EINDOPDRACHT DATA ANALYTICS WITH PYTHON 2.0

Ingrid Dirckx

24-04-2024

Ingrid Dirckx

Introductie

Dit document is het eindrapport ter afsluiting van de opleiding “Python for Data Analytics”. Dit rapport vormt, samen met een Jupyter Notebook, de uitwerking van drie vragen. Alle drie de vragen hebben betrekking op CO2-emissie. In drie hoofdstukken worden de uitwerkingen toegelicht door eerst de data-analyse te becommentariëren en daarna de vraag te beantwoorden.

De drie vragen zijn:

1) Wat is de grootste voorspeller van een grote uitstoot van CO2-emissie per land, uitgedrukt in CO2-emissie per inwoner?
2) Welke landen maken de grootste stappen in het verminderen van CO2-emissie?
3) Welke non-fossiele energie technologie heeft in de toekomst de beste prijs?

Voor de uitwerking van de data-analyse wordt gebruikt gemaakt van de beschikbare databronnen afkomstig op de volgende website: “Our world in data” (https://ourworldindata.org/).

Aan het einde van elk hoofdstuk is een link opgenomen naar het Jupyter Notebook (zowel de link naar Google Colab als de link naar Github). In de code zijn alle genomen stappen voorzien van een korte toelichting.

Hoofdstuk 1 Wat is de grootste voorspeller van een grote uitstoot van CO2-emissie per land, uitgedrukt in CO2-emissie per inwoner?

Dataverzameling

Om de vraag te kunnen beantwoorden is gezocht op de website “Our World in Data” naar data betrekking hebbend op CO2-emissies.

Gestart is met zoeken naar informatie over de CO2-emissie per capita per land. Hierover is een CSV-datafile gevonden en geïmporteerd in het notebook. De file bevat historische data over de CO2-emissie per capita van1949 – 2022.

Bij het zoeken naar potentiële voorspellende variabelen, zijn de volgende categorieën gevonden:

- GDP per land, per capita, per jaar
- Primaire energieconsumptie per land, per capita, per jaar
- Dieetsamenstelling per land
- Aantal geregistreerde auto’s per 1000 inwoners van een land

Met deze potentiële voorspellers zal de vraag worden uitgewerkt. De gebruikte CSV-files zijn allemaal afkomstig van de website “Our World in Data”.

Opschonen en bewerken datafiles

Alle geïmporteerde files zijn opgeschoond door eerst de overbodige kolommen te verwijderen en de namen van de kolommen korter te maken (nieuwe naam geven). Vervolgens zijn alle namen omgezet naar kleine letters zodat het samenvoegen van de tabellen in een latere fase eenvoudiger zal zijn.

Een analyse van de rijen laat zien dat niet alleen landen zijn opgenomen in de tabel, maar ook subtotalen (bijvoorbeeld de werelddelen en groepen op basis van inkomen) en totalen (world). Deze (sub-)totalen zijn verwijderd uit alle dataframes.

Voor dit onderzoek is het van belang dat de periode, waarop de data betrekking heeft, voldoende lang is. Dit geeft meer betrouwbare resultaten. Echter de verwachting is dat de kwaliteit van de CO2-emissiedata, met het actueler worden van dit onderwerp, pas recentelijk in kwaliteit is verbeterd. Daarom is gekozen voor de onderzoeksperiode 2010 t/m 2022.

Van het bestand ‘Dieetsamenstelling per land’, is ervoor gekozen om alleen de informatie te gebruiken over de consumptie van dierlijke producten (vlees en van melk/eieren). Van veeteelt is immers bekend dat deze een grote bijdrage hebben aan de CO2-emissie.

Om de onderzoeksvraag te kunnen beantwoorden moet de focus worden gelegd op de landen met een grote uitstoot per capita. Daarom is gekozen om te werken met alleen de data van de landen met de grootste uitstoot per capita. Hun impact in de totale CO2-emissie is namelijk het grootst. Ook is de keuze gemaakt om als meetmoment, voor de bepaling van de grootste uitstoters, het meest actuele jaar te nemen, namelijk 2022. Aan de hand van het dataframe CO2-emissie per capita, is een top tien van grootste landen gemaakt (hoogste CO2-emissie per capita per jaar) en in een apart dataframe opgeslagen.

Combineren en analyseren datafiles

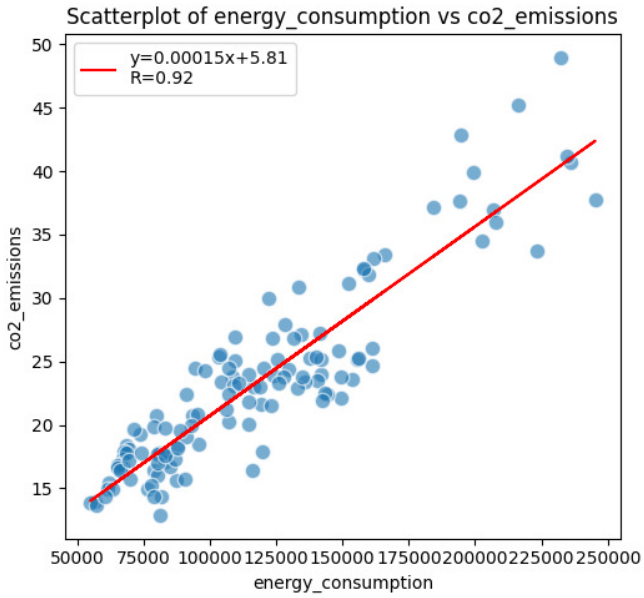
Op basis van de landen in de top 10, wordt het dataframe van CO2-emissie per capita gekoppeld aan de afzonderlijke dataframes van de potentiële voorspellers. Vervolgens worden alle waardes uitgezet in een scatterdiagram en aangevuld met de regressielijn (inclusief de functie en de correlatiecoëfficiënt).

Resultaat

De potentiële voorspeller is de categorie met de grootste correlatiecoëfficiënt. Per variabele is de correlatiecoëfficiënt als volgt:

1. GDP per land, per capita, per jaar: R = 0.7
2. Primaire energieconsumptie per land, per capita, per jaar: R = 0.92
3. Dieet van landen in de wereld (alleen consumptie van dierlijke producten: R = -0.28
4. Aantal geregistreerde auto’s per 1000 inwoners van een land: R = 0.18

De categorie ‘primaire consumptie per capita’ is de grootste voorspeller van een grote uitstoot van CO2-emissie per land, uitgedrukt in CO2-emissie per inwoner. Zie onderstaande grafiek voor een grafische weergave. Let op; een correlatie is niet per se een causaal verband tussen beiden. Opvallend is ook de negatieve correlatie tussen de consumptie van dierlijke producten en de CO2 emissie. Dit kan zijn omdat de productie in een ander land plaatsvindt dan het land waar de consumptie is. Nader onderzoek is hiervoor nodig.



Notebook met code van de uitwerking in Google Colab

Alle beschreven stappen zijn terug te vinden in het Notebook Eindopdracht vraag 1. Zie onderstaande link.

https://colab.research.google.com/drive/1T90ri9j07zlDRZhFjx43t-5X7XSYje3G?usp=sharing

https://github.com/IngridDirckx/Datafiles-eindopdracht/blob/main/Eindopdracht_vraag_1.ipynb

Hoofdstuk 2 Welke landen maken de grootste stappen in het verminderen van CO2-emissie?

Dataverzameling

Voor de uitwerking van deze vraag is gezocht naar een datafile met de CO2 uitstoot per land met historische data. Historische data zijn nodig om te kunnen meten of er sprake is van een verbetering of een verslechtering.

Het aantal inwoners van een land heeft invloed op de CO2 uitstoot. Daarom moet er voor de beantwoording van deze vraag ook rekening worden gehouden met deze het aantal inwoners in een land. Om deze reden is gekozen voor de datafile: “CO2 emissions per capita” (bron: “Our World in Data”). In deze file worden de CO2-emissies gedeeld door het aantal inwoners **en** is er historische data beschikbaar.

Opschonen en bewerken datafile

De tabel is geïmporteerd als een dataframe in een Jupyter Notebook (Google Colab) voor verder analyse en bewerking.

De eerste stap in de bewerking is een analyse op de kolommen. Overbodige kolommen zijn verwijderd. De titels van de overgebleven kolommen zijn voorzien van een kortere naam.

Een analyse van de rijen laat zien dat niet alleen landen zijn opgenomen, maar ook subtotalen (bijvoorbeeld de werelddelen en groepen op basis van inkomen) en totalen (“world”). Deze (sub-)totalen zijn verwijderd uit de selectie. Daarna zijn alleen de rijen met data uit het jaar 2021 en het jaar 2022 geselecteerd. De overige rijen met oudere jaren worden niet gebruikt in de analyse.

Na het opschonen van de kolommen en rijen is het resultaat een tabel met drie kolommen: “Entitity, Year, Co2 capita. Deze tabel is vervolgens gecontroleerd op ontbrekende waarden. Dit is niet het geval, dus actie is hier niet nodig.

Analyse

Om de vraag te beantwoorden welke landen de grootste progressie hebben geboekt moet een vergelijking worden gemaakt tussen jaren. Om een ranglijst te kunnen maken van landen met de grootste CO2-reductie, worden 2 jaren met elkaar vergeleken. In de vraag is geen periode gespecificeerd, daarom is gekozen voor de meest recente periode die beschikbaar is. Namelijk de jaren 2021 en 2022.

Om een vergelijking te maken tussen de CO2 uitstoot tussen 2021 en 2022 moet een tabel worden gemaakt met drie kolommen: land; CO2 uitstoot in 2021; CO2 uitstoot in 2022.

Hiervoor zijn eerst, door een copy te maken van df_capita_recent, twee nieuwe tabellen gemaakt, een met alleen 2021 waarden en een met alleen 2022 waarden. Deze twee tabellen zijn vervolgens samengevoegd tot een nieuwe tabel (merged_df).

In de tabel merged_df is een nieuwe kolom toegevoegd met de procentuele verandering tussen 2021 en 2022. De tabel geeft nu het antwoord op de vraag welk land de grootste progressie heeft laten zien.

Er is gekozen om een top 10 van landen op te nemen die de grootste afname in CO2-emissie hebben gerealiseerd in 2022 t.o.v. 2021. Deze landen zijn opgenomen in de grafiek.

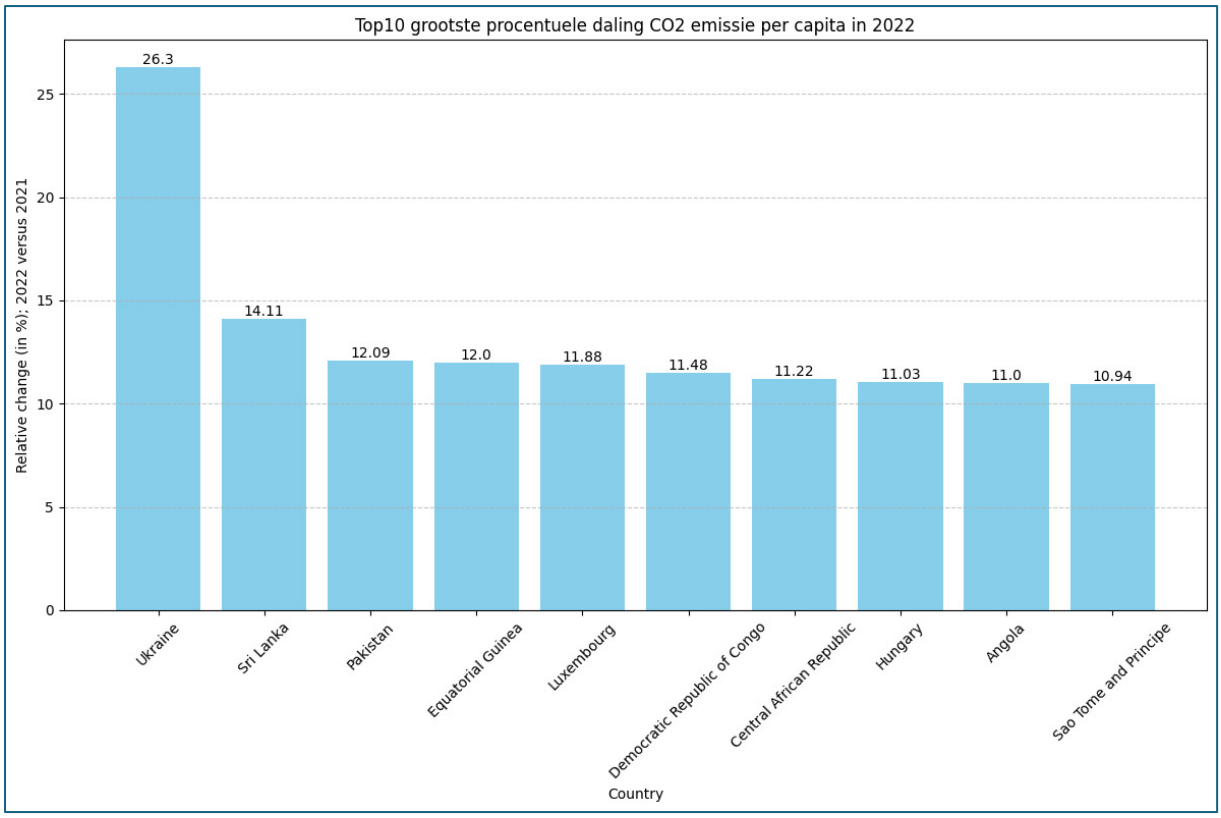
Resultaat

Het resultaat van de analyse is onderstaande top 10 van landen met de grootste CO2-reductie in 2022.

1) Oekraïne
2) Sri Lanka
3) Pakistan
4) Equatoriaal-Guinea
5) Luxemburg
6) Democratische Republiek van Congo
7) Centraal Afrikaanse Republiek
8) Hongarije
9) Angola
10) Sao Tomé en Principe

Zie onderstaande grafiek voor de grafische weergave van de procentuele progressie. Omdat de verschillen tussen de landen klein zijn, is ervoor gekozen om de waarden (procentuele verandering 2022 t.o.v. 2021) op te nemen in de grafiek. De grootste daler van CO2-emissie in 2022 is Oekraïne.

Het risico van een vergelijking van slechts twee meetmomenten (2021 en 2022) is dat er sprake kan zijn van een incidentele waarden. Bijvoorbeeld een land dat jaarlijks een reductie realiseert, behalve in de periode 2021 en 2022. In het specifieke geval van de Oekraïne kan de inval van Rusland (24-2-2022) een rol hebben gespeeld. Dit is niet nader onderzocht.



Notebook met code van de uitwerking in Google Colab

Alle beschreven stappen zijn terug te vinden in het Notebook Eindopdracht 2. Zie onderstaande link.

https://colab.research.google.com/drive/1YajfjISy8BYg5Gv6JDkdLBdhmg3oaXQg?usp=sharing

https://github.com/IngridDirckx/Datafiles-eindopdracht/blob/main/Eindopdracht_vraag_2.ipynb

Hoofdstuk 3 Welke non-fossiele energie technologie heeft in de toekomst de beste prijs?

Dataverzameling

Voor de uitwerking van deze vraag is gezocht naar een datafile met non-fossiele energiebronnen, de prijs van deze energiebronnen, inclusief historische prijzen. Historische data zijn belangrijk om een goede voorspelling te kunnen doen met behulp van lineaire regressie.

De volgende file is voor de uitwerking gebruikt: “levelized cost of energy by technology” (Bron: “Our World in Data”).

De prijzen in deze tabel zijn gecorrigeerd voor inflatie. Dit heeft als voordeel dat de prijsontwikkeling tussen de jaren met elkaar vergelijkbaar zijn.

Opschonen en bewerken datafile

De tabel is geïmporteerd als een dataframe in een Jupyter Notebook. De kolomnamen zijn gewijzigd naar kortere namen. De kolom met de landcode is verwijderd.

Voor deze analyse zijn de waardes van de individuele landen niet van belang. Daarom is ervoor gekozen om alleen de data van ‘world’ te gebruiken. Alle rijen welke ongelijk zijn aan ‘world’ zijn daarom verwijderd uit de selectie.

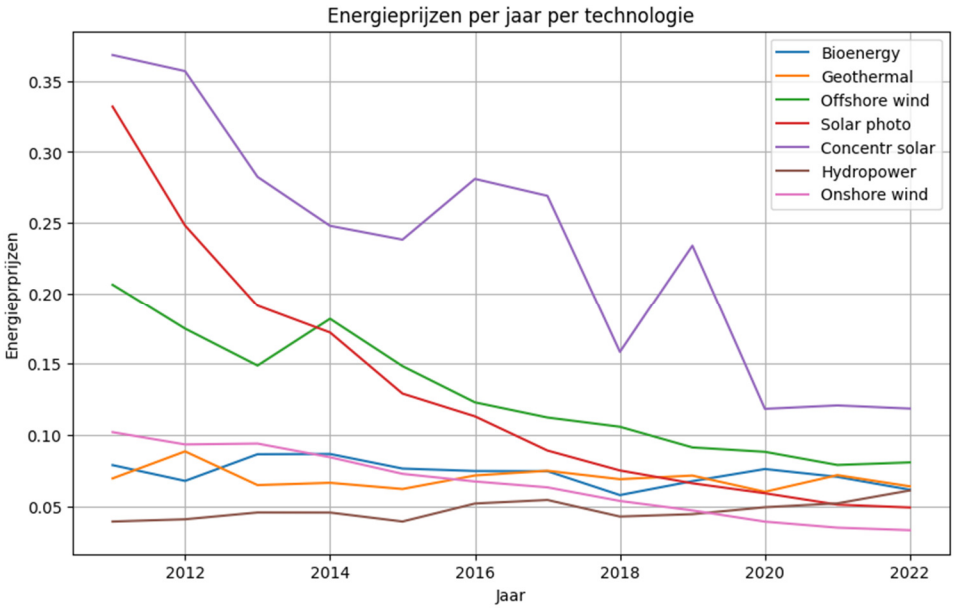
In de dataframe df_world zit nog een groot aantal NaN waarden. Een analyse laat zien dat tot 1999 van alle vormen van non-fossiele energie, de prijzen ontbreken (met uitzondering van ‘on shore windenergie’). Omdat de informatie bovendien relatief oud is worden de jaren 1984 – 1999 uit de tabel verwijderd.

Na deze filtering wordt zichtbaar dat ook in de jaren 2000 – 2010 de data bij de meeste vormen van energie nog ontbreekt (energie waarvan de prijzen (deels) wel aanwezig zijn, zijn: ‘onshore wind’, ‘geothermal’, ‘off shore wind’). Vanaf 2010 zijn wel alle prijzen beschikbaar. Daarom wordt de bovenstaande filtering aangepast naar 2010. Dus alle waarden uit de jaren 1984-2010 worden verwijderd. In de data van 2011 – 2022 is er nog slechts 1 null waarde (‘geothermal’). Deze waarde is vervangen door de gemiddelde prijs van ‘geothermal’.

Het resultaat is een tabel met energieprijzen per technologie over de periode 2011-2022. Dit is het uitgangspunt voor een extrapolatie.

Analyse

Gestart is om de waarden uit te zetten in een lijngrafiek. Dit geeft inzicht in de aanwezigheid van mogelijke uitschieters en hiermee de betrouwbaarheid van een voorspelling. Zie onderstaande grafiek.

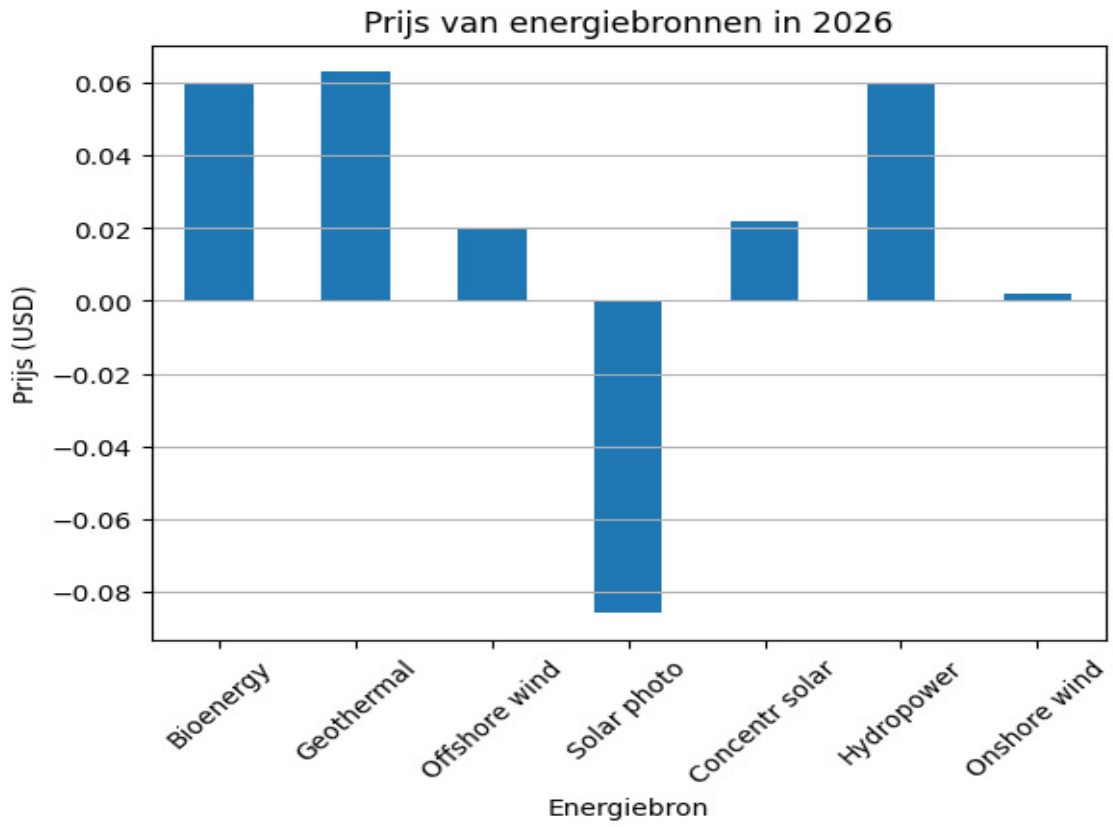


In de grafiek is te zien dat de prijzen van ‘Concentrated solar power’ erg fluctueren. Besloten is om alle waarden te ongewijzigd te laten en om de bestaande waarden mee te nemen in de vervolganalyse.

Voorspelling van de toekomstige prijzen

De vraag die moet worden beantwoord betreft een voorspeling in de toekomst. Er is gekozen om de prijzen voor het jaar 2026 te voorspellen. Op basis van de prijzen in de jaren 2011 – 2022 is een extrapolatie uitgevoerd om de toekomstige prijzen voor het jaar 2026 te berekenen.

De resultaten van de uitgevoerde analyse zijn opgenomen in onderstaande grafiek



De berekende toekomstige prijs van de energie technologie ‘Solar photovoltaic’ is negatief. Het is onwaarschijnlijk dat de toekomstige prijs negatief zal zijn. Daarom is besloten om deze energiebron verder buiten de analyse te houden. Het advies is om deze energiebron nader te onderzoeken in een vervolgonderzoek, gezien de huidige lage prijs en de dalende prijsontwikkeling.

Van de overgebleven zes energiebronnen heeft de technologie ‘Onshore wind’ de laagste verwachte prijs in 2026.

Notebook met code van de uitwerking in Google Colab

Alle beschreven stappen zijn terug te vinden in het Notebook Eindopdracht 3. Zie onderstaande link.

https://github.com/IngridDirckx/Datafiles-eindopdracht/blob/main/Eindopdracht_vraag_3.ipynb

https://colab.research.google.com/drive/1NkX_hYzlZ3KUMGThGNwYZP8EbtUJu4KW?usp=sharing