# R. verantwoordingsdocument

Ik heb een dataset gekozen waaruit blijkt welke scores een land scoort op het Bruto Nationaal Geluk.

## Korte toelichting

De World Happiness Report is een jaarlijkse publicatie door het United Nations Sustainable Development Solutions Network. Het artikel bevat een globaal beeld van het bruto nationaal geluk van ieder land. De UNSDSN baseert zijn scores op de volgende punten

1: GDP per capita,

2: Social support,

3: Healthy life expectancy,

4: Freedom to make life choices,

5: Generosity,

6: Perceptions of corruption

De waardering van het Bruto Nationaal Geluk is gebaseerd op de [Cantrill ladder](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5778415/#:~:text=The%20Cantril%20Scale%20(CS)%20is,and%20quality%20of%20social%20relations.) enquête. Een groep respondenten worden dan gevraagd om te denken aan een ladder waar de best mogelijke leven die iemand kan hebben als een 10 beoordeeld kan worden en de slechtste mogelijke leven die iemand kan hebben beoordeeld kan worden met een 0. Vervolgens worden de respondenten gevraagd om hun eigen leven te beoordelen op de Cantrill ladder beoordeling waarin hun zelf aangegeven hebben wat een waarde van een 10 of een 0 heeft. Daarmee kunnen experts de enquêtes analyseren en beoordelen op welke velden de respondenten een hoge score hebben gegeven. De velden zijn economisch, psychologisch en nationale statistieken.

## Uitleg/beredenering

Er waren in totaal 5 CSV bestanden, met de jaren 2015-2016-2017-2018-2019. Ik heb ervoor gekozen om enkel een vergelijking te maken met de voorgaande recente jaren, dus 2018 en 2019. Uit de analyse wil ik achterhalen of geluk uit te drukken is in geld (GDP) van een land. Naar mijn verwachting heeft geld een enorme impact op de kwaliteit van leven voor mensen. Daarom vond ik dit zo interessant om te analyseren. In mijn R-script heb ik ook opmerkingen geplaatst zodat de lezer in 1 á 2 regels kan aflezen waarvoor de regels code dienen.

Ik gebruik de packages: “dplyr”, “ggplot2”, “ggthemes” en “caTools”.

Regel 6 & 7: Hier laad ik het CSV-bestand in mijn R-Studio project.

Regel 10 & 11: Ter controle of de datasets correct ingeladen zijn.

Regel 14 & 15: Een globaal beeld van de structuur van de datasets.

Regel 18 & 19: Hier kijk ik of er eventuele NA waardes opgenomen zijn in de datasets.

Regel 22 & 23: Een globale samenvatting van de datasets.

Sommige regels van code heb ik er enkel ingezet om spelenderwijs “R” beter te begrijpen. Daarom plaats ik soms in aanhalingstekens (spelenderwijs) tussen, dit dient enkel als conceptueel en visueel beeld over mijn kennis over “R”. Ik vermeld vervolgens ook 2 regels op, lezer wordt hiermee op te hoogte gebracht dat hij/zij dit dient te lezen als regel … EN regel … en dient dit niet te lezen als regel … TOT regel ….

(Spelenderwijs)Regel 28 & regel 46: Datasets: “df\_2018 & df\_2019” waren aflopend van de scores van een land. Dit wou ik niet en heb daarom gekozen om de datasets aflopen te maken van GDP en Score. Hiermee wou ik kijken of een land met veel geld (GDP) een hogere score (Score) had.

Regel 30 & regel 48: Van deze regels wou ik enkel een selectie maken van de kolommen van waar ik denk dat de hoofdvraag (is geluk uit te drukken in geld) wel beoordeeld kan worden.

(Spelenderwijs)Regel 33 & regel 51: Hier heb ik een nieuwe kolom gemaakt met de “mutate()” functie van R. Ik heb simpelweg alle positieve velden bij elkaar opgeteld en enkel één negatief veld (Perceptions of corruption) afgetrokken van het geheel.

Regel 35 & regel 53: Hier bereken ik de mean waarde van de kolom “Score” en ik laat eventuele NA waardes uit de berekening.

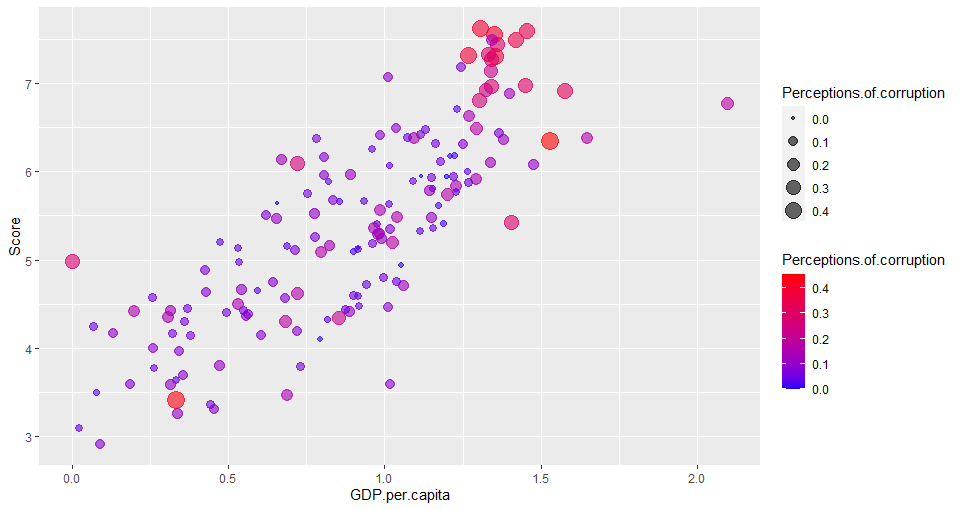
Regel 37 & regel 55: Hier bereken ik de standaarddeviatie van de kolom “Score”.

Regel 39 & regel 57: Hier bereken ik de Z score, hier kan je zien hoeveel een land afwijkt van de standaarddeviatie.

## Visualisatie

Regel 64 – regel 65: Bij figuur 1 heb ik een “scatterplot” gemaakt. Ik heb op de x as de GDP per persoon geplaatst en op de y as de Scores van een land geplaatst. Op regel 65 heb ik enkel parameters van “geom\_point” aangepast om mijn scatterplot visueel aantrekkelijker uit te laten zien. Een voorbeeld hiervan is de “aes()” functie van geom\_point. De aesthetics laag van geom\_point heb ik de cirkels de grootte (alpha=0.6) en de kleur gegeven van de kolom Perceptions of corruption. Hiermee wou ik aantonen of de respondenten die de enquêtes hebben ingevuld meer of minder het gevoel hadden van corruptie in hun land, vergeleken de hoeveelheid GDP dat een land heeft en ter idee hoe hoog een land scoorde met hun Bruto Nationaal Geluk. Uit het scatterplot kan je concluderen dat de respondenten een hoger Perceptions of corruption hadden naarmate een land meer geld (GDP) verdiende. Maar ondanks het gevoel dat er corruptie bestaat in een land, scoren de landen nog redelijk hoog op hun Bruto Nationaal Geluk. Er zijn enkele uitschieters links onderin in de scatterplot, wat betekent dat een land een lage Bruto Nationaal Geluk score heeft, weinig geld verdiend (GDP) en een hoge Perceptions of corruption heeft.

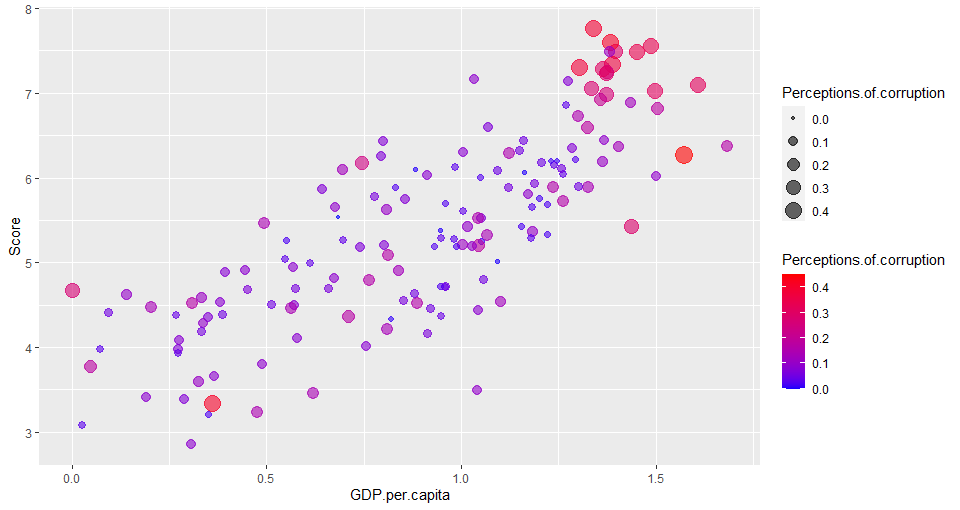




Figuur 1: Scatterplot 2018, visualisatie tbv de perceptie van corruptie, afwegend tegen de GDP per land en de algehele score

Regel 93 – regel 94: Ter vergelijking van figuur 1 heb ik bij figuur 2 een scatterplot over het jaar 2019 gemaakt. Enkel zie je in de scatterplot van 2019 geen radicale verschillen vergeleken 2018.

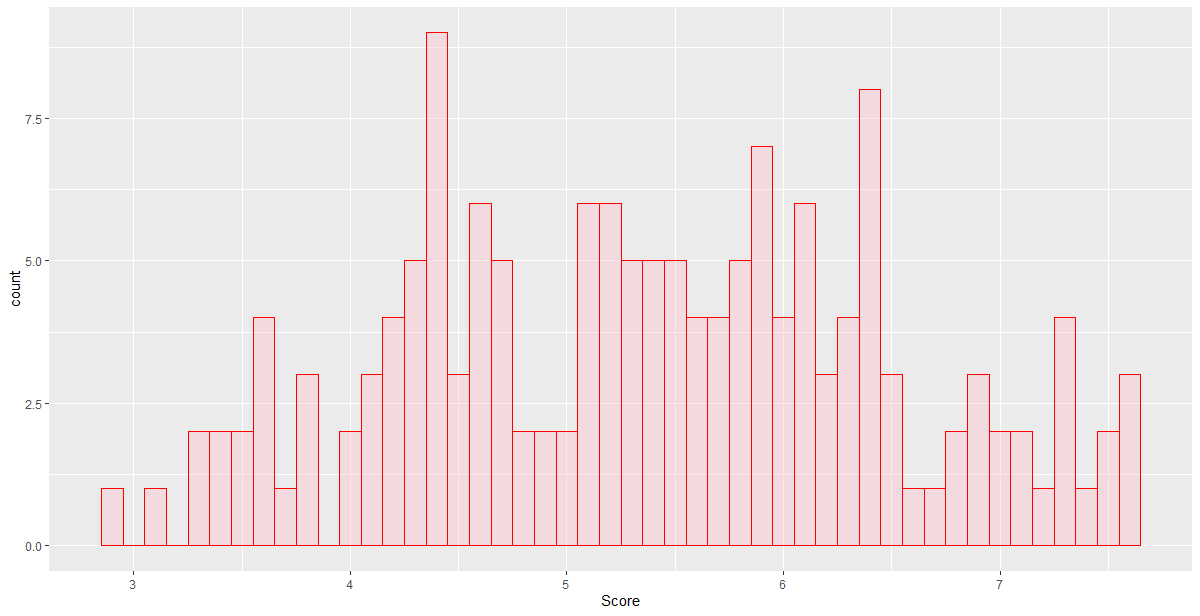




Figuur 2: Scatterplot 2019, visualisatie tbv de perceptie van corruptie, afwegend tegen de GDP per land en de algehele score

Regel 68- regel 69: Bij figuur 3 heb ik enkel een histogram gemaakt van de kolom “Score”. Ik heb de histogrammen gemaakt met de “geom\_histogram()” functie van ggplot2.

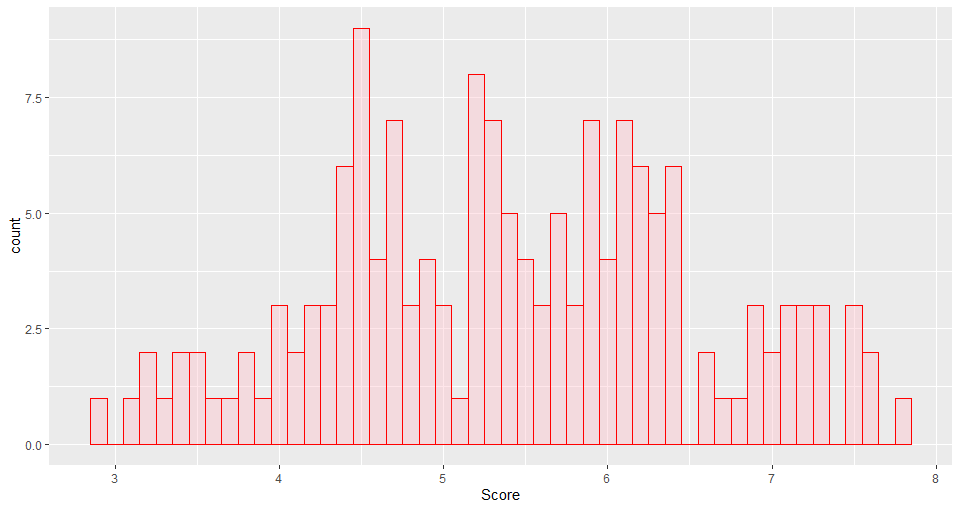




Figuur 3: Histogram 2018, waaruit de score visueel naar voren wordt gebracht.

Regel 97 – regel 98: Ter vergelijking over 2018 heb ik een histogram (figuur 4) over het jaar 2019 gemaakt. Hieruit kun je concluderen dat er over 2019 een betere score is gegeven vergeleken 2018.

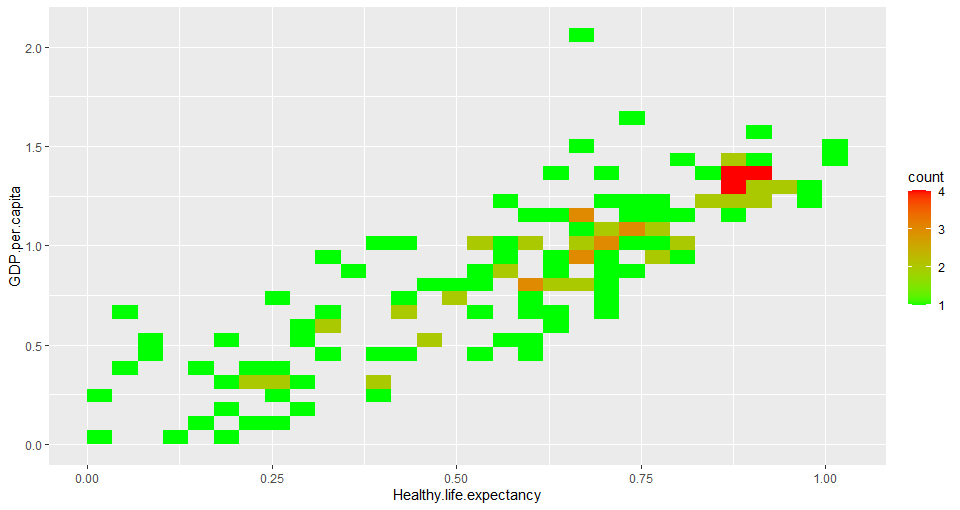




Figuur 4: Histogram 2019, waaruit de score visueel naar voren wordt gebracht.

Regel 72 – regel 73: Bij figuur 5 heb ik een “Barplot” gemaakt dat enkel fungereert als een soort heat map om te zien of de Healthy life expectancy hoger is naarmate de GDP van een land ook hoog is. Hieruit kun je concluderen dat dit wel het geval is. Een rijker land heeft een hoger levensverwachting/levensjaren. Ik heb de barplots gemaakt met de “geom\_bind2d()” functie van ggplot2. Vervolgens “scale\_fill\_gradient()” functie toegepast om het visueel aantrekkelijk te maken voor de lezer.

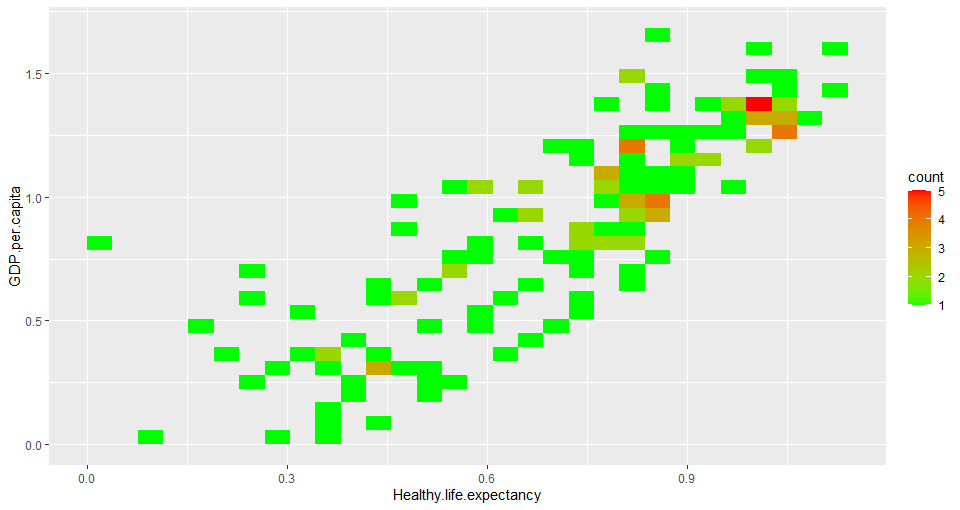




Figuur 5: Barplot 2018, wat visueel zichtbaar moet maken of de levensverwachting van iemand hoger is in rijkere landen. X as staat de Healthy Life Expectancy en op de y as staat de GDP per capita.

Regel 101 – regel 102: Ter vergelijking over 2018 heb ik een barplot (figuur 6) over het jaar 2019 gemaakt. De barplots van 2018 en 2019 zijn bijna identiek aan elkaar wat erop wijst dat er op het gebied van deze analyse weinig verschillen zitten. Je ziet enkel een lichte verschuiving van uiterst links naar rechts.





Figuur 6: Barplot 2019, wat visueel zichtbaar moet maken of de levensverwachting van iemand hoger is in rijkere landen. X as staat de Healthy Life Expectancy en op de y as staat de GDP per capita.

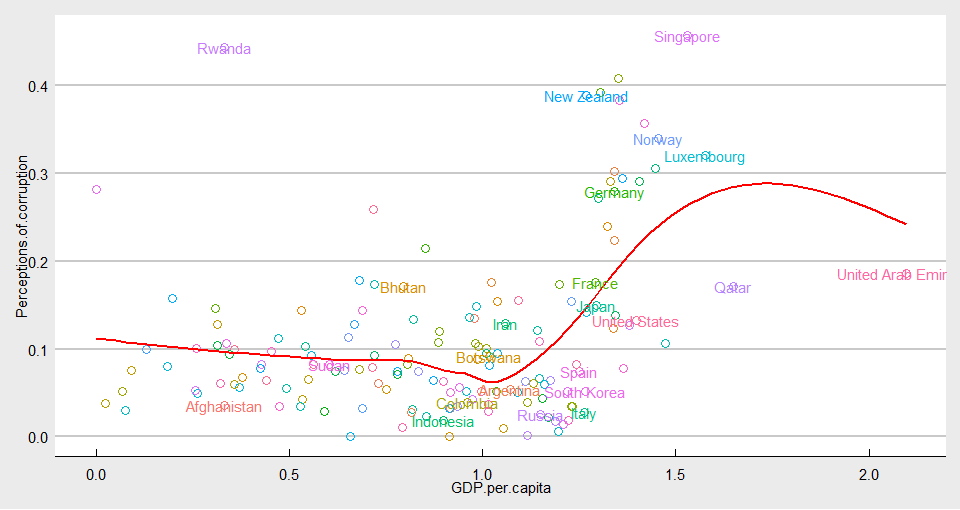
Bij het laatste visualisatie van de datasets heb ik een iets uitgebreider plot gemaakt als aanvulling op mijn eerste plot. Hier gebeurt er heel veel en hopelijk leg ik het in dit document helder genoeg uit.

Regel 84 – regel 88: Bij deze plot heb ik ieder land aangeduid als een cirkel. De plaatsing van de cirkels geven aan hoeveel een land verdiend en hoe hoog de perceptie van corruptie is in dat land. Vervolgens heb ik door middel van een lijn de gemiddelde waarde van de tabel aangegeven.

Door middel van “geom\_point()” functie van ggplot2 creëer ik alle cirkels. Vervolgens gebruik ik de “geom\_smooth()” functie om een gemiddelde lijn weer te geven van alle cirkels. In de smooth functie heb ik de formule y ~ x toegepast (dit staat vermeld in de documentatie), de reden ervan is dat de lijn zijn referentiepunten krijgt. Ik haal daarnaast altijd de legenda weg, omdat dit onwijs veel plaats neemt en daardoor is de plot minder goed zichtbaar. Vervolgens gebruik ik de “geom\_text()” functie om labels te creëren voor mijn cirkels. Ik heb op regel 76 een vector gecreëerd met namen van landen. Die vector heb ik vervolgens gelinkt aan mijn cirkels, zodat enkel de namen verschijnen bij de cirkels die overeenkomen met de namen zoals benoemd in de vector. Zoals u al kunt zien, zijn er onwijs veel cirkels en om alle namen te vermelden is niet handig. Door gebruik de maken van de “ggthemes” package en library, heb ik een theme toegepast op mijn plot. De theme heet “theme\_economist\_white()”. Dit heb ik enkel gedaan om het visueel aantrekkelijker te maken voor de lezer.

Conclusie over het plot. Zoals u kunt zien is er vanaf een bepaalde punt op de x-as van GDP.per.capita een verschuiving omhoog. Hieruit kan ik concluderen dat naarmate een land meer verdiend, de respondenten een hoger perceptie van corruptie over een land heeft.

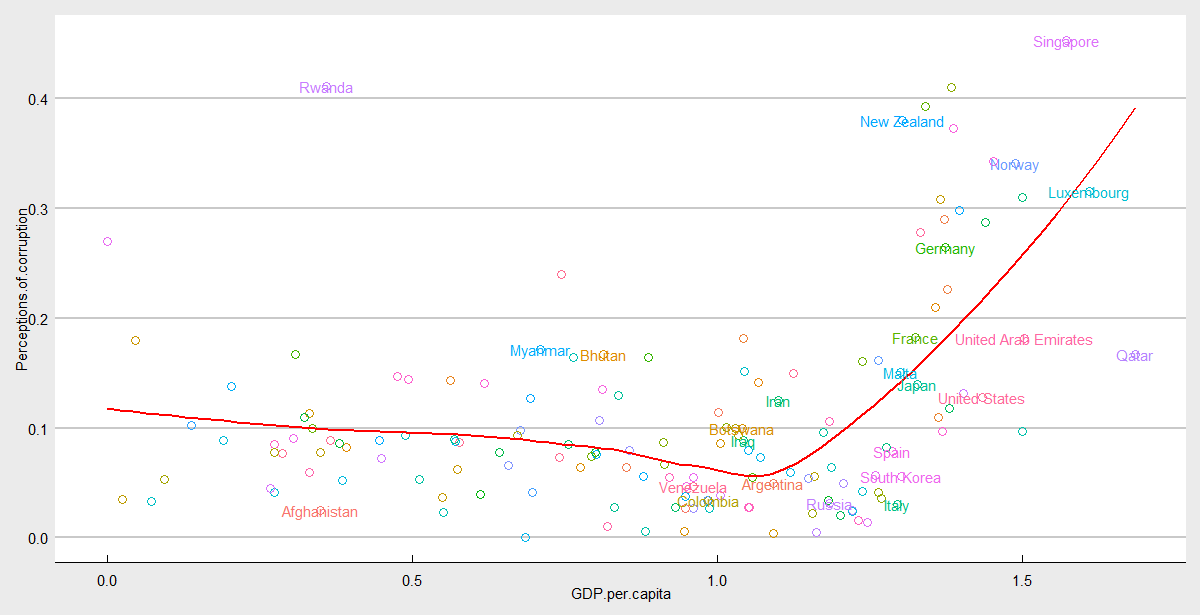




Regel 113 – regel 117: Ter vergelijking over 2018 heb ik een plot over het jaar 2019 gemaakt.

Conclusie over plot: Het interessante van deze plot is dat er in 1 jaar tijd, de perceptie van corruptie over een land heel erg gestegen is. Vergeleken 2018 is 2019 een jaar waarin meer respondenten van landen het idee hebben dat er in hun land veel corruptie plaats vond. Daar waar ik uit mijn eerste scatterplot het onderscheid niet heel goed kon maken, is het nu visueel zeer goed zichtbaar.





Machine Learning (Ik gebruik library “caTools” voor ML)

De datasets genaamd: “2018” & “2019”, geven aan de hand van 6 kolommen, een beeld over hoe een land presteert op het gebied van de Bruto Nationaal Geluk. De kolommen bestaan uit:

* Score (Hoe een land scoort),
* GDP per capita (Hoe rijk een land is),
* Social support (Perceptie dat er voor een persoon gezorgd wordt),
* Healthy life expectancy (Levensverwachting),
* Freedom to make life choices (Perceptie van vrijheid),
* Generosity (Gulheid),
* Perception of corruption (Perceptie dat er corruptie bestaat).

Naast al deze punten, heb ik een extra kolom aangemaakt (zoals te zien op regel 33 en regel 51). Deze nieuwe kolom, moet naar mijn mening de kolom “Score” onderbouwen. De nieuwe kolom genaamd “Overall Scores” zou evenredig moeten lopen met de kolom “Score”. Hoe hoger een land scoort op de “Score” kolom, des te hoger de “Overall Scores” cijfer dan wordt. De berekening van de kolom “Overall Scores” is simpel gezegd alle kolommen dat een positief effect heeft op de kolom “Score”, heb ik bij elkaar opgeteld. Zoals u hierboven kunt opmerken, heeft perceptie van corruptie niet een geheel positief effect op een mensenleven. Daarom heb ik ervoor gekozen om de perceptie van corruptie kolom van de andere kolommen af te trekken. Dit alles resulteert in het feit dat er een kolom (“Overall.Scores”) wordt gemaakt, wat de kolom “Score” onderbouwt.

Ik heb gekozen om lineaire regressie toe te passen op de datasets. Voordat ik dit heb gedaan, heb ik gekeken of de kolom “Overall Scores” en de resterende kolommen enigszins een correlatie had met elkaar. Hier komt mijn berekening van regel 33 & regel 51 aan te pas. Ik heb op deze regels een Overall Score aangemaakt, waarbij ik de kolommen dat een positief effect hebben op de dataset opgeteld bij elkaar en de kolom dat een negatief effect heeft erbij afgetrokken. Op regel 124 – regel 127 controleer ik de correlatie tussen mijn kolommen en mijn eigen gecreëerde kolom (Overall.scores).

Overall.scores

Overall.rank -0.9687329

Score 0.9746639

GDP.per.capita 0.8778592

Social.support 0.8220934

Healthy.life.expectancy 0.8545861

Freedom.to.make.life.choices 0.5600995

Generosity 0.1376585

Perceptions.of.corruption 0.3765084

Overall.scores 1.0000000

Zoals u kunt zien, is er wel degelijk een correlatie tussen de kolom “Overall Scores” en de rest van de kolommen. Daarom is het toepassen van een lineaire regressie zeker van toepassing. Dichter bij de 1, betekent dat de kolommen gecorreleerd zijn met elkaar.

Regel 134 tot regel 138: Hier heb ik de dataset opgedeeld in 2 sets. Hierbij pas ik de 80-20 regel toe, waarbij er 80% van de data beschikbaar wordt gesteld voor trainingsdoeleinden en 20% van de data wordt beschikbaar gesteld voor testdoeleinden.

Regel 141: Hier bouw ik mijn model, waarbij ik de formule van berekening in de “lm()” functie vermeld. Ik weet wat de uiteindelijke uitkomst moet zijn, waardoor supervised learning een correcte uitkomst kan bieden voor eventuele vervolg data waar ik de uiteindelijke uitkomst niet van weet. (Mocht je geïnteresseerd zijn, dan kun je de standaard meegeleverde plots (QQ-plot, Residual vs Fitted plot, Scale location-plot en Residuals vs Leverage plot) over de residuals vinden op regel 142).

Regel 145: Hier bereken ik de te verwachtte uitkomst. Ik gebruik hierbij de “predict()” functie van R. Hier plaats ik in de parameters mijn model en testdata. Mijn predict() functie weet wat ik wil berekenen, omdat ik mijn model doorgeef als parameter.

Regel 148 tot regel 153: Hier maak ik een dataframe aan met de uitkomst van de lineaire regressie.

Predicted Actual

11 10.185132 10.194

14 11.803938 11.644

17 11.535350 11.510

22 11.227272 11.159

25 10.186054 10.141

36 10.068683 10.198

40 9.711129 9.762

48 9.055456 9.018

61 7.702069 7.698

64 6.747870 6.705

65 10.009040 10.184

Dit is maar een kleine deel ervan en zoals je kunt zien, zijn de voorspellingen bijna identiek aan de daadwerkelijke berekening. De nummers van de rijen zijn niet in volgorde, dit komt omdat ik 80% van mijn data als trainingsdata beschikbaar heb gesteld en 20% ervan als testdata.

Regel 156: Hier bereken ik, hoe groot de afwijking van de lineaire regressie is.