World University Ranking

Data Analysis and Visualization

Charlotte Kaandorp - 11800232
Guilly Kolkman - 11822465
Nienke Reints - 11899239
Tobias Teule - 11790091
Tutor: Nora Schinkel
28 juni 2018

Inleiding

De Universiteit van Amsterdam (UvA) wordt gezien als een van de beste universiteiten in Nederland en wereldwijd staat de UvA in de top 100. Maar wat zijn de eigenschappen van een goede universiteit? En hoe verhouden deze eigenschappen zich tot elkaar?

In dit verslag wordt beschreven hoe er onderzoek is gedaan naar de verbanden tussen eigenschappen van de universiteiten en de verschillen in jaren uit de Times Higher Education's World Ranking [4]. Met deze dataset en de interesse naar de eigenschappen van goede universiteiten zijn er drie deelvragen opgesteld. Ten eerste wordt er inzicht verkregen in de data door op zoek te gaan naar benoemingswaardige verschillen en/of afwijkingen. Ten tweede wordt er gekeken naar de consistentie van de verschillende universiteiten, landen en continenten. Als laatste worden verdere verbanden gezocht in de data van de Times Higher Education's World Ranking [4].

Deze drie vragen worden beantwoord door de data op de website van Times Higher Education's World Ranking te analyseren. In deze data is de ranking van de universiteiten gebaseerd op een *overall score*. De verdere opbouw van de data wordt uitgelegd in de methode.

Om de data, dat gepubliceerd staat op een website, te kunnen analyseren wordt deze eerst van de website gehaald en omgeschreven naar een bestand dat eenvoudig uit te lezen is (scrapen). Vervolgens kan met behulp van Numpy, Bokeh en Sklearn de data geanalyseerd worden. Alle drie deze hulpmiddelen zijn opensource-uitbreidingen van python. Als eerste wordt met Numpy ingewikkelde berekeningen uitgevoerd. Daarnaast kunnen met Bokeh grafieken geplot worden, hierdoor kan de data inzichtelijk worden gemaakt. Als laatste wordt met behulp van Sklearn lineaire regressie toegepast.

Methode

Allereerst werd door middel van het scrapen met *Beautiful Soup*[2] van de Times Higher Education World University Rankings data verzameld uit de jaren 2016, 2017 en 2018. Hier waren de universiteiten gerangschikt op basis van *overall score* die is opgebouwd uit 30% *teaching*, 30% *research*, 30% *citations*, 7.5% *international outlook* en 2.5% *industry income*. Deze scores zijn wederom opgebouwd uit verschillende gegevens. Allereerst is er *Teaching*, deze is opgebouwd uit de resultaten van een enquête over de reputatie van de universiteit, verhouding medewerkers-studenten, verhouding doktoraten-werknemers, verhouding doktoraten-bachelors en tot slot subsidies. *Research* is ook opgebouwd uit een enquête over de reputatie, daarnaast wordt er gekeken naar het inkomen gegenereerd door onderzoek en het aantal gepubliceerde artikelen. Verder is *citations* gebruikt om vast te leggen hoeveel invloed een universiteit heeft op onderzoek in het algemeen. Dit wordt gemeten door te kijken naar hoe vaak gepubliceerde artikelen worden geciteerd door geleerden wereldwijd. Daarnaast is de *international outlook* die rekening houdt met de verhoudingen buitenlandse studenten tegenover binnenlandse studenten en buitenlandse werknemers tegenover binnenlandse werknemers en naar hoeveel procent van de gepubliceerde artikelen een buitenlandse coauteur hebben. Als laatste is de *industry income* gebaseerd op hoeveel een universiteit kan bijdragen aan innovaties, uitvindingen en consultancy en hoeveel geld een universiteit hiermee verdient.

Naast deze scores en percentages was er informatie over het aantal studenten, de verhouding werknemers-studenten, het aantal internationale studenten en de geslachtsverhouding. De locatie van de universiteiten stond aanvankelijk bij de naam, maar deze is voor bruikbaarheid van de data verplaatst naar een eigen kolom.

Daarna werd de data opgeschoond door missende waardes in te vullen en uitschieters te onderzoeken. Zo waren er 4 universiteiten (Washington University, Griffith University, Czech technical university en AGH Poland) waar de ratio verkeerd stond omdat er geen correcte data was doorgegeven. Ook Anadolu university is uit de dataverwerking gehaald wanneer er werd gekeken naar het aantal studenten. Dit omdat Anadolu gemiddeld 1.2 miljoen studenten heeft. Een overzicht van hoe de missende data is ingevuld is de vinden in tabel 1.

Lege waarde	**Oplossing**	**Reden**
Ratio	Gemiddelde van andere jaren waar mogelijk, anders gemiddelde van de kolom	De geslachtsverhouding verandert niet drastisch in 3 jaar, dus zou een gemiddelde een goede benadering zijn.
Overall	Per universiteit berekend	De formule en de waardes waarmee de overall score werd berekend waren bekend
Industry income	Gemiddelde van de kolom	Per universiteit per jaar lagen deze waardes erg uit elkaar, daarom is er geen benadering kunnen maken en is het gemiddelde van de kolom aangehouden.

Tabel 1: Veranderingen in data

Vervolgens is er extra data toegevoegd, namelijk: het continent waar de universiteit zich bevindt, de grootte op basis van studenten (klein ≤ 5000, $5000 <$ medium ≤ 15000 en groot > 15000) en of de

universiteit in een Engelstalig land staat (Australië, Canada, Nieuw-Zeeland, Verenigd Koninkrijk, Verenigde Staten, Ierland en Jamaica). Ook is er gebruik gemaakt van een tweede dataset[1] van alle universiteiten op de wereld en het bijbehorende land. Tot slot is er door middel van de Google Geocoding API de coördinaten van alle universiteiten verkregen.

Om een algeheel overzicht te krijgen van hoe de universiteiten verspreid zijn over de wereld is er een kaart gemaakt. Hiervoor is de Google Maps functie in Bokeh gebruikt. Vervolgens zijn de verkregen coördinaten geplot in deze kaart. Er is onderscheid gemaakt tussen de top universiteiten en overige universiteiten per jaar.

Voor het onderzoeken van de consistentie werd er gekeken naar de gemiddelde *overall score* van elke universiteit op basis van alle drie de jaren. Vervolgens is de standaarddeviatie per universiteit berekend met de *mean* en *std* build-ins van Numpy. Dit is ook gedaan per land en continent. Er is het gemiddelde genomen van *overall score* per land voor elk jaar, op basis van het gemiddelde van de jaren is de standaarddeviatie berekend. Voor het onderzoeken van de continenten is dezelfde aanpak gehanteerd als bij de landen, alleen werd het gemiddelde per land gebruikt in plaats van het gemiddelde van de universiteiten. Dit werd gevisualiseerd met een *scatterplot* en trendlijn.

Om de data te analyseren voor afwijkingen van universiteiten zijn er radar plots[3] gemaakt. In deze plots werden zes variabelen die invloed hadden op de rang bekeken. Dit waren: research, teaching, citation, industry income, international outlook en overall. De plots die zijn bekeken zijn de top 200 van alle jaren. Als een universiteit een afwijking had werd hier naderhand dieper in gekeken wat de exacte afwijking was. Deze afwijkingen zijn geordend in, stijging van rang, daling in rang, inconsistentie bij eigenschappen en niet aanwezig in bepaalde jaren. Ook is er onderzoek gedaan naar de oorzaak van de afwijkingen.

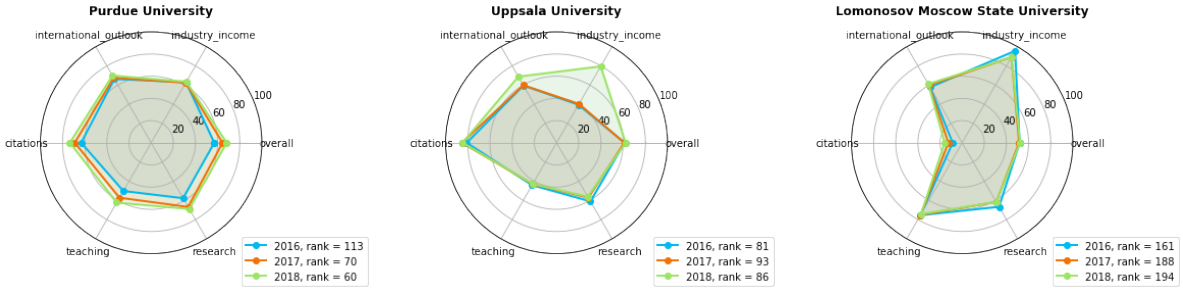
Bij het zoeken naar verbanden is er allereerst gekeken naar de volledige dataset voor elk jaar. De gegevens van alle kolommen zijn als *scatterplots* weergegeven voor overzicht. Vervolgens waren deze variabelen ook tegen elkaar uitgezet voor elk jaar. Interessant ogende grafieken werden nader onderzocht door het toepassen van lineaire regressie om het verband duidelijk vast te stellen. Dit is gedaan met behulp van Numpy en Sklearn. Allereest werd de kleinste kwadraten methode toegepast met Numpy's linalg.lstsq built-in. Vervolgens werd de *mean squared error* (MSE) met Sklearn's *mean_ squared_ error* built-in berekend. Met deze error is ook de *root mean squared error* (RMSE) berekend.

Verder is er gekeken naar de gemiddelde *overall score* door de jaren heen door dit telkens tot een bepaalde rank te berekenen en vervolgens te plotten in een lijngrafiek.

Tot slot is er gekeken naar de verhouding tussen Engelstalige landen en de andere landen. Hiervoor is de *overall score* per jaar berekend voor beide categoriën en tegen elkaar uitgezet in een staafdiagram.

Resultaten

Ten eerste is er gekeken naar benoemingswaardige verschillen en afwijkingen; zoals gezegd in de methode is hierbij gebruik gemaakt van radar plots. In de bijlagen zijn de radar plots opgedeeld in vijf categorieën (figuren 6, 7, 8, 9 en 10) waarvan er in figuur 1 de drie duidelijkste zijn weergegeven.

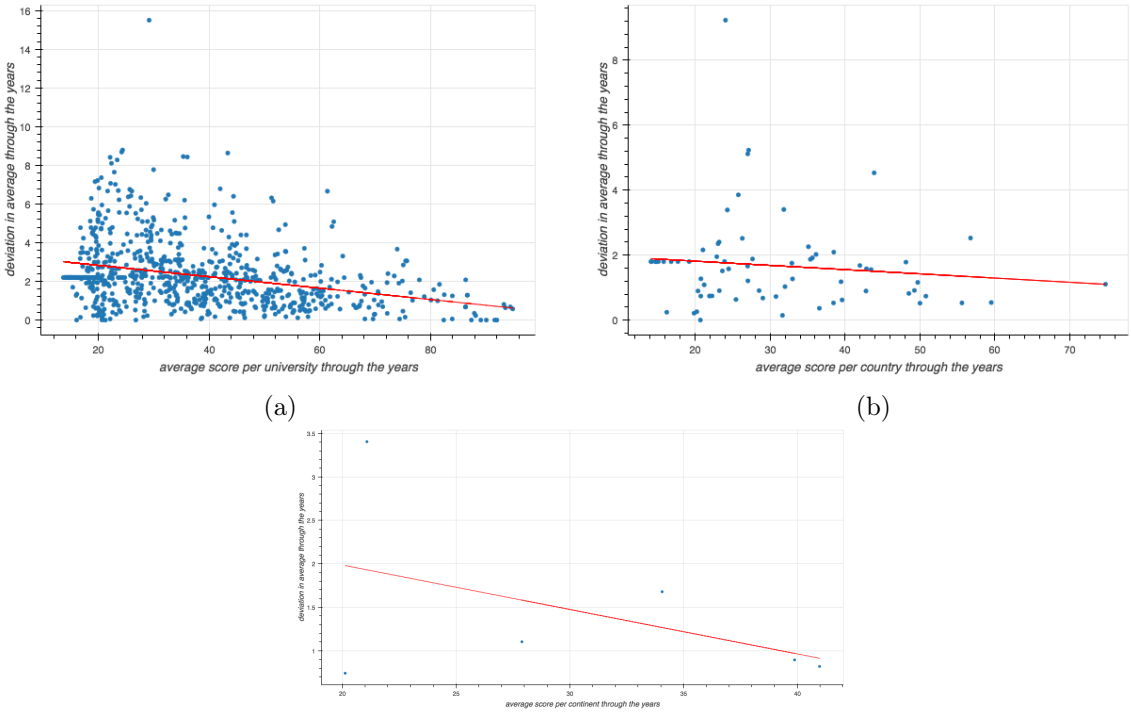


(a) Verbetering (b) Stijging industry income (c) Lage citations

Figuur 1: Radar plots

Ten eerste is in figuur 1a te zien dat Purdue University gestegen is in rank. Deze universiteit heeft in het jaar 2015 een dieptepunt gekend. Na dit dieptepunt is de universiteit zich gaan focussen op de innovatie en het uitbreiden van verschillende faciliteiten.
Ten tweede is in figuur 1b de radar plot van Uppsala University te zien. Hier valt goed op dat het industry income van deze universiteit zeer toenam in 2018. Wat hiervoor de verklaring is, is niet gevonden.
Als laatste staat in figuur 1c de verdeling voor de Lomonosov Moscow State University. In dit figuur is te zien dat de universiteit laag scoort op *citations* maar hoog op *industry income* en *teaching*. Het is bekend dat de Lomonosov Moscow State University zich ook meer focust op onderzoek met vijftien onderzoeksinstituten.

Voor het beantwoorden van de tweede deelvraag over de consistentie van universiteiten, landen en continenten is de standaarddeviatie uitgezet tegen de gemiddelde *overall score* per universiteit land en continent, zie respectievelijk figuren 2a, 2b en 2c. Meer informatie over deze figuren staat in tabel 2.



(a) (b)

(c)

Figuur 2: Standaarddeviatie

Figuur	**MSE**	**RMSE**	**Aantal datapunten**
2a	2.66	1.63	800
2b	1.92	1.38	70
2c	0.69	0.83	6

Tabel 2: Extra informatie over figuur 2

Opvallend is dat in de figuren 2a, 2b en 2c te zien is dat de trendlijn door de datapunten een dalende lijn is.

Naast het maken van de figuren 2a, 2b en 2c zijn ook de meest en minst consistente universiteiten, landen en continenten bepaald. Deze staan in de tabellen 3, 4 en 5.

Universiteit	**Score 2016**	**Score 2017**	**Score 2018**	**Standaarddeviatie**
Massachusetts Institute of Technology	92.0	92.0	92.0	0.0
University of Michigan	82.4	82.4	82.4	0.0
Chinese University of Hong Kong	54.5	61.9	67.8	6.66
Northeastern University	47.3	33.5	49.4	8.64
Peter the Great St Petersburg Polytechnic University	47.1	19.8	20.7	15.51

Tabel 3: Hoogste & laagste standaarddeviatie (universiteiten)

Land	**Score 2016**	**Score 2017**	**Score 2018**	**Standaarddeviatie**
Verenigde Staten	48.9	49.0	47.6	0.82
Zwitserland	56.2	55.1	55.5	0.52
Russische Federatie	31.7	31.5	31.7	0.14
IJsland	46.6	46.4	38.7	4.53
Portugal	31.8	27.6	21.6	5.11
Jordanië	16.2	21.8	34.2	9.23

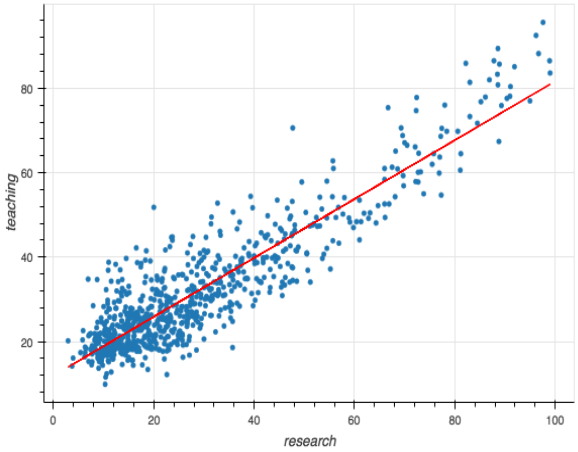
Tabel 4: Hoogste & laagste standaarddeviatie (landen)

Land	**Score 2016**	**Score 2017**	**Score 2018**	**Standaarddeviatie**
Zuid-Amerika	21.5	21.5	22.7	0.74
Noord-Amerika	48.9	49.0	47.6	0.82
Oceanië	43.8	42.0	42.6	0.90
Azië	73.7	74.65	75.9	1.10
Europa	43.9	41.0	41.0	1.68
Afrika	35.6	28.9	31.0	3.41

Tabel 5: Hoogste & laagste standaarddeviatie (continenten)

Naast de afwijkingen en de consistentie van universiteiten is als derde gekeken naar andere verbanden tussen de variabelen uit de dataset.

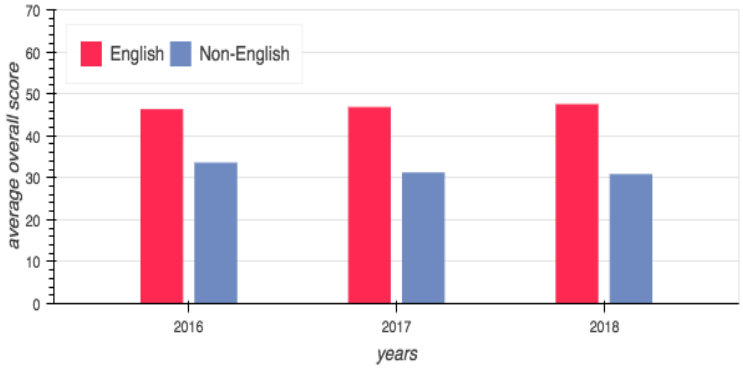
Ten eerste leidde het maken van *scatterplots* tot verder onderzoek naar teaching en research. Deze *scatterplot* inclusief lineaire regressie is te zien in figuur 3.



Figuur 3: Teaching research

In figuur 3 is de *mean squared error* (MSE) 39.75 en is de *rooted mean squared error* (RMSE) 6.30.

Ten tweede is er gekeken naar de gemiddelde *overall score* voor universiteiten uit Engelstalige en niet-Engelstalige landen. Hiervan staan de resultaten in figuur 4 en tabel 6. Vooral opvallend in figuur 4 is de gelijkenis in de verschillende jaren. Tevens is in tabel 6 toegevoegd hoeveel procent van de universiteiten zich in Engelstalige landen bevond.

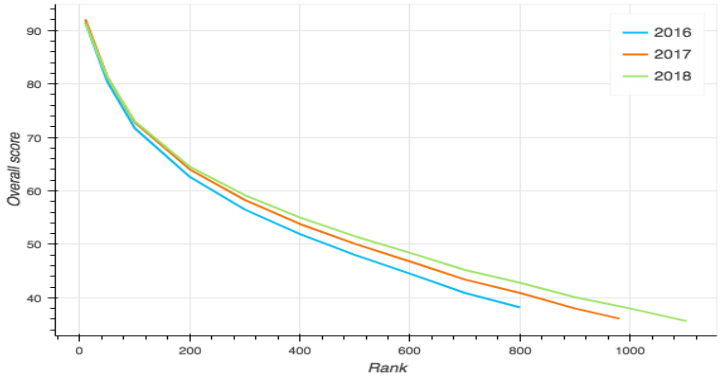


Figuur 4: Gemiddelde overall score

Jaar	**Percentage universiteiten in Engelstalige landen**	**Gemiddelde score Engelstalige landen**	**Gemiddelde score niet-Engelstalige landen**
2016	37.1%	46.16	33.34
2017	32.4%	46.69	31.06
2018	29.7%	47.39	30.63

Tabel 6: Gemiddelde overall score

Ten derde is de *overall score* tegenover de *rank* per jaar bekeken. Het verloop hiervan is weergegeven in figuur 5. In dit figuur is te zien dat de *overall score* ten opzichte van de *rank* niet in elk jaar hetzelfde is.



Figuur 5: Rank per jaar

Discussie

Bij het kijken naar benoemingswaardige verschillen en/of afwijkingen is naar voren gekomen dat er 22 universiteiten in de top 200 van de jaren 2016, 2017 en 2018 afwijkende resultaten vertoonden. Wel is er voor veel universiteiten een gemiddeld *industry income* ingevuld waar deze onbekend was. Hoewel dit in geen gevallen de rang heeft veranderd en er in de *overall scores* geen grote verschillen zijn ontstaan omdat *industry income* voor slechts 2.5 % meetelt in de berekening, kunnen er in het kijken naar afwijkende radar plots universiteiten over het hoofd gezien zijn. Ook kan er onterecht geconstateerd zijn dat universiteiten hun *industry income* sterk is gestegen/gedaald. In een eventueel vervolgonderzoek zou er duidelijk bijgehouden moeten worden welke waardes zijn vervangen door gemiddeldes. Daarnaast zou er dieper in worden gegaan op de oorzaken van deze afwijkingen.

Vervolgens is er gekeken naar de consistentie van de universiteiten, landen en continenten. Hierbij kan uit de *scatterplots* van figuur 2 de conclusie worden getrokken dat de *overall score* een correlatie heeft met hoe consistent de universiteit/ het land/ het continent is. Deze conclusie kan echter niet met zekerheid gesteld worden, omdat de standaarddeviaties bij een *overall score* van ongeveer 25 bij de universiteiten in figuur 2a ver uit elkaar liggen. Hetzelfde geldt voor de landen en continenten. Naast figuur 2 is er een overzicht van de meest en minst consistente universiteiten, landen en continenten in de tabellen 3, 4 en 5.

Als laatste zijn er op de vraag of er nog verdere verbanden in de data aanwezig zijn verschillende antwoorden gevonden.

Ten eerste is bij het kijken naar de gemaakte *scatterplots* een lineair verband gevonden tussen *teaching* en *research*, zie hiervoor figuur 3. Uit dit figuur kan de conclusie worden getrokken dat *teaching* en *research* een correlatie met elkaar hebben. Het is namelijk gemiddeld zo dat beide variabelen tegelijk toenemen of afnemen. In een vervolgonderzoek zou uitgezocht kunnen worden waarom *teaching* en *research* een sterk verband met elkaar vertonen. Hierin zou dan dieper ingegaan kunnen worden op de definities van *teaching* en *research*.

Ten tweede is er bij het kijken naar de universiteiten uit Engelstalige en niet-Engelstalige landen een consistentie gevonden in het scoren van de universiteiten. Uit figuur 4 blijkt dat universiteiten uit Engelstalige landen gemiddeld hoger scoren op de *overall score* dan universiteiten uit niet-Engelstalige landen. Daarnaast is in tabel 6 te zien dat de gemiddelde score van universiteiten uit Engelstalige landen steeg met 1.23 en de gemiddelde score van universiteiten uit niet-Engelstalige landen daalde met 2.71. Hieruit zou de conclusie getrokken kunnen worden dat de universiteiten uit Engelstalige landen zeker verbeterd ten opzichte van de universiteiten uit niet-Engelstalige landen. Toch is het zo dat dit niet valide is aangezien er geen standaarddiviatie is berekend.

Ten derde is tijdens het bestuderen van de *overall score* gevonden dat deze gemiddeld beter wordt over de jaren heen. Terwijl in de top 100 de gemiddeldes nog dicht bij elkaar liggen onstaat er een duidelijk verschil in de lager scorende universiteiten. Hieruit is de conclusie te trekken dat universiteiten het gemiddeld beter gaan doen over de jaren.Er is hier niet gekeken of specifieke univeriteiten verbeteren, alleen het totaal gemiddelde. In een vervolgonderzoek kan gekeken worden of dit daadwerkelijk zo is door soortgelijke rankings (die bijvoorbeeld op basis van andere gegevens de kwaliteit van universiteiten meet) te vergelijken.

Referenties

Referenties

[1] All universities in the world dataset. https://github.com/endSly/world-universities-csv.

[2] Beautiful soup 4 python. http://www.pythonforbeginners.com/beautifulsoup/beautifulsoup-4-python.

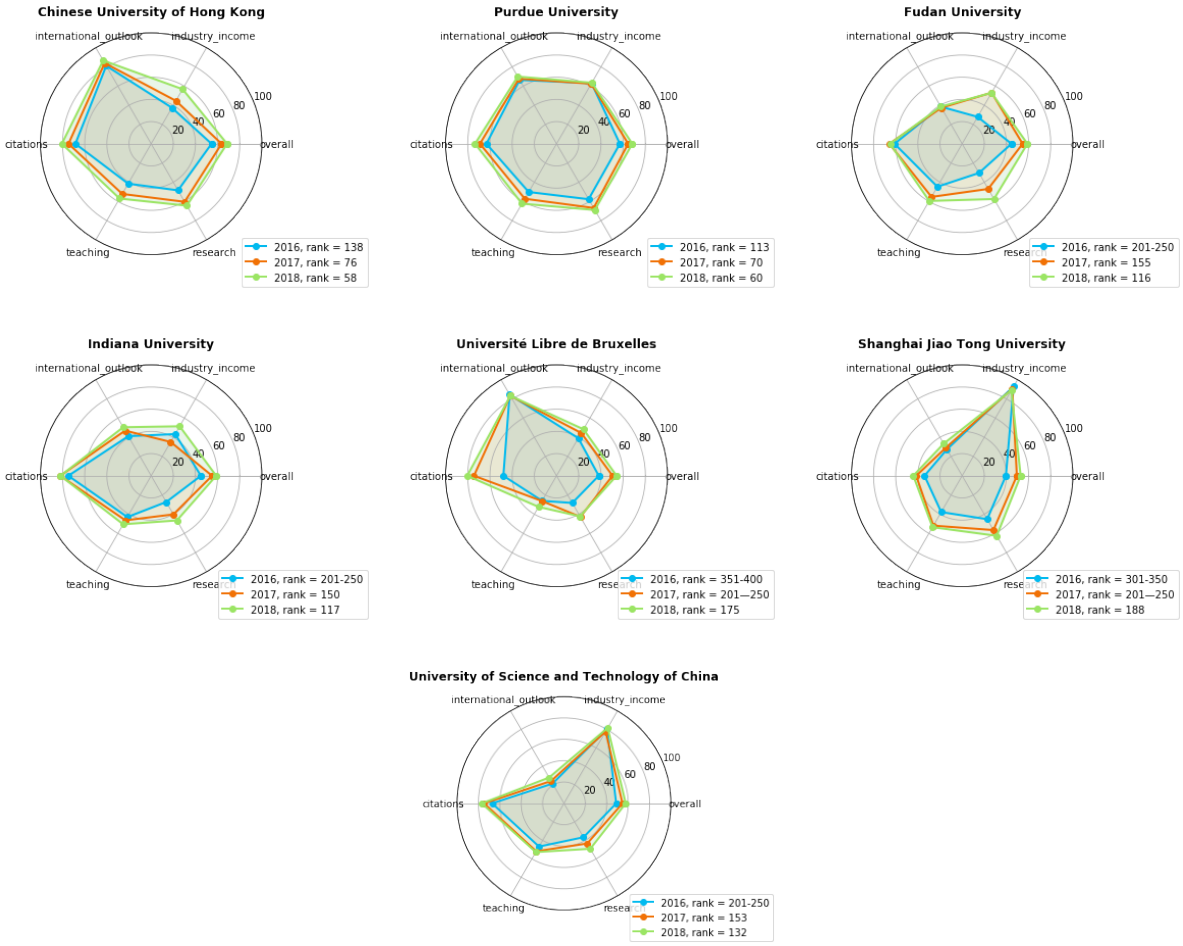
[3] Draw radar charts in python. https://www.kaggle.com/typewind/draw-a-radar-chart-with-python-in-a-simple-way.

[4] Times higher education's world ranking. https://www.timeshighereducation.com/world-university-rankings/2018/world-ranking!/page/0/length/25/sort$_b$y/$rank$/$sort_o$$rder$/$asc$/$cols$/$stats$.

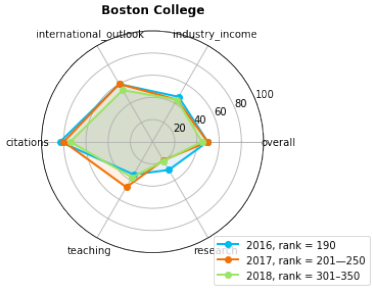
Bijlagen

Lijst van figuren

Lijst van tabellen

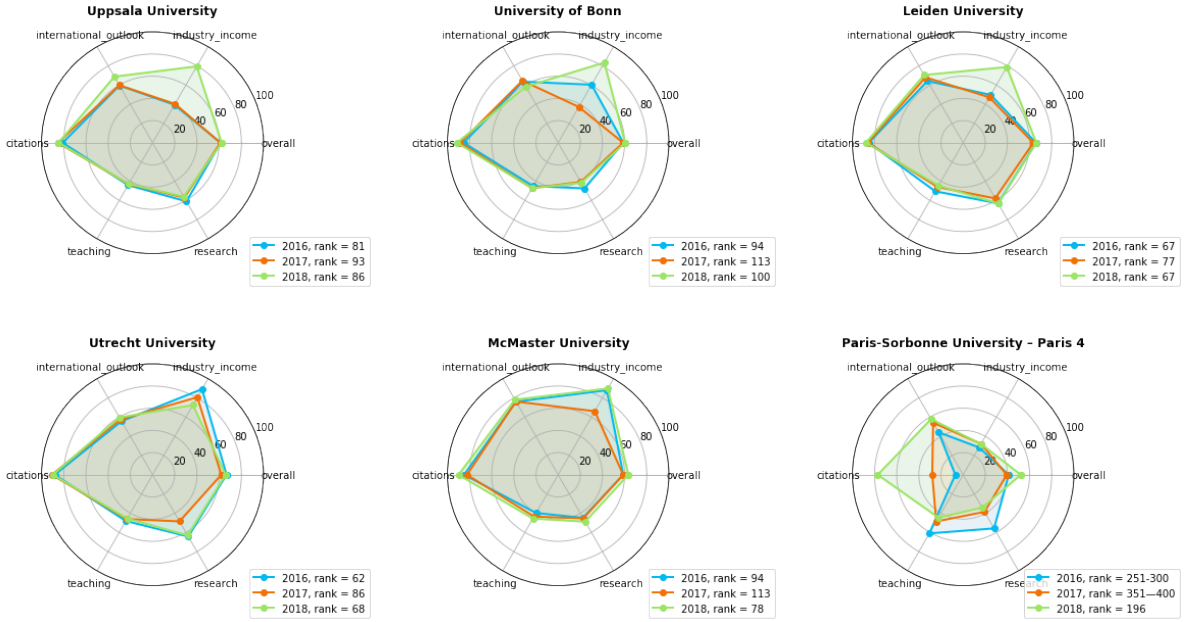


Figuur 6: Stijging in rang

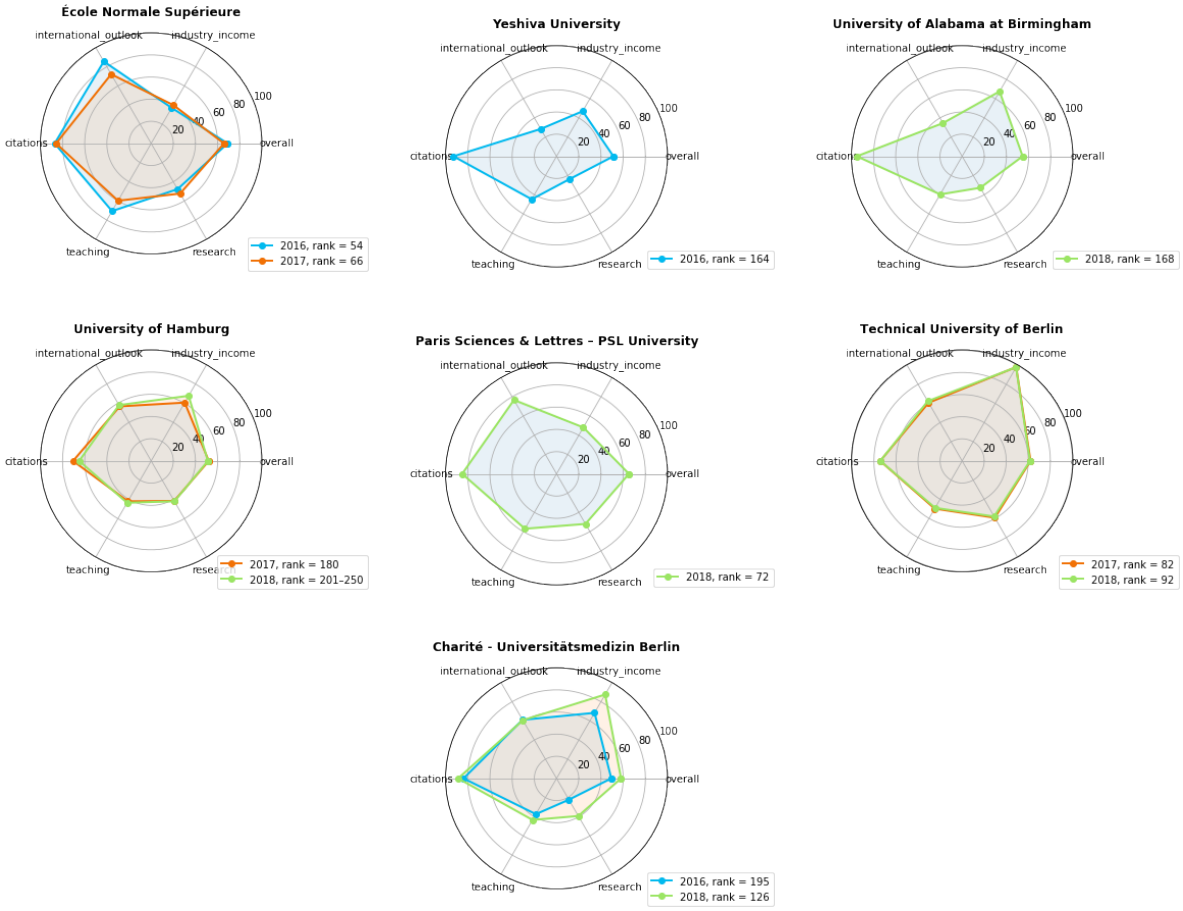


Figuur 7: Daling in rang

Figuur 8: Verassende score



Figuur 9: Inconsistent door de jaren



Figuur 10: Missende jaren