Happy Planet Index clustering script

HarvardX Data Science Capstone Own Project

Codrin Kruijne

2019-06-03

This report has been prepared for the HarvardX Data Science program Capstone course final project submission and consists of three files of original work that can be found on GitHub (with all commits): this report in PDF, this report in R markdown and an R markdown script file with all the code (also knitted to PDF with results for convenience). Results are read in from data objects stored locally (also saved in the GitHub repository), that were the output of the script file. We were limited in our visualisations by the requirement to publish to PDF.

Summary

Humankind uses diverse progress indicators, most currently Growth Domestic Product (GDP) to guage economic growth. It is becoming more apparant that this is a limited and myopic view and alternative measures have been emerging over te last years, like the United Nation's Human Development Index. I wanted to have a closer look at two concepts that follow; the Social Progress Index (SPI) and the Happy Planet index (HPI), which take into account the state of our ecosystems. This report describes merging the last avaialbele data for both from the year 2016, applying clustering to see how certain countries group, and having a look at their differences. Resulting clusters show that social progress seems to come at the cost of ecological systems that cannot sustain levels of consumption.

Analysis

Raw data

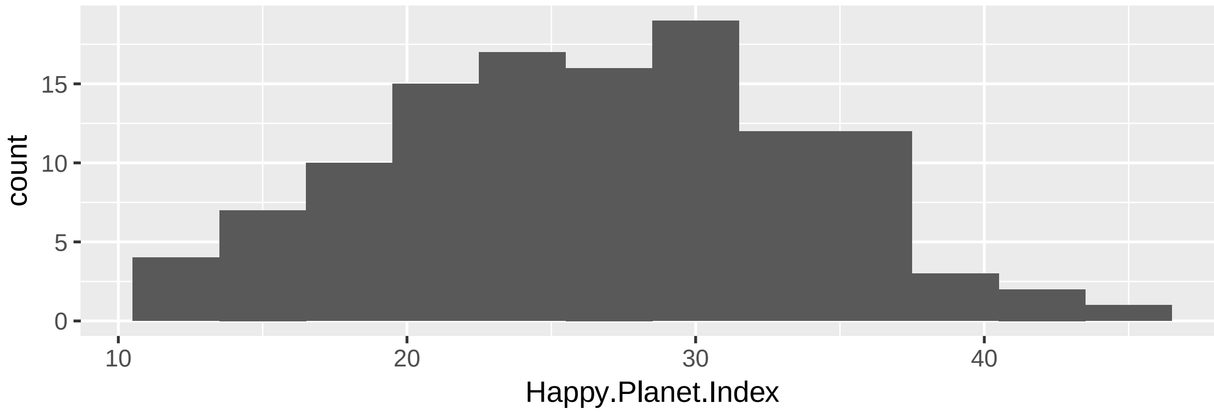
Data was downloaded from the official websites as Excel files and the last common year, 2016 was read in. HPI data in short looks at the amount of happy life years peole in a country achive, divided by the ecological footprint or 'ecosystem costs' to obtain that to balance means and ends. SPI data includes more diverse indicators with the main categories being Basic Human needs, like health and safety, Foundations for Wellbeing, like access to knowledge and quality environment and, last, Opportuniy with freedom, rights and inclusiveness.

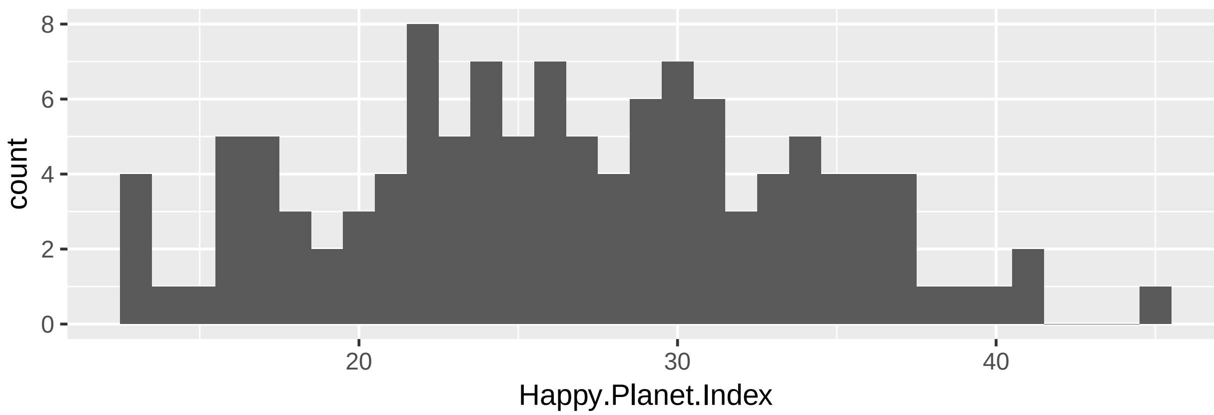
Data preparation

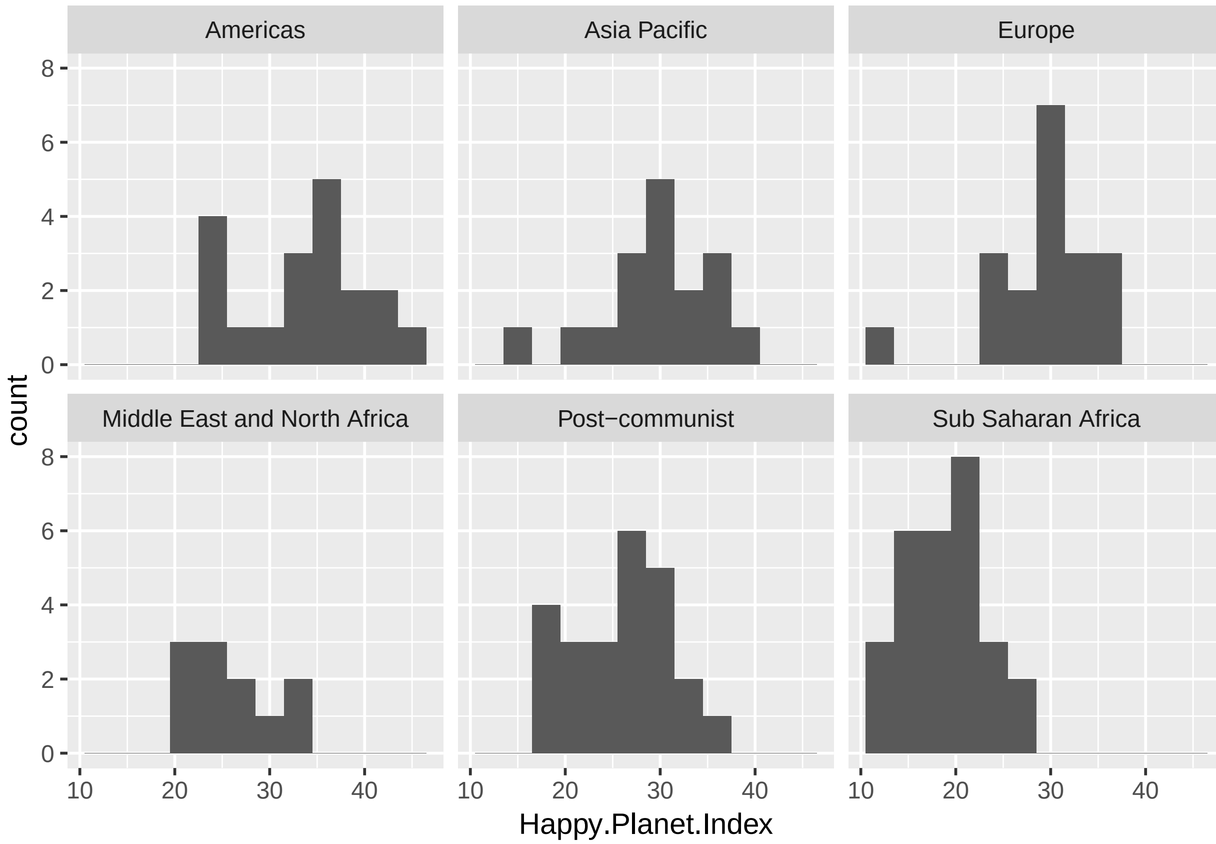
Data from HPI and SPI were joined by Country name. Missing values imputed by mean values. All numeric variables were scaled in preparation for clustering. This leaves us with 118 observations of 86 variables. We explore the data next.

Data exploration

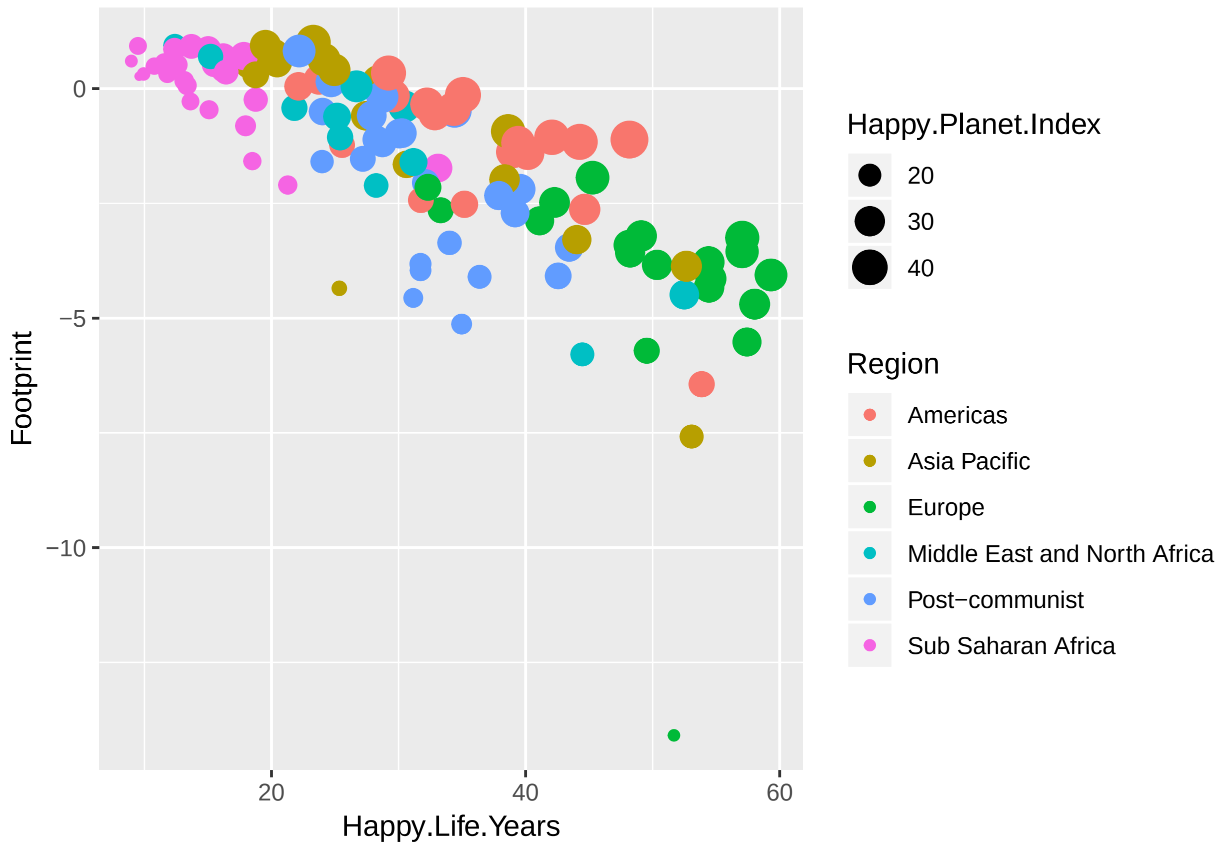
To get an initial view of the data we plot histograms of HPI scores with different bin withs and a scatter plot with major variables.


count
15
10
5
0
10
20
30
40
Happy.Planet.Index


count
8
6
4
2
0
20
30
40
Happy.Planet.Index



Americas
Asia Pacific
Europe
Middle East and North Africa
Post−communist
Sub Saharan Africa
count
0
2
4
6
8
10
20
30
40
Happy.Planet.Index

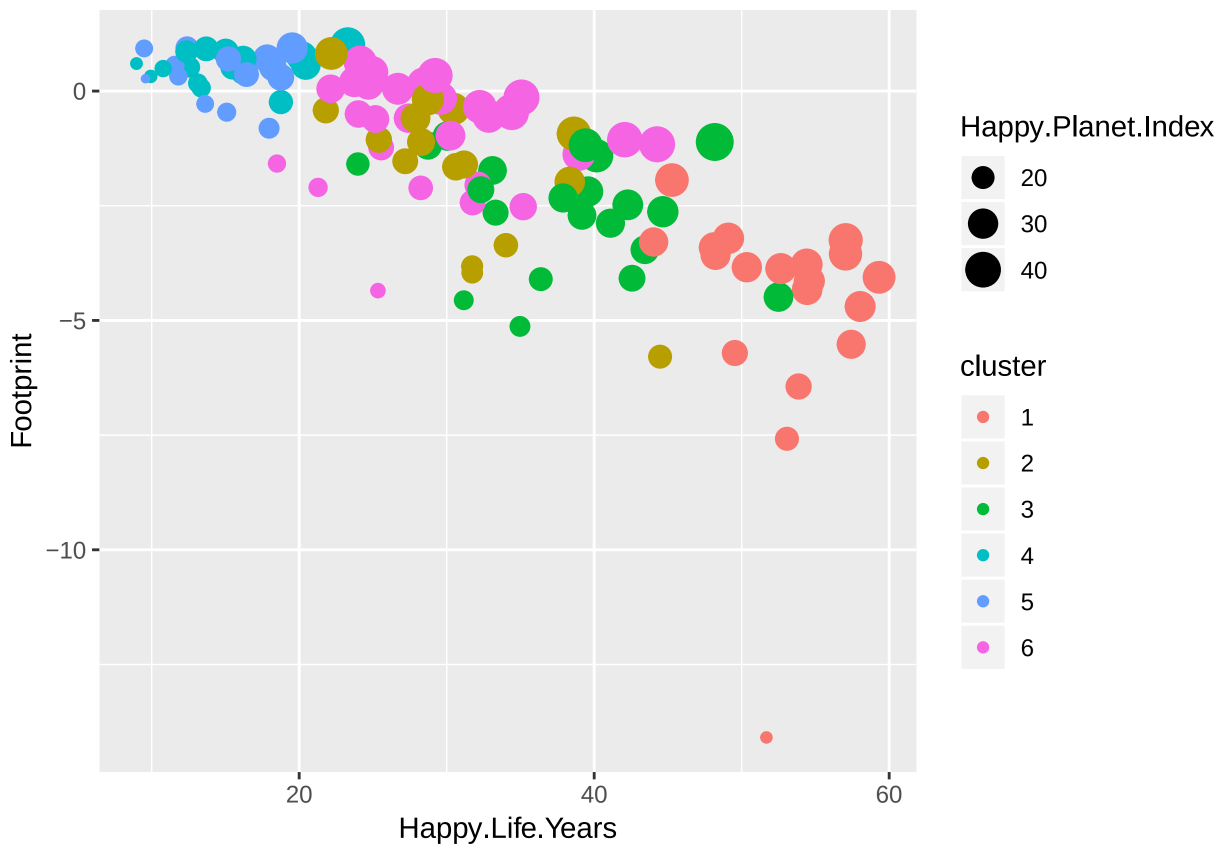


Looking at all HPI scores the histogram seems to show a somewhat normal distribution. The HPI data is already labeled with regions and separating them clearly shows differences, most notably contrasts like the low scores in Sub Saharan Africa and the distributed, yet high scores of the Americas.

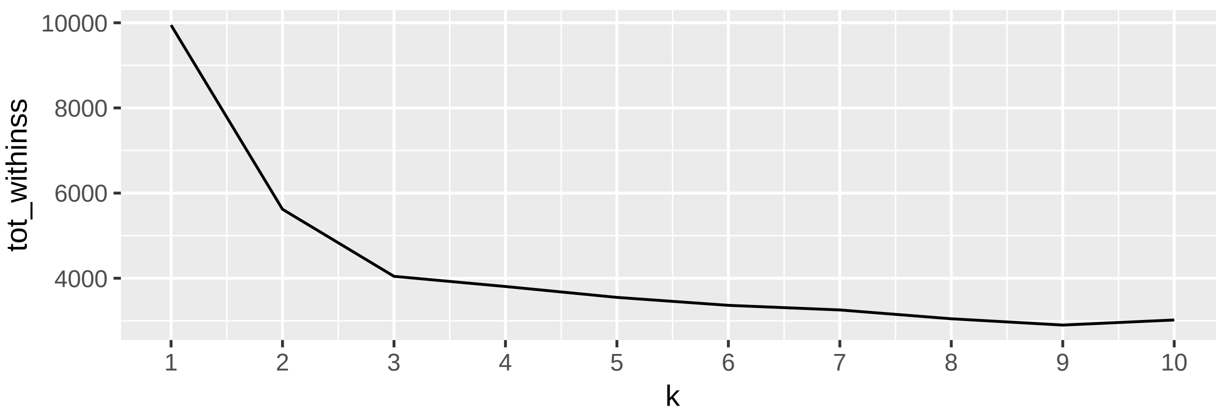
The scatterplot shows HPI on the x-axis and the effective footprint on the y-axis. Footprint data represents how big an area of ecosystems is needed to realise the consumption levels in that country (corrected for many things). 1.73 is the theoretical amount our wolrd would be able to sustain. I chose to show the difference on the y-axis; a positive value means countries consume within planetary boundaries; a negative value shows a comsumption rate requires more ecosystem capacity than is available. Here you see differences and the underlying values with notabley Sub Saharan africa and Europe at opposite ends. As we have more data from the SPI I wanted to explore what clustering based on all variables beyond region would yield for insights.

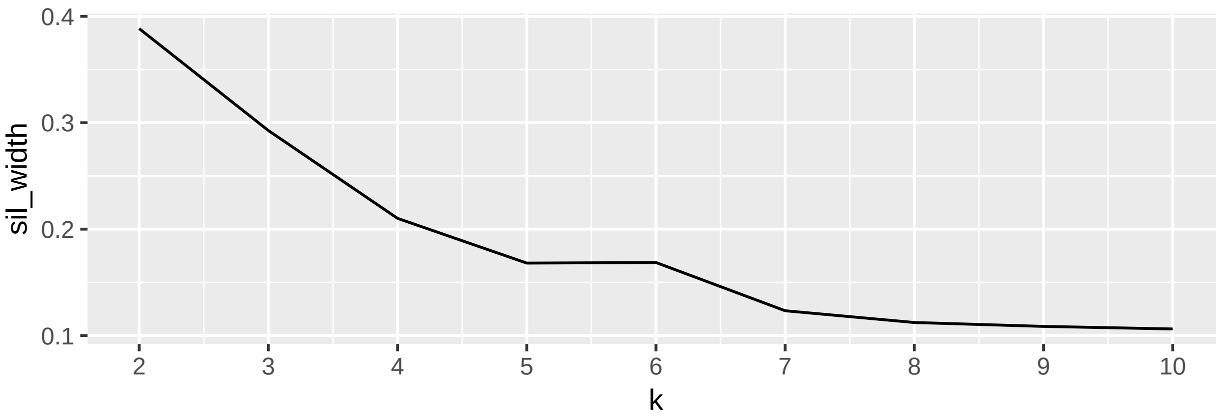
Clustering modeling

As the data provides six regions, we use that for the amount of initial clusters, after which we will calculate the most appropriate amount of clusters. We used K-means clustering.



These clusters differ quite a bit from the geographical regions. Lets calculate the optimal amount of clusters using the Elbow method and Silhouette width.



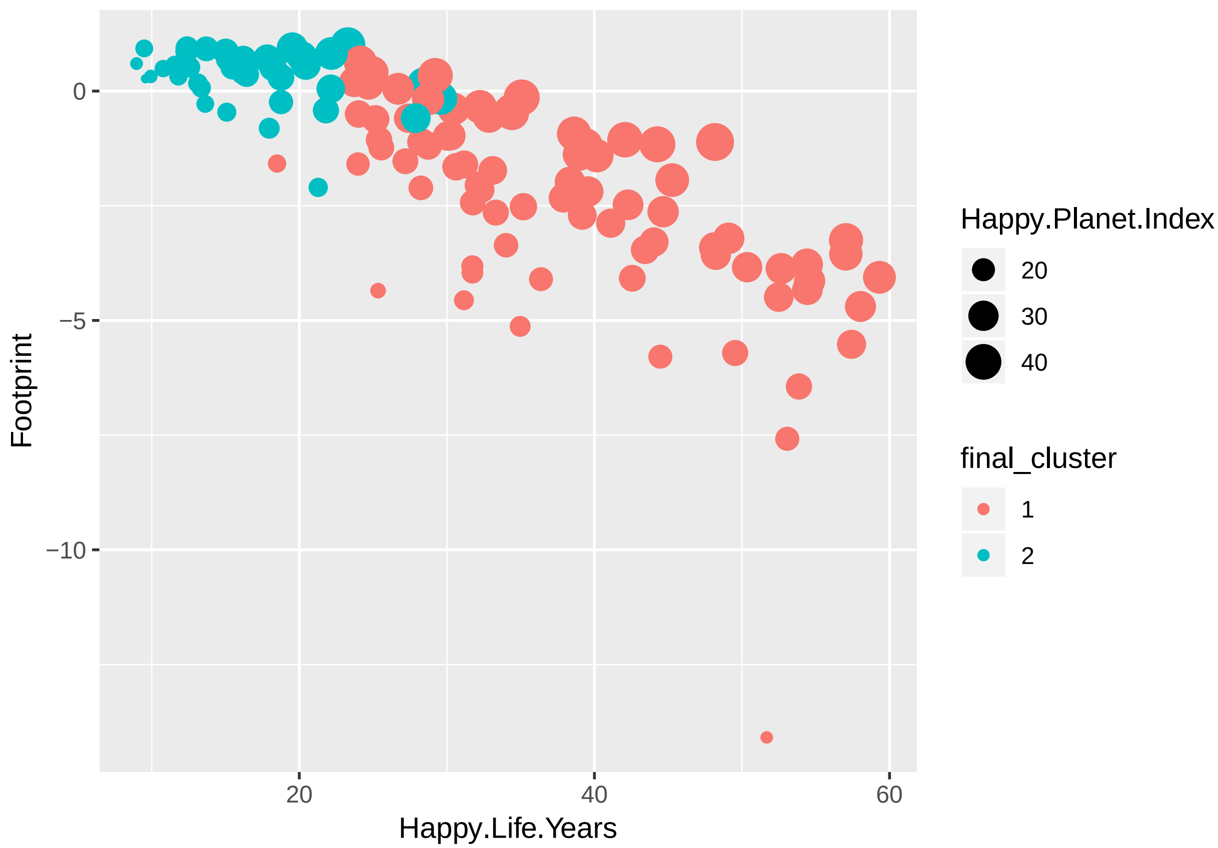


Fitting several cluster models shows two clusters seems to be the best separation. At k = 2 we see the elbow bend and the average silhouette with is highest when data is separated in two.

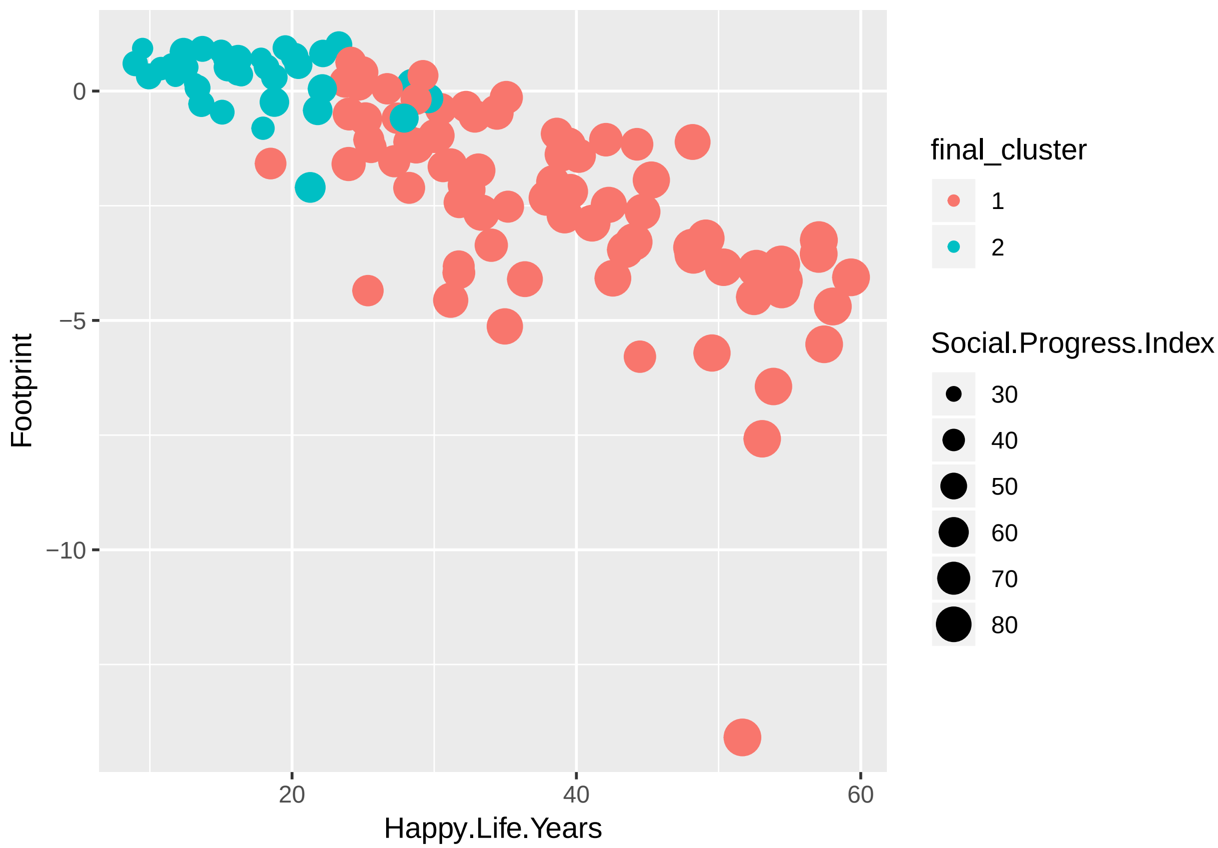
Results

Visualising clustered data

The figures below show the data with clusters identified by color and size representing first HPI and then SPI.



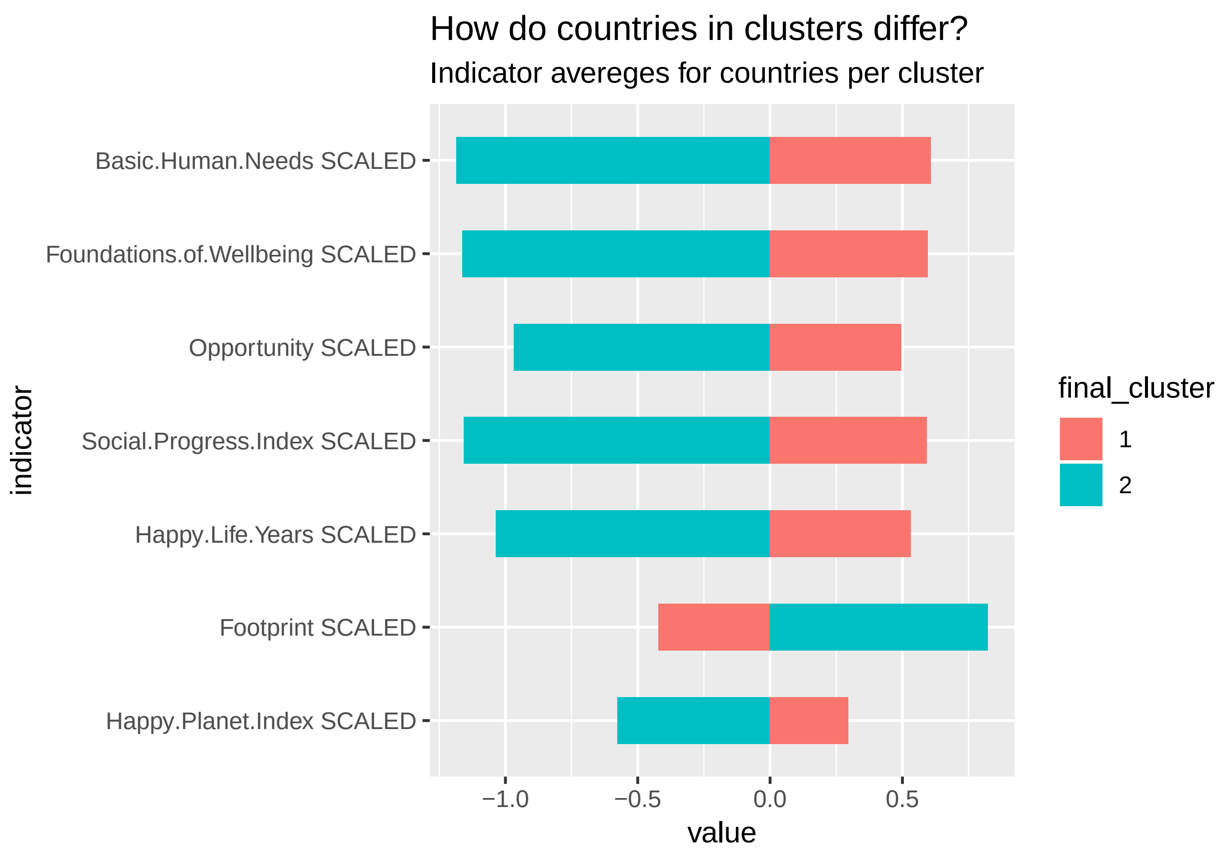
Footprint
0
−5
−10
20
40
60
Happy.Life.Years
Happy.Planet.Index
20
30
40
final_cluster
1
2



The first cluster seems to have most coutries with a positive footprint, meaning their comsumption can be sustained, while the second cluster has increasingly Happy Life Years and more negative footprint. Using HPI as size for datapoints accentuates that the lower happy life years and the higher negative footprint, the less is HPI score, whereas SPI score visually seems similar putting an emphasis on social progress.

Cluster analysis

To better understand the differences in countries we plot the average scores per cluster on the major components of both the HPI and SPI indicators.



Conclusions and discussion

We can see a clear distinction between clusters on average scores of a select number of indicators. Countries in cluster 1 score above average except for their environmental footprint; they are doing wel socially, but at the cost of our ecological systems. Countries in cluster 2 score well below averages, except for their footprint. Although this graph shows relative performance of clusters, a major concern is the high nimber of countries with negative net footprint; meaning we are comsuming not only the services of out ecosystems, but the production capacity itself, which will lead to increasing footprint pressure. Further analysis into all measures may provide additional insights into differences in these clusters. Particularly interesting woud be to explore countries with positive net footprint.

References

1. How is the Happy Planet Index calculated?
2. Social Progress Imperative

Appendix

Cluster 1	countries
Albania	Latvia
Algeria	Lebanon
Argentina	Lithuania
Armenia	Luxembourg
Australia	Macedonia
Austria	Malaysia
Belarus	Mauritius
Belgium	Mexico
Bhutan	Mongolia
Bolivia	Montenegro
Brazil	Morocco
Bulgaria	Netherlands
Canada	New Zealand
Chile	Nicaragua
China	Norway
Colombia	Oman
Costa Rica	Panama
Croatia	Paraguay
Cyprus	Peru
Czech Republic	Philippines
Denmark	Poland
Dominican Republic	Portugal
Ecuador	Romania
El Salvador	Russia
Estonia	Serbia
Finland	Slovakia
France	Slovenia
Georgia	South Africa
Germany	Spain
Greece	Sri Lanka
Hungary	Suriname
Iceland	Sweden
Iran	Switzerland
Ireland	Thailand
Israel	Tunisia
Italy	Turkey
Japan	Ukraine
Kazakhstan	United Kingdom
Kyrgyzstan	Uruguay

Cluster 2	countries
Afghanistan	Lesotho
Bangladesh	Liberia
Benin	Malawi
Botswana	Mauritania
Burkina Faso	Mozambique
Burundi	Myanmar
Cambodia	Nepal
Cameroon	Niger
Chad	Nigeria
Comoros	Pakistan
Djibouti	Rwanda
Egypt	Senegal
Ethiopia	Sierra Leone
Ghana	Swaziland
Guatemala	Tajikistan
Guinea	Tanzania
Honduras	Togo
India	Uzbekistan
Indonesia	Yemen
Kenya	Zimbabwe