Universiteit van Amsterdam

Naam van het samen te vatten boek

Titel van het document

28 juni 2018

Studenten:
Kalle Janssen
11882107

Wijnand Potthoff
10780726

Daniel Jensen Perez
11610964

Anthony Thieme
11872934

Tutor:
Mattijs Blankesteijn

Practicumgroep:
Naam van de groep

Cursus:
Data Analyse en Visualisatie

Vakcode:
5082DAAV6Y

1 Introductie

Het genieten van een goede opleiding is voor velen van groot belang. Voor het vinden van de opleiding met de hoogste kwaliteit kijkt men vaak naar de status van de universiteit waar deze aangeboden wordt. Er zijn echter veel aspecten waarop een universiteit in kwaliteit beoordeeld kan worden. Ook is het zo dat niet alles directe invloed heeft op het niveau van de opleidingen die een instituut aanbied.

In dit onderzoek zal er worden gekeken naar de invloed van de scores van verscheidene aspecten op de uiteindelijke ranking van een universiteit. Deze scores zijn verkregen van de Times Higher Education World University Ranking [**Times**], van de jaren 2016, 2016 en 2018. Deze rangijst maakt gebruik van dertien verschillende factoren om een zo uitgebreid mogelijk beeld te kunnen geven van de status van 's werelds De vraag is of er specifieke factoren zijn die een sterke invloed uitoefenen op de uiteindelijke kwaliteit van een universiteit. Door middel van diepgaande data-analyse zal er vanuit de verkregen scores naar opvallende verbanden worden gezocht. Deze worden dan grafisch weergegeven zodat de resultaten gemakkelijk te lezen en te begrijpen zijn, ongeacht de educatieve achtergrond van de lezer.

Ook zal er worden gezocht naar de meest consistente en inconsistente universiteiten en regio's binnen deze ranglijst. Daarnaast zal er ook naar de meest opvallende en relevante verschillen in de top tweehonderd universiteiten worden gezocht, en uiteindelijk of er zich andere interessante patronen voordoen binnen de dataset. Hiervoor zal deze vergeleken worden met een dataset betreffende het bruto nationaal product van de landen waarin de gerankte universiteiten zich bevinden.

HYPOTHESE. VERWACHTINGEN.

1.1 Definities

definities??

1.2 Vraagstelling

vraagstelling??

2 Methode

De data die gebruikt is in dit onderzoek is verkregen van de website van de Times Higher Education World University Rankings [**Times**]. Op deze website wordt er vanaf 2011 elk jaar een ranglijst bij gehouden met 's werelds top universiteiten. Voor dit onderzoek is er alleen gekeken naar de top achthonderd rannglijsten uit de jaren 2016, 2017 en 2018. Er is voor de top achthonderd gekozen omdat er in de ranglijst uit het jaar 2016 maar 800 universiteiten stonden. In de jaren 2017 en 2018 waren dit er veel meer. Om de jaren zo goed mogelijk te vergelijken wordt er dus alleen naar de top achthonderd gekeken.

Voor het scrapen van de data werd er gebruik gemaakt van de Python requests package. Het opgevraagde HTML-bestand bevatte echter niet de benodigde tavel, aangezien deze interactief door JavaScript werd geladen. Er is vervolgens een webdriver gebruikt om de tabel uit te lezen en zo de relevante data te scrapen.

Omdat we veel data-punten misten in de kolommen *score_industry pct_intl_student, male/female ratio en score_overall* zochten we naar manieren om deze in te vullen. Uiteindelijk werd er gekozen om de missende data-punten in de kolommen *score_industry en score_overall* in te vullen door middel van machine learning. Hiervoor werd de python module scikit-learn[referentie naar scikit-learn] gebruikt. De data-punten werden ingevuld door middel van *multiple* lineaire regressie. Dit is een vorm van machine learning die geleend is van statistiek. Om onze *classifier* te trainen gebruikten we alle rijen in ons data-set die geen data-punten misten in de kolommen *ranking, score_overall, score_teaching, score_research, score_citation en score_int_outlook*. Deze kolommen waren de *features* voor onze classifier, met de *features* kan je proberen de waarde van een gekozen *label* te voorspellen. Deze classifier werd vervolgens gebruikt om alle rijen waarin er missende waarden waren voor de kolommen score_overall en score_industry in te vullen. Dit deed hij met een nauwkeurigheid van 99.9% en 52% respectievelijk. De *male/female ratio* bleek slechts 15 procent accuraat, en de verkregen data is dan ook niet gebruikt. Voor deze kolom is uiteindelijk de gemiddelde man/vrouw verdeling van alle universiteiten per land genomen voor elke universiteit waar deze data miste

Om een duidelijke weergave van de verbanden binnen de dataset te kunnen bieden, is er de keuze gemaakt om deze in meerdere grafische voorstellingen te verwerken met behulp van *Bokeh*. *Bokeh* is een interactieve visualisatie-bibliotheek die zich richt op moderne webbrowsers voor presentaties met als doel het bieden van elegante veelzijdige grafische voorstellingen. Er zijn met *Bokeh* een aantal plots gemaakt.Er zijn staafdiagrammen van het aantal universiteiten per continent, voor alle drie de jaren, gemaakt om de consistentie van de universiteitenvbinnen deze gebieden weer te geven. Een staafdiagram van de universiteiten per regio werden toegevoegd voor een specifiekere blik. Om een mogelijk verband in de top tweehonderd universiteiten aan te tonen is een tornadodiagram gemaakt van de verdeling tussen de mannen en vrouwen die aan deze universiteiten studeerden, en een scatterplot voor het percentage intenationale studenten per universiteit. Voor eventuele andere verband zijn er een histogram met de verdeling van de scores per universiteit gemaakt, evenals een radar plot betreffende het percentage internationale studenten, het percentage mannelijke studenten, de hoeveelheid studenten en het percentage staf per student.

Voor het maken van een interactieve wereldkaart die het aantal gerankte universiteiten per land weergeeft is geen gebruik gemaakt van *Bokeh* maar van *Plotly*. *Plotly* is net als *bokeh* een interactieve visualisatie-bibliotheek die zich richt op moderne webbrowsers voor presentaties, alleen het maken van een wereldkaart was hiermee aanzienlijk eenvoudiger.

Om dit duidelijk online weer te kunnen geven, is er op de aangewezen server de mogelijkheid gecreeerd om de informatie per werelddeel per jaar weer te kunnen geven. Ook zijn

slechts de relevantste grafieken weergegeven, om een zo duidelijk mogelijk beeld te kunnen schetsen van eventuele verbanden.

Daarnaast wordt er ook nog gekeken of een hoog Bruto Nationaal Product (BNP oftwel GDP) een positieve invloed heeft op de gemiddelde scores en ranking van universiteiten van een land. Hiervoor wordt extra dataset gebruikt van de *Cental Intelligence Agency* [**CIA**] die het GDP weergeeft. Om dit te kunnen vergelijken moet voor elk land de gemiddelde ranking en de gemiddelde scores berekend worden. Hierna wordt het vergeleken met het GDP en vervolgens weergegeven in een scatterplot.

2.1 Algoritme

2.2 Diagram

2.3 Procedure

2.4 Software en apparatuur

3 Resultaten

resultaten

4 Discussie

Vanuit de diagrammen van de hoeveelheid universiteiten per continent en per regio zijn wat opvallende trends op te merken. Zo is te zien dat er alleen in Europa (4.65%) en Oceanie (13.2%) de afgelopen drie jaar een stijging in het aantal universiteiten op de ranglijst is. Dit laat zien dat in deze gebieden de kwaliteit van de universiteit genoeg is gestegen om die van de andere regios uit de lijst te duwen. Daarentegen daalt de hoeveelheid universiteiten in Azie (-7.25%), specifiek in Oost-Azie. In de top tweehonderd universiteiten zijn helaas weinig verschillende aspecten te zien. Het precentage internationale studenten verschilt weinig, tenzij er naar de volledige ranglijst wordt gekeken, er zijn dan in de top 200 aanzienlijk meer internationale studenten dan bij de universiteiten die een lagere rank hebben. In de verdeling tussen man en vrouw is ook weinig opvallends te zien. Er zijn enkele uitschieters, maar er is geen duidelijke trend aan te wijzen, en bij de meeste universiteiten zijn er evenveel mannen als vrouwen. Wat betreft andere opvallende verbanden, blijkt dat de grote meerderheid van de universiteiten een score onder 60 uit 100 heeft gekregen, in alle drie de jaren. Hoe hoger deze score is, hoe hoger ook de ranking is. Dit laat zien dat de scores niet normaal verdeeld zijn, en het niveau van een gemiddelde universiteit waarschijnlijk lager ligt dan men verwacht. Ook is te zien dat de radar plots van de top vijf universiteiten per jaar behoorlijk overeenkomen in vorm. Uiteindelijk is er helaas geen duidelijke sterkste invloed op de rank van een universiteit naar voren gekomen. Dit kan gelegen hebben aan een tekort aan relevante plots, of er is simpelweg gewoon niet een specifiek aspect dat het belangrijkst is voor de ranking van een universiteit.

Daarnaast is er ook nog het Bruto Nationaal Product (BNP oftewel GDP) met de gemiddelde ranking en scores van universiteiten per land vergeleken. De Verenigde Staten en China zijn hier niet in meegenomen omdat het GDP van deze landen zo groot was dat dit de hele plot onoverzichtelijk maakt. Deze plot laat zien dat het GDP een positief effect heeft op de ranking en scores van de universiteiten in een land. Hoe hoger het GDP is, hoe hoger ook de ranking en scores van de universiteiten zijn.

4.1 Implicaties en aanbevelingen

Implicaties

4.2 Conclusie

conclusie

5 Referenties

Referenties

[1] Times Higher Education. *World University Rankings 2016, 2017 & 2018.* `https://www.timeshighereducation.com/world-university-rankings/2018/world-ranking#!/page/0/length/25/sort_by/rank/sort_order/asc/cols/stats.` (geraadpleegd 28 juni 2018)

[2] Central Intelligence Agency. *The World Factbook (2014).* `https://www.cia.gov/library/publications/the-world-factbook/fields/2195.html` (geraadpleegd 28 juni 2018)