Centraal Amerika en West Afrika zijn beiden geen rijke gebieden maar ook hier laat de staafdiagram een verschil zien. De staafdiagram is naast de boxplots toegevoegd om snel en duidelijk de verschillen weer te geven.



-

- Bij deze analyse is rijst als goed gekozen omdat dit het meest voorkomende product was en daarom een groot aantal landen met elkaar vergeleken konden worden.
- correlaties tussen regenval en productprijzen berekend om te kijken of de prijs van producten verklaard kan worden aan de hand van regenval. Dit is van belang voor de beantwoording van deelvraag 3.
- correlaties tussen populatie en productprijzen berekend om te kijken of de prijs van producten verklaard kan worden aan de hand van populatie. Dit is van belang voor de beantwoording van deelvraag 3.
- overzicht gemaakt met aantal landen die een hoge correlatie hebben bij een bepaalde productcombinatie, dit omgezet in een csv bestand om de gegevens makkelijk via pandas te analyseren.

- Voorlopige bubble plot gemaakt waarbij de x-as de populatie laat zien, de y-as de gemiddelde rijstprijzen van die landen. De grootte van de bubbles representeert het BNP en het hover-effect toont welke bubbles welke landen voorstellen.
-


-

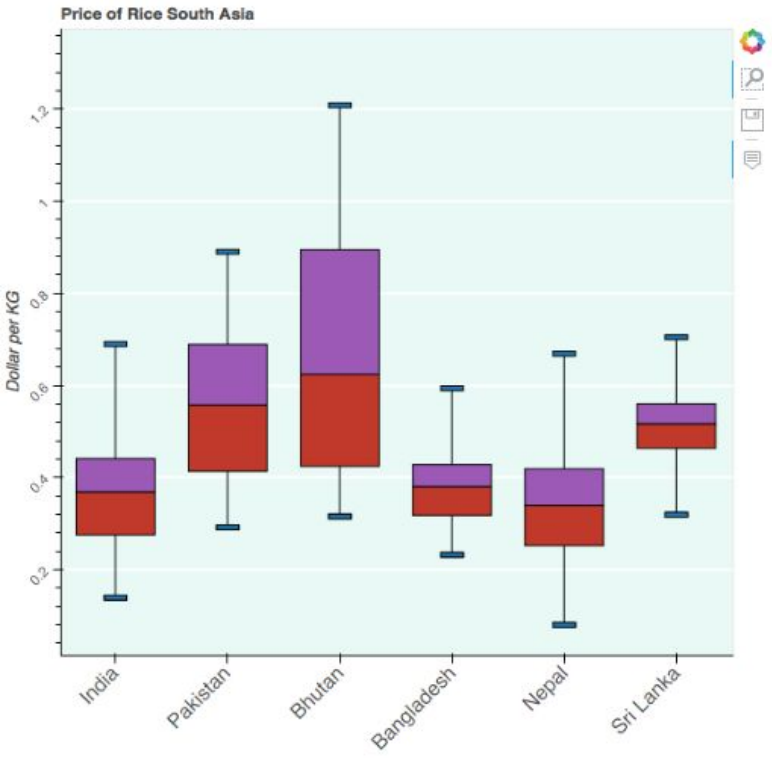
Week 4

Actiepunten

- Boxplot visualisatie afmaken
- Staafdiagram maken die de verschillende prijsverandering in rijst per regio snel en duidelijk laat zien
- Bubble chart visualisatie afmaken, inclusief slider door de jaren heen
- Correlaties netjes ordenen in tabellen
- scatter plots en lineaire regressie maken voor de correlaties
- t-SNE visualisatie maken voor
- Report schrijven
- Website maken

Gemaakte beslissingen en belangrijke inzagen

- Boxplot van Centraal Amerika is ingeruild voor Oost Afrika, omdat er voor het verslag, voor de tweede deelvraag, op de volgende regio's de focus is gelegd: Oost Afrika, West Afrika, Zuid Azië en het Midden Oosten. Deze regio's zijn gekozen omdat hier de meeste data beschikbaar van is.
- Boxplot Zuid Azië:



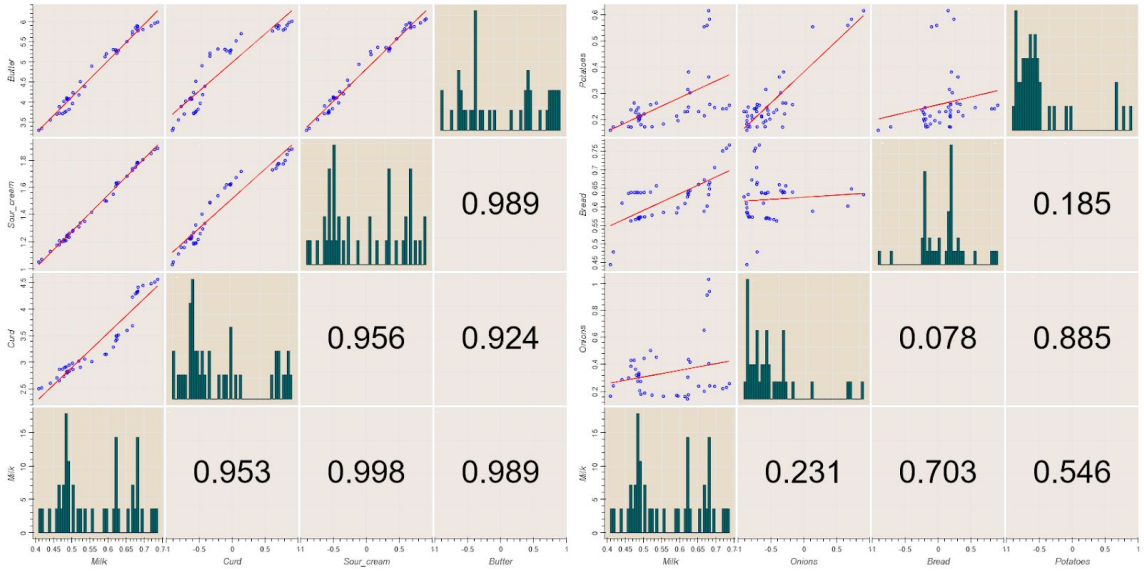
-

- De landen Palestina en Mauritania zijn niet meegenomen in de boxplots en staafdiagram omdat deze erg afwijkende waardes hadden ten opzichte van de rest van hun gebied.
- gemiddelde mediaan prijs rijst 1992-2016 uitgerekend voor de volgende vier regio's:
  Midden Oosten:1.6833722807349696 US dollar per KG
  West Afrika:0.8163391925119723 US dollar per KG
  Oost Afrika:0.9037541711966105 US dollar per KG
  Zuid Azië: 0.46402898290547173 US dollar per KG
  - Deze waardes zijn significant verschillend, dus dit kan gebruikt worden voor de beantwoording van deelvraag 2.
- Uiteindelijk zijn de landen waar mee gewerkt gaat worden, de landen uit de volgende regio's:
  - Oost Afrika = Mozambique, Zambia, Unted Republic of Tanzania, Madagascar, Malawi, Burundi, Ethiopia, Djibouti, Kenya, Rwanda, Somalia, Uganda, Sudan, South Sudan
  - Midden Oosten = Armenia, Iraq, Iran, Turkey, Syrian Arab Republic, Jordan, Yemen, Afghanistan
  - West Afrika = Mali, Algeria, Cote d'Ivoire, Burkina Faso, Niger, Guinea, Guinea-Bissau, Ghana, Cameroon, Gambia, Mauritania, Nigeria
  - Zuid Azie = India, Pakistan, Bhutan, Bangladesh, Nepal, Sri Lanka, Tajikistan
- Uit de boxplots en staafdiagram komt naar voren dat landen in hetzelfde gebied, dezelfde prijsverandering hebben. In Guterres & Peter (2012) is gesteld dat de vluchtelingenstroom in een land gekoppeld kan worden aan de prijzen van voedsel. Daarom zijn we correlaties gaan bekijken tussen de factor vluchtelingenstroom in een

land en de graanprijzen (de prijs van rijst, maïs, millet en sorghum) van dat land. Dit hebben we voor de 10 landen waar de meeste vluchtelingen in voorkomen. Dit is te zien in onderstaande tabel:

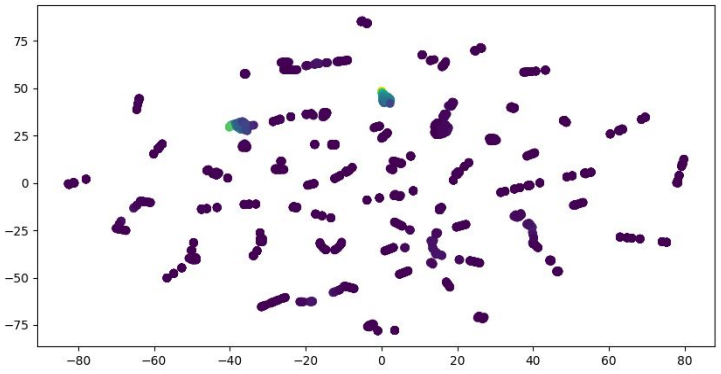
#	Country Name	Correlation Rice Price and Incoming Refugees
0	Lebanon	0.403
1	Pakistan	0.684
2	Ethiopia	0.379
3	Chad	0.826
4	Jordan	0.979
5	Uganda	0.421
6	Burundi	0.792
7	Somalia	0.995
8	Iraq	0.881
9	Mali	0.443
10	Turkey	-0.603
11	Cote d'Ivoire	-0.693

- Deze tabel geeft dus grotendeels aan dat de vluchtelingenstroom een verklaring kan geven voor deelvraag 2.
- Alle tabellen omgezet in lateX tabellen, zodat de tabellen in het report een consequente lay-out hebben.
- Functie geschreven die commodity uitzet tegen BNP.
- Scatter matrix gemaakt voor Oekraïne, Oekraïne gekozen omdat dit land gegevens had over veel verschillende producten. In de matrix links zijn de producten melk, wrongel, zure room en boter gebruikt, hier is ook te zien dat de prijsveranderingen van deze producten hele hoge correlaties met elkaar hebben. In de scatter matrix rechts, is als controle een matrix gemaakt waarin de producten melk, ui, brood en aardappels zijn gebruikt. Deze matrix geeft, in lijn is met onze verwachtingen, lage correlaties aan. Deze matrixen kunnen gebruikt worden in het report voor de beantwoording van deelvraag 1.



-

- Het bleek dat de BNP correlaties van Yemen met suiker en rijst een hoge negatieve correlaties had. Om dit verder te onderzoeken is de BNP van Yemen geplot, hieruit is gekomen dat de prijs van deze producten de laatste jaren heel sterk is gedaald. Aangezien dit gegeven niet helemaal aansluit bij de deelvragen van dit verslag, is hier verder geen verklaring voor gezocht.
- Er zijn scatterplots gemaakt met lineaire regressie en tabellen met gemiddelde rijst prijzen, olieprijzen, en mean square error van de scatterplots voor alle landen in de eerder gekozen regio's. Dit is gedaan om de beantwoording van deelvraag 2 met extra visualisaties te bevestigen.
- In week 3 was er een clustering gemaakt van alle databestanden bij elkaar door middel van K-means. Om deze clustering in 2D weer te geven zonder data te verliezen hebben we deze data gevisualiseerd door middel van t-SNE. Hier zijn echter geen interessante gegevens voor de deelvragen uitgekomen.



-