De patronen van de Economie

Data-analyse en visualisatie

Karst Backer, Remco Jacobs, Franklin Willemen, Louis van Zutphen
11700173, 11860316, 10992693, 11851910

Inhoud

Inleiding

In dit onderzoek wordt gekeken naar de dataset *Productprijzen per markt* [1] van het CBS. In deze dataset is vanaf 1981 voor een groot aantal branches de procentuele verhouding van de prijs voor de afzet naar binnenland, buitenland en totaal ten opzichte van 2010 weergegeven. We weten dat er in de afgelopen jaren een kredietcrisis heeft plaatsgevonden. Er wordt gekeken of dit in de data is af te lezen en of er indicaties zijn die alle crises gemeen hebben. Er wordt verwacht dat voor de dips die zullen plaatsvinden tijdens de crisis, de prijzen eerst stijgen. Dit kan verklaard worden aan de hand van het bubbel effect. Voordat een crisis zich voordoet zullen de prijzen eerst oplopen en daarna snel kelderen.
Verder wordt gekeken of er dips in de markt aanwezig zijn die niet kunnen worden verklaard in vergelijking tot andere markten. Er wordt gepoogd historische gebeurtenissen te linken aan deze dips. De verwachting is dat rond 2008 een aantal branches zal instorten en dat de prijzen zullen dalen.
Als laatst zal worden gezocht naar patronen die interessante resultaten op zouden kunnen leveren. Hier wordt ook gekeken naar naar onverwachte patronen die men niet zou verwachten. Met de opkomst van duurzame energie zal de ontwikkelingen van andere energiebronnen dalen. Dit wordt gedacht omdat de energieconsumptie niet veel zal veranderen terwijl de verhoudingen tussen duurzame energiebronnen en andere energiebronnen we zullen verschuiven.
Deze data dient geanalyseerd, verwerkt en vervolgens gebruikt te worden om de drie onderzoeksvragen te beantwoorden. Verder wordt er gepoogd de deelvragen volledig te beantwoorden. Ook wordt naar relaties gezocht tussen bepaalde branches. Als laatste dient de data duidelijk en op een mooie manier gevisualiseerd te worden.

Methode

Preprocessing (week 1)

Afzet	BedrijfstakkenBranchesSBI2008	ID	OntwikkelingTOV1JaarEerder_2	OntwikkelingTOV1MaandEerder_3	Perioden	ProducentenprijsindexPPI_1	Wegingcoefficient_4
A6	305700	0			1981MM0	53.8	0
A6	305700	1		0.6	1981MM0	54.1	0
A6	305700	2		0.4	1981MM0	54.3	0
A6	305700	3		0	1981MM0	54.3	0
A6	305700	4		0.4	1981MM0	54.5	0
A6	305700	5		1	1981MM0	55	0
A6	305700	6		1	1981MM0	55.6	0
A6	305700	7		-0.1	1981MM0	55.5	0
A6	305700	8		-0.2	1981MM0	55.4	0
A6	305700	9		-0.1	1981MM1	55.4	0
A6	305700	10		0	1981MM1	55.4	0
A6	305700	11		0	1981MM1	55.4	0
A6	305700	12			1981JJ00	54.9	0
A6	305700	13	7.6	4.5	1982MM0	57.9	0
A6	305700	14	7.8	0.8	1982MM0	58.3	0
A6	305700	15	8.1	0.7	1982MM0	58.7	0
A6	305700	16	9.4	1.2	1982MM0	59.4	0
A6	305700	17	9.4	0.3	1982MM0	59.6	0
A6	305700	18	8	-0.3	1982MM0	59.4	0
A6	305700	19	6.5	-0.3	1982MM0	59.2	0
A6	305700	20	6.2	-0.4	1982MM0	58.9	0
A6	305700	21	5.7	-0.6	1982MM0	58.6	0
A6	305700	22	5.7	0	1982MM1	58.5	0
A6	305700	23	6.6	0.8	1982MM1	59	0
A6	305700	24	6.7	0.2	1982MM1	59.1	0
A6	305700	25	7.3		1982JJ00	58.9	0
A6	305700	26	2.5	0.3	1983MM0	59.3	0
A6	305700	27	1.7	-0.1	1983MM0	59.3	0
A6	305700	28	0.9	0	1983MM0	59.2	0
A6	305700	29	-0.3	0	1983MM0	59.2	0
A6	305700	30	-0.4	0.2	1983MM0	59.3	0
A6	305700	31	0.5	0.5	1983MM0	59.7	0
A6	305700	32	1.1	0.3	1983MM0	59.8	0
A6	305700	33	2.9	1.4	1983MM0	60.7	0
A6	305700	34	5.5	1.8	1983MM0	61.8	0
A6	305700	35	5.6	0.1	1983MM1	61.8	0
A6	305700	36	5.5	0.7	1983MM1	62.3	0
A6	305700	37	6.1	0.7	1983MM1	62.7	0
A6	305700	38	2.6		1983JJ00	60.4	0
A6	305700	39	8.6	2.7	1984MM0	64.4	0
A6	305700	40	12.1	3.2	1984MM0	66.5	0
A6	305700	41	12.9	0.6	1984MM0	66.9	0
A6	305700	42	13.4	0.4	1984MM0	67.2	0
A6	305700	43	12.7	-0.5	1984MM0	66.9	0
A6	305700	44	12.4	0.3	1984MM0	67.1	0

Figuur 1: Een visualisatie van het ruwe dataframe.

De dataset bestaat uit 96681 entries die de relatieve prijzen per branche ten opzichte van 2010 bevatten. Verder zijn er kolommen die de prijs ten opzichte van de vorige maand en het vorige jaar bevatten. Een voorbeeld hiervan is te zien in figuur 1.Ook is er een wegingscoëfficiënt dat bepaalt hoeveel het cijfer meetelt in het totaalplaatje. De datatypes van alle kolommen in de dataset zijn te zien in figuur 2.

Kolomnaam	Datatype
Afzet	Nominal
BedrijfstakkenBranchesSBI2008	Nominal
ID	Ratio
OntwikkelingTOV1JaarEerder_2	Ratio
OntwikkelingTOV1MaandEerder_3	Ratio

Perioden	Ordinal
ProductenprijsindexPPI_1	Ratio
Wegingscoëfficiënt	Ratio

Figuur 2: De datatypes per kolom.

Verder kan over de data gesteld worden dat alle data al genormaliseerd is. Alle prijzen zijn namelijk relatief aan de prijzen van die branche in 2010. Ook valt op dat er veel datapunten ontbreken. Een groot aantal hiervan komt doordat er op dat moment (nog) niet gemeten is. Bij veel branches die deze ontbrekende waarden bevatten beginnen de eerste metingen op hetzelfde tijdstip. Verder bevat de data overkoepelende sectoren die een aantal branches bevat. Als laatste bevatten sommige maanden meerdere meetpunten. Om de data gelijk te houden is per maand het gemiddelde genomen indien er meerdere meetpunten aanwezig zijn.

Bij het preprocessen van de data is allereerst alle data van JSON omgezet naar een PANDAS dataframe. Vervolgens is de data bekeken en zijn onnodige kolommen verwijderd. De kolommen Afzet, BedrijfstakkenBranchesSBI2008, ID, OntwikkelingTOV1JaarEerder_2 en OntwikkelingTOV1MaandEerder_3 zijn uit de dataset verwijderd omdat Afzet en BedrijfstakkenBranchesSBI2008 voor de hele subset gelijk zijn. ID omdat deze er 2x in stond. OntwikkelingTOV1JaarEerder_2 en OntwikkelingTOV1MaandEerder_3 omdat deze waarschijnlijk niet nodig zijn en als ze nodig zijn kunnen ze berekend worden. Verder zijn in de kolom Perioden de entries met een jaartal weggehaald. Dit is gedaan omdat deze afwijkende data geven tussen de maanden en als ze nodig zijn kunnen ze worden berekend. De weggehaalde kolommen zullen later nog wel worden gebruikt maar zijn gesplitst van de gecleande dataset. Hierna is de dataset opgedeeld in kleinere datasets die over 1 branch en 1 afzet gaan. Hier is vervolgens pandas.Dataframe.interpolate op toegepast om missende data aan te vullen. Als laatst zijn entries die nog steeds leeg waren uit de data verwijderd. Dit is gedaan omdat deze overgebleven entries punten in de tijd zijn waar geen metingen zijn gedaan.

Exploratory data analysis (week 2)
Om een globaal beeld te krijgen van het aantal datapunten dat er ontbreekt bij deze dataset hebben we de pandas isnull() en sum() functies toegepast op onze dataframe. Het aantal ontbrekende waardes is te zien in figuur 3.

Afzet	0
BedrijfstakkenBranchesSBI2008	0
ID	0
OntwikkelingTOV1JaarEerder_2	39885
OntwikkelingTOV1MaandEerder_3	41120
Perioden	0
ProductenprijsindexPPI_1	36179
Wegingscoëfficiënt	0

Figuur 3: Aantal ontbrekende waardes per kolom.

Figuur 4 geeft een vertekend beeld, namelijk als er per bedrijfsbranche wordt gekeken valt er iets bijzonders op, namelijk veeltal branches delen exact hetzelfde aantal datapunten dat "missen" voor OntwikkelingTOV1JaarEerder_2(Ojaar), OntwikkelingTOV1MaandEerder_3(Omaand) en PPI. Zo is hieronder te zien van links naar rechts: branchecode, totaal datapunten van branche, data punten waarbij alle drie de features ontbreken en het totaal aantal datapunten waarbij er iets ontbreekt van de drie features. Er is hier gebruik gemaakt van de .loc functie van pandas om deze data uit de dataframe te halen.

Branche	Aantal Datapunten	3 features ontbreken	1 feature ontbreekt
305700	1443	50	231
305800	1443	52	231
306300	1443	0	183
307500	1443	0	183
307610	1443	936	1011
307600	1443	0	183
307700	1443	936	1011
307800	1443	936	1011

307900	1443	936	1011
308000	1443	936	1011
308100	1443	940	1017
308300	1443	936	1011
308400	1443	936	1011
308500	1443	936	1011
308600	1443	936	1011

Figuur 4: Aantal ontbrekende datapunten per kolom van een aantal branches.

Door de dataset te splitsen op basis van overeenkomende ontbrekende data werd het mogelijk om de reden ervan te achterhalen en op de juiste wijze te imputen. Zo blijkt dat alle branches waarbij er rond de 1011 rows data ontbrak er later dan 1981 is begonnen met meten, het overgrote deel van de ontbrekende data is dus hiermee te verklaren. Voor de rows waar er wel data ontbreekt is er gebruik gemaakt van de interpolate pandas functie. Dit is te zien in figuur 5.

```
removed936 = newDF2.loc[newDF2["BedrijfstakkenBranchesSBI2008"] == branches936[0]]
i = 1
while i < len(branches936):
    temp = newDF2.loc[newDF2["BedrijfstakkenBranchesSBI2008"] == branches936[i]]
    removed936 = removed936.append(temp)
    i += 1

removed936 = removed936.interpolate()
removed936 = removed936.loc[(removed936["OntwikkelingTOV1JaarEerder_2"].notnull()) |

removed936.isnull().sum()
```

```
Afzet                          0
BedrijfstakkenBranchesSBI2008  0
ID                             0
OntwikkelingTOV1JaarEerder_2  13
OntwikkelingTOV1MaandEerder_3  1
Perioden                       0
ProducentenprijsindexPPI_1     0
Wegingcoefficient_4            0
dtype: int64
```

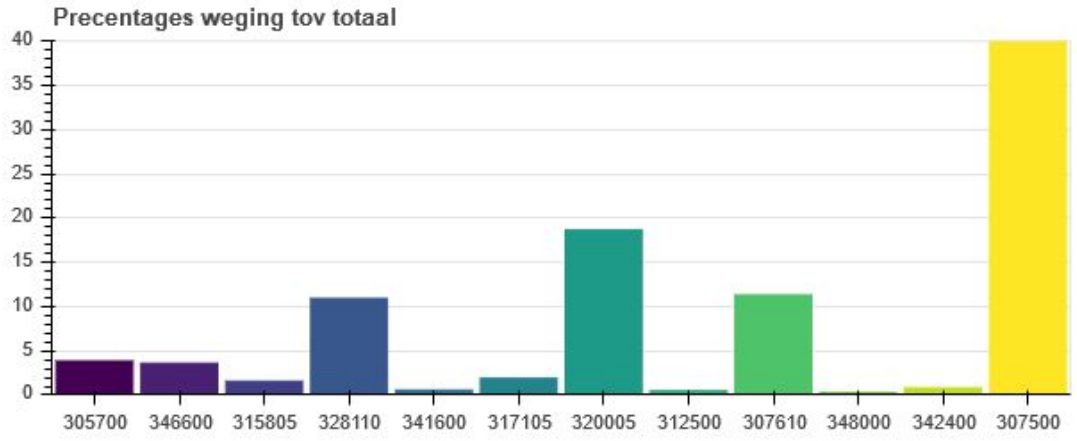
Figuur 5: Het aantal ontbrekende waarden per kolom na het cleanen.

De nog ontbrekende data volgens figuur 4 heeft te maken met het feit dat er geen metingen zijn voor de eerste maand of de eerste jaar. Er zijn ook een aantal branches waar er alles ontbrak of waarbij er maar voor een paar jaar metingen zijn opgenomen.

Ook de feature wegingscoëfficiënt is nader bekeken, zo blijkt het dat alleen vanaf 2005 deze feature is toegevoegd. Er is per overkoepelende branche gekeken wat het percentage is van zijn aandeel in het totaal (figuur 6).

Branche	Percentage van totaal
305700 Delfstoffenwinning	0.0383
307500 Industrie	0.3622
307610 Voedings-, genotmiddelenindustrie	0.1130
312500 Textielindustrie	0.0047
315805 Hout- en bouwmaterialenindustrie	0.0155
317105 Papier- en grafische industrie	0.0188
320005 Raffinaderijen en chemie	0.1864
328110 Metalektro	0.1090
341600 Meubelindustrie	0.0055
342400 Overige industrie	0.0079
346600 Energievoorziening	0.0358
348000 Waterbedrijven en afvalbeheer	0.0029

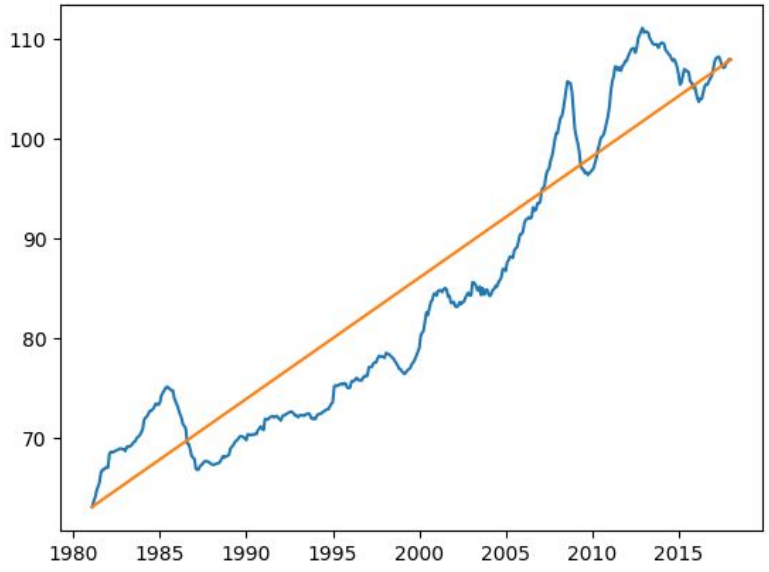
Figuur 6: Gemiddelde wegingscoëfficiënt van de overkoepelende branches.



Figuur 7: Wegingspercentages ten opzichte van het totaal.

Uit figuur 7 blijkt dat Industrie en Raffinaderijen en chemie de twee grootste aandeel hebben in de Nederlandse binnenlands en buitenlands afzet. Ook is er gekeken naar de percentages van de weging van de binnenland en buitenlands afzet ten opzichte van de totaal per branche, hieruit kon direct uit opgemaakt worden of de desbetreffende bedrijfsbranche een grotere binnenlands of buitenlands afzet had.

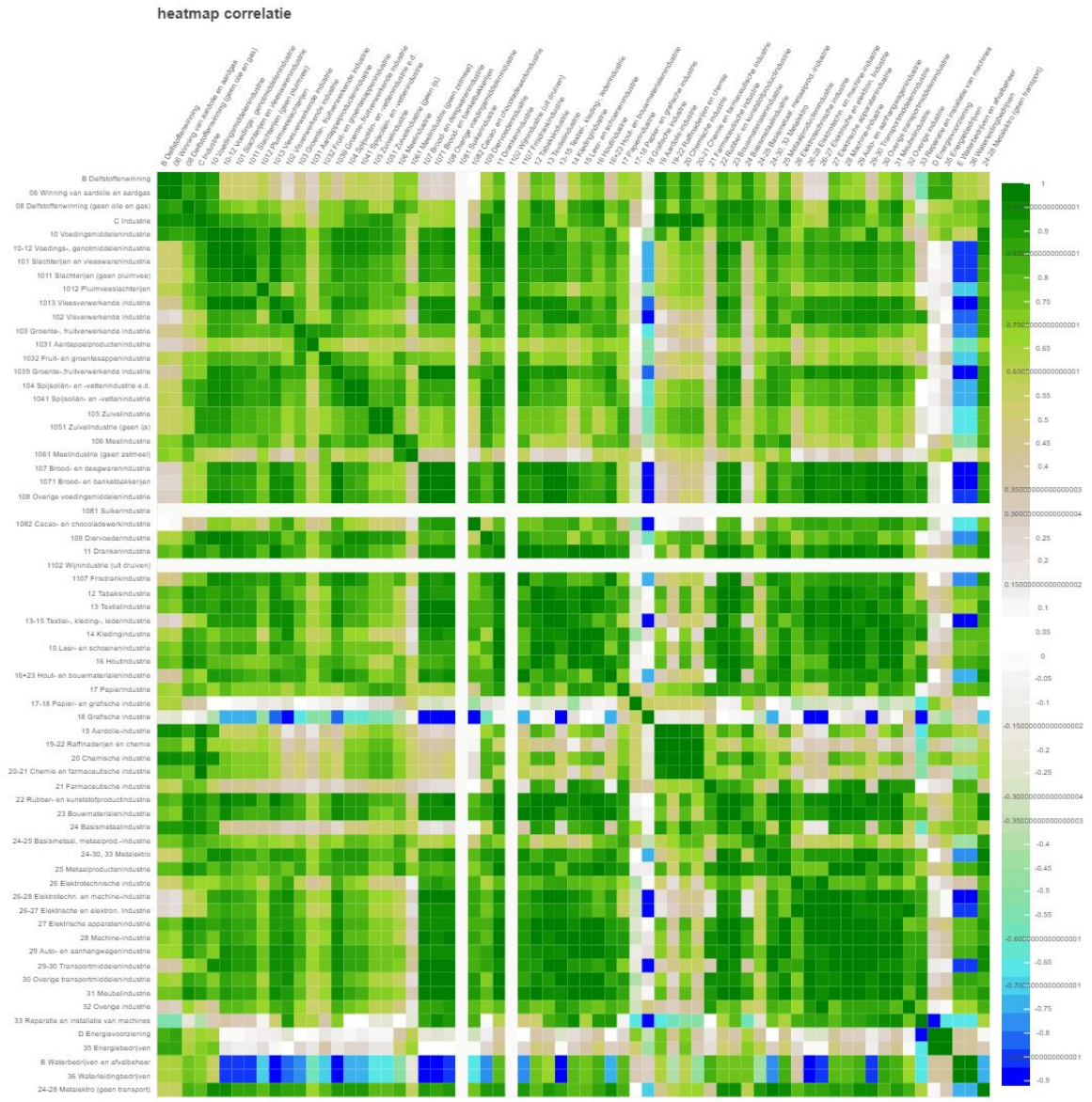
Om een beter idee te krijgen wat de prijs groei van de bedrijven betekent, kunnen ze in relatie gesteld worden tot het gemiddelde. In dit geval is dit de gemiddelde prijs groei van Nederland. In de volgende grafiek is deze gemiddelde prijs groei te zien samen met de gemiddelde groei van alle prijzen in nederland.



Figuur 8: Totale en gemiddelde stijging van alle branches.

Met behulp van de gemiddelde prijs groei in nederland zijn er grafieken gegenereerd die relatief staan aan het gemiddelde van nederland (figuur 8). Bij deze grafieken is het gemiddelde dus de nullijn en dus kun je makkelijk zien wanneer een branche boven of onder het gemiddelde zit.

Om te zien hoe de branches aan elkaar relateren, wordt de correlatie tussen de branches berekend. Om deze overzichtelijk weer te geven is een heatmap gemaakt, waar de correlatie wordt weergegeven door de kleur. In figuur 9 is te zien dat er mogelijk groeperingen zijn in de branches.



Figuur 9: Op de x en y as staan alle branches. de kleur geeft de correlatie tussen die twee branches aan.

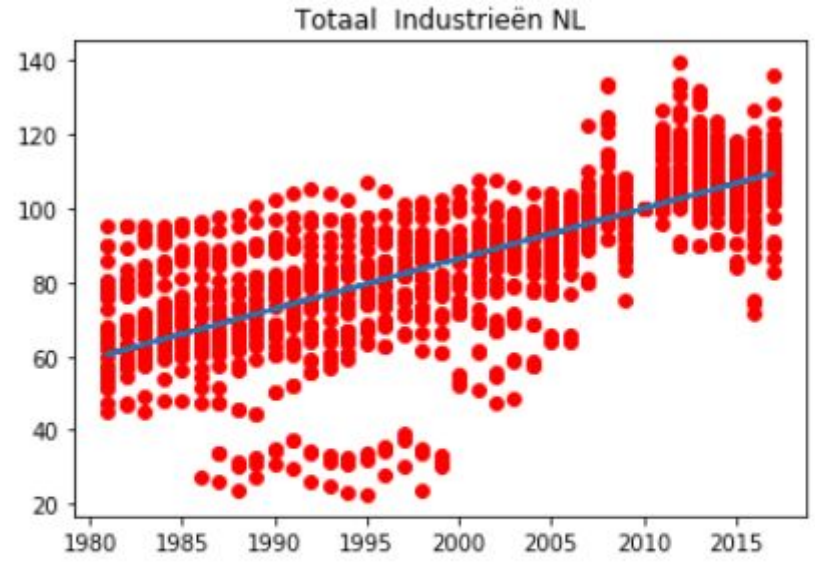
In-depth analysis (week 3)

Er zijn nieuwe features gemaakt op basis van de wegingscoëfficiënt waaronder de percentages van de weging t.o.v. de branche zelf en het totaal en ook een feature om aan te geven welke afzet het grootst is bij de desbetreffende branche (figuur 10). (pOB = percentage Overkoepelende Branche, wegingB = weging van desbetreffende afzet dus 1 als het totaal betreft, pTovTotaal = percentage ten opzichte van totaal.)

	Afzet	grootsteAfzet	pOB	wegingB	pTovTotaal
0	2	2.0	0.112967	1.0	0.112967
1	2	2.0	0.112967	1.0	0.112967
2	2	2.0	0.112967	1.0	0.112967
3	2	2.0	0.112967	1.0	0.112967
4	2	2.0	0.112967	1.0	0.112967

Figuur 10: Nieuwe features.

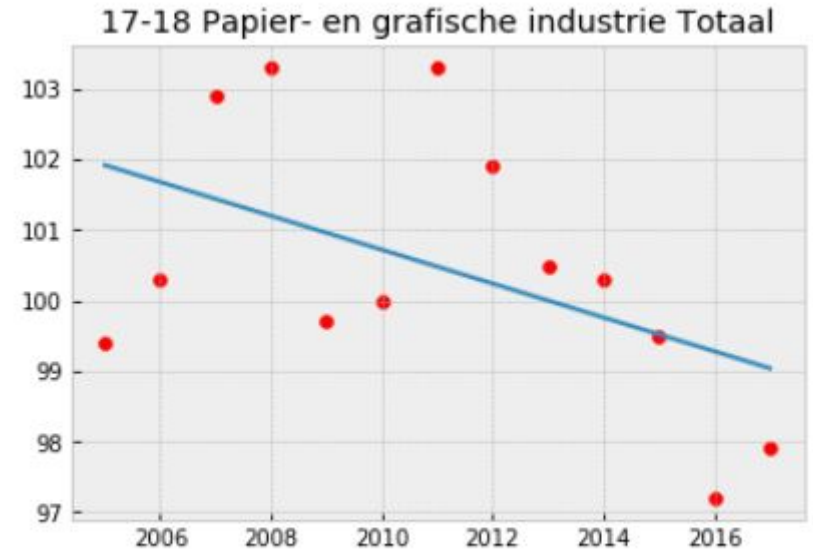
Helaas hebben deze nieuwe features het niet mogelijk gemaakt te clusteren of multivariate regressie toe te passen. Tot nu toe blijkt het zo te zijn dat de dataset alleen tweedimensionale regressie toestaat, wegens tijdsdruk zijn geen nieuwe datasets erbij betrokken en de is huidige dataset niet verder geëxploreerd. Het was wel mogelijk geweest een eenvoudige regressie lijn te berekenen en deze te plotten met bijhorende slope (figuur 11).




```
RSME: 13.076696830622021
Slope: 1.36
```

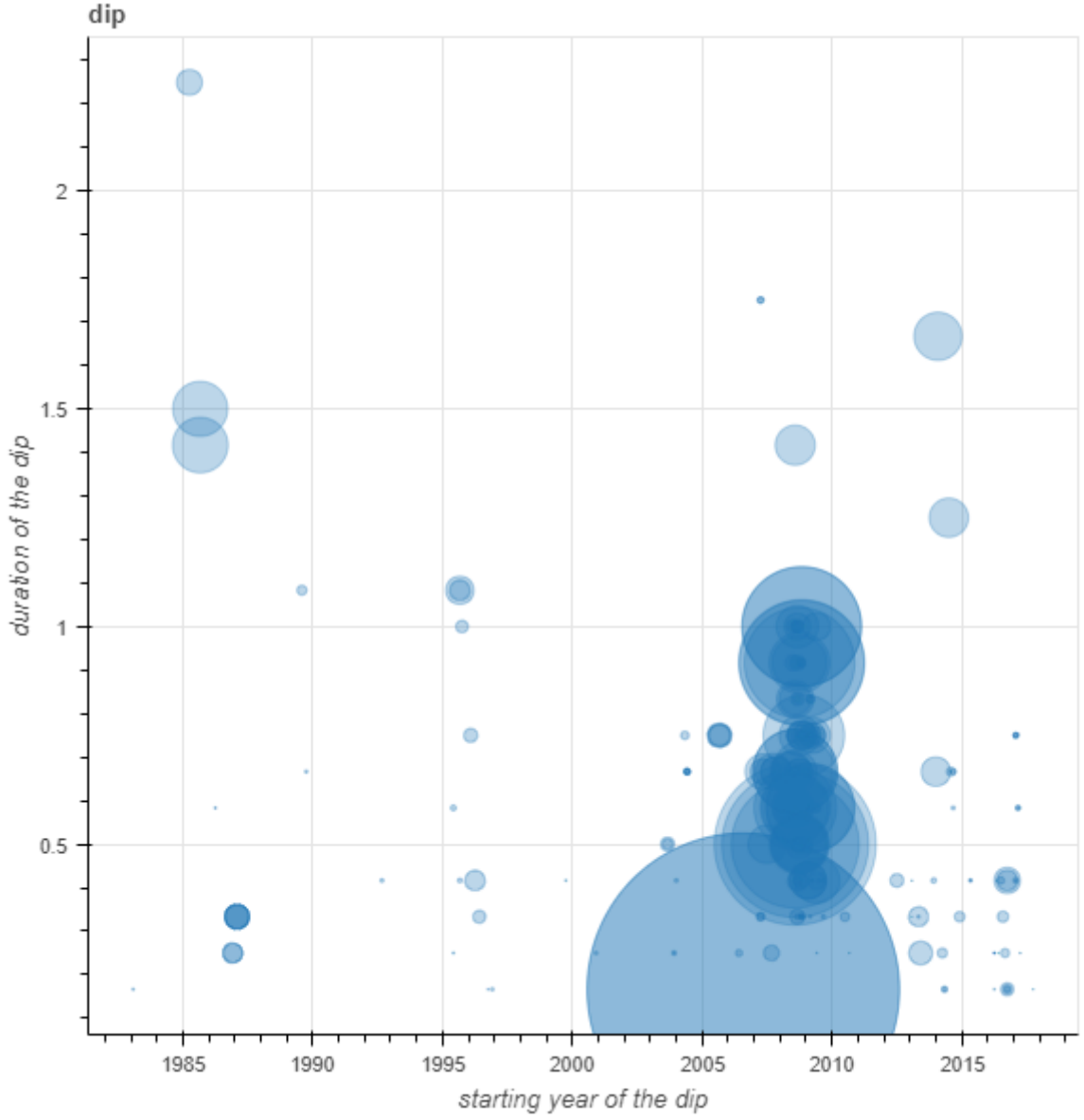
Figuur 11: Regressie toegepast op het totaal.

De enige branche waar er sprake was van een daling in de loop van 1988 tot 2016 is de Papier en grafische industrie (figuur 12).



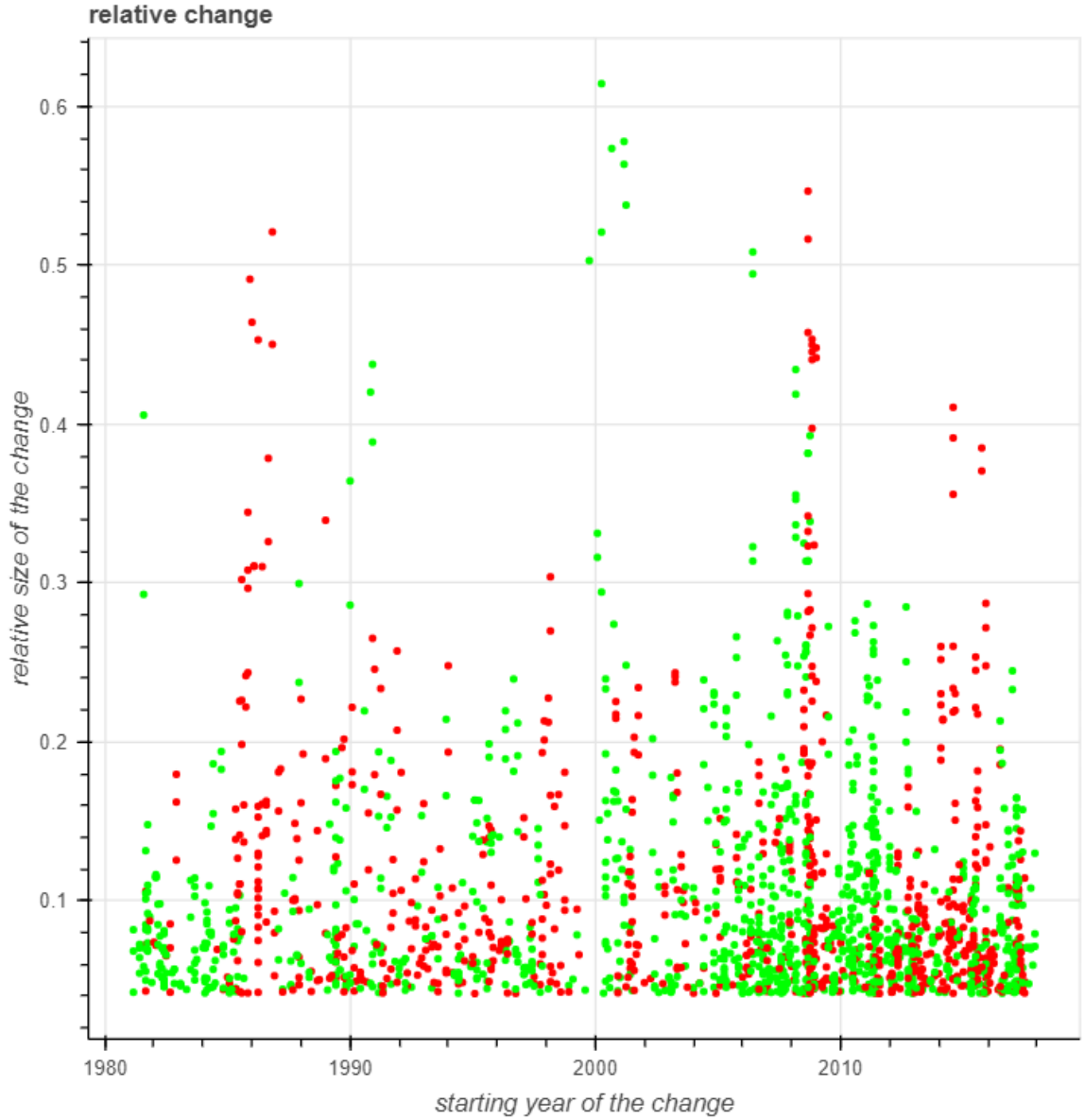
Figuur 12: Verloop van papier- en grafische industrie.

Om een beter idee te krijgen van crises, is het handig om te kijken naar grote veranderingen. tijdens crises zijn er vaak zeer grote veranderingen, dus er werd gekeken naar de grootste verandering in elke branche. In figuur 13 is een grafiek te zien dat er een sterke groepering van deze grootste dalingen is rond 2009. de grootste cirkel in deze grafiek hoort bij de energievoorziening, wiens prijsindex erg veel omhoog en omlaag gaat.



Figuur 13: Elk punt op deze grafiek geeft een aaneensluitende daling weer. op de y as is weergegeven hoe lang deze daling duurt en op de x is de tijd van het begin van de daling. de grootte van de cirkel representeert de grootte van de daling.

Figuur 13 geeft geen volledig beeld van de crises, omdat het veel grote veranderingen mist, inclusief grote stijgingen. In figuur 14 is een nieuwe grafiek gemaakt die positieve en negatieve veranderingen beschrijft. Deze grafiek geeft een groot deel van alle verandering weer. De veranderingen die niet zijn meegenomen zijn zeer klein en zullen niet veel effect hebben op de economie. Om de veranderingen beter te kunnen vergelijken is de y coördinaat van deze grafiek de verandering relatief aan zichzelf voor de verandering, in plaats van de verandering in procentpunten relatief aan 2010.



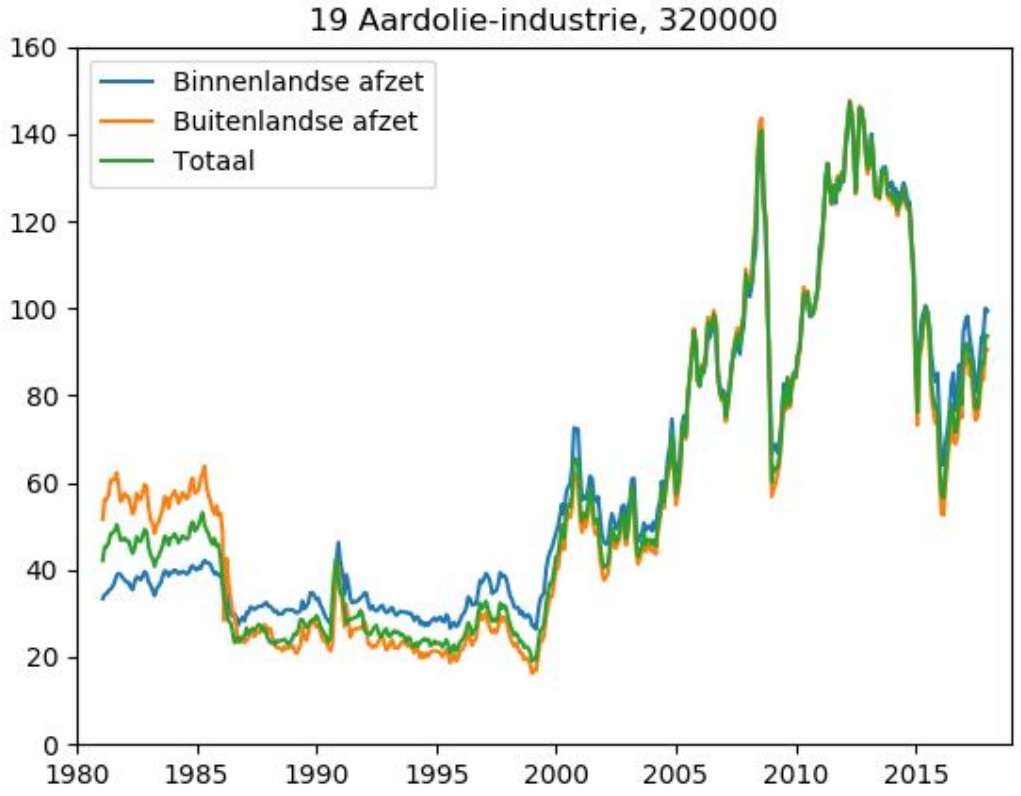
Figuur 14: Een bokeh plot van de aaneengesloten veranderingen in prijs. Jaar van de verandering is hier uitgezet tegen de relatieve hoeveelheid van de verandering, dus als iets de helft van zijn waarde verliest is de y coördinaat van dat punt 0,5. De rode stippen staan hier voor dalingen en de groene stippen voor stijgingen in de prijs.

Resultaten

In figuur 13 valt op dat er een aantal momenten zijn geweest waarop veel branches tegelijk in waarde zijn gedaald.

In figuur 14 valt ook weer op dat er een aantal momenten in het verleden zijn geweest waarop veel branches tegelijkertijd daalden. Verder valt op dat voor deze dalingen vaak een grote hoeveelheid stijgingen te vinden is.

In figuur 15 valt op dat er vlak voor het jaar 2001 en vlak voor 2008 twee grote pieken te vinden zijn. In 2008 is te zien dat de grafiek na de piek in één keer 80 procentpunten van zijn prijs verliest.



Figuur 15: Een grafiek van het prijsverloop tussen 1981 en 2017 van de Nederlandse aardolie-industrie. Op de verticale as is de relatieve prijs ten opzichte van 2010 uitgezet.

Discussie

In de resultaten is te zien hoe veel branches op gelijke tijdstippen pieken en dalen vertonen. Dit kan verklaard worden aan de hand van gebeurtenissen die in het verleden hebben plaatsgevonden.
Om de eerste deelvraag te beantwoorden wordt er gekeken naar de verschillende branches en of zij dalingen vertonen tijdens de crisis. De crisis is terug te vinden in de branches Delfstofwinning, Industrie, Raffinaderijen en chemie, Voedings- en genotmiddelen. Het valt op dat bij veel branches vlak voor de daling een piek te vinden is. Dit kan worden verklaard doordat de vraag wordt overtroffen door het aanbod. Ook is zoals verwacht duidelijk te zien dat er voor elke crisis een piek aanwezig is in de grafieken.
Daarnaast is gekeken naar een dip bij de aardolie-industrie die niet in andere branches te vinden is of aan de hand van een andere branche te verklaren is. Deze dip kan worden verklaard aan de hand van het volgende citaat van het CBS: *"De productie van de aardoliewinning daalde met ruim 5% en wist zich dus nog niet te herstellen van de wereldwijde prijsdaling van ruwe aardolie met ongeveer 30% in 1998. Deze daling was er de oorzaak van dat een deel van de aardoliewinning op het continentaal plat minder rendabel werd, waardoor de maatschappijen tot productiebeperking overgingen. In maart 1999 besloten de elf OPEC-landen tot een productiebeperking ten einde de prijsval van aardolie te stoppen. Aangezien deze landen zich beter dan voorheen aan de afspraken hielden, steeg de olieprijs spectaculair: van $10.32 voor een barrel North Sea Brent in december 1998 naar $25.08 in december 1999. Voor de Nederlandse aardoliewinning betekende dit dat ondanks een afname van de geproduceerde hoeveelheden de waarde van de productie wel toenam."* [2]
Energievoorziening vertoont veel onverwachte patronen in de buitenlandse afzet. De afzet stijgt van de ene maand op de andere 50%. Het patroon van deze industrie is onvoorspelbaar en de data van het CBS zou inaccuraat kunnen zijn.
Een ander interessant patroon zou Brexit kunnen zijn. De Rabobank had hierover het volgende te melden: *“Niet elk bedrijf in de maakindustrie moet zich zorgen maken na een Brexit. Het zijn vooral de leer-, schoenen- en textielindustrie en de elektrische en optische apparatenindustrie (bijvoorbeeld brillenglazen en microscopen, maar ook computers en wasmachines) die het meest gaan voelen van een vaarwel van de Britten, vermoedt Rabobank.”* [3]
Volgens de Rabobank zou het duidelijk te zien zijn in de leer- en schoenenindustrie na een Brexit. Op 23 juni 2016 vond een referendum voor een Brexit plaats. Deze was echter, tegen de verwachting in, niet te zien in de leer- en schoenenindustrie.

Verder is in dit onderzoek ook gepoogd relaties te vinden tussen branches waarvan werd verwacht dat deze ongeveer dezelfde beweging vertoonde. Dit is echter niet gevonden waardoor deze vraag onbeantwoord blijft. Ook werd in dit onderzoek gekeken naar clusters in de heatmaps. Het is echter onzeker of deze clusters werkelijk bestaan aangezien de volgorde

van de kolommen en rijen willekeurig is. Een goed idee voor een vervolgonderzoek zou zijn om data van meer branches te gebruiken en om datapunten over een langere tijd te gebruiken. Verder kan het interessant zijn om andere data als politiek in het onderzoek te betrekken. Dit zal meer duidelijkheid geven over de rol die de politiek speelt in de economie van Nederland.

Referenties

[1] https://data.overheid.nl/data/dataset/producentenprijsindex-afzetprijzen-bedrijfstak-sbi-2008-2010100
[2] De Nederlandse economie 1999, Centraal Bureau voor de Statistiek
[3] https://www.businessinsider.nl/deze-bedrijven-krijgen-flinke-last-van-een-brexit-646230/