World Happiness – How happy was your country in 2015?

Data Exploration

Data Source

Merging different datasets:
- United Nations Report
- Our World in Data
- World Bank

32 variables, 132 countries

Variables

Quality of Life · Health · Population · Infrastructure · Environment · Education · Economy · Politics

Methodology

- Data preparation
- Dimension Reduction (PCA, FA)
- Clustering
- Variable Importance (RFE)
- Linear Regression

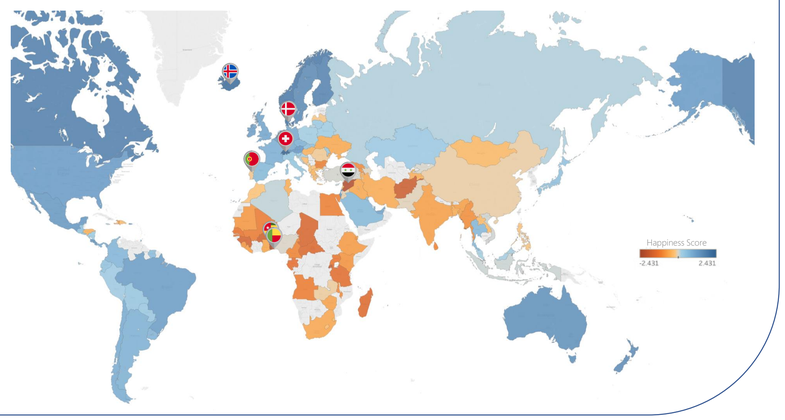
Objectives

- Analysis of happiness around the world
  - Reduce dimensionality to facilitate interpretation
  - Discover the factors that influence happiness
  - Prediction of happiness
- Analysis of similarity among countries

World Happiness Rank

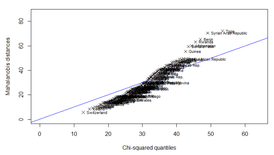
Switzerland
Iceland
Denmark
...
Benin
Syria
Togo

The worlds happiest countries are represented by Switzerland, Iceland and Denmark (Western Europe), whereas the least happy countries are Benin, Syria and Togo (Asia and Africa). In general, people living in the countries of Western Europe, Australia and North America tend to be happier than the people living in Eastern Europe, Africa and Asia. Portugal was ranked as number 84 out of 132, which results in being one of the least happy countries in Europe. This ranking is based on a Happiness Score that is conducted by a poll survey. People from all around the world were asked to evaluate their life on a scale from 0 to 10.

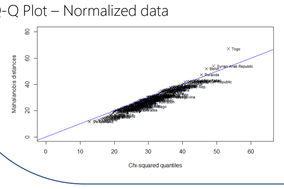


Data Pre-processing

Q-Q Plot – Assessment of multivariate normality



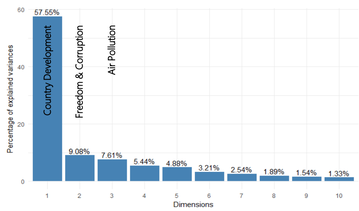
Q-Q Plot – Normalized data



Dimension Reduction

Principal Component Analysis (PCA)

PCA Scree Plot



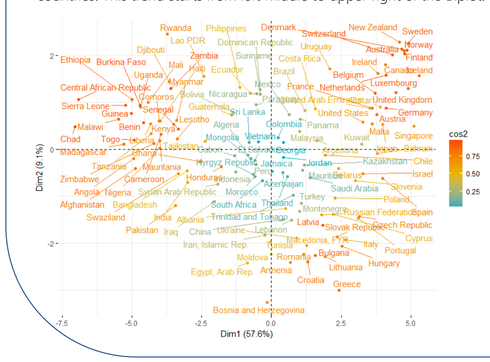
PCA Loadings

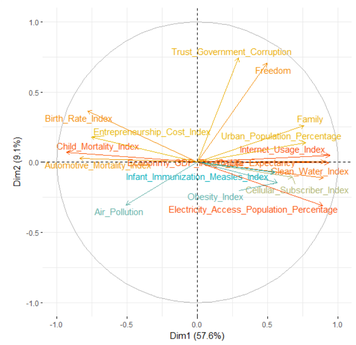
	Country Development	Freedom & Corruption	Air Pollution
Internet Usage	0.94	0.05	0.03
Economy GDP per Capita	0.93	0	0.17
Health Life Expectancy	0.91	0	-0.14
Clean Water	0.9	-0.11	-0.03
Electricity Access	0.89	-0.31	0.03
Urban Population	0.77	0.14	0.38
Family	0.76	0.26	-0.15
Cellular Subscriber	0.68	-0.11	0.37
Infant Immunization Measles	0.55	-0.07	0.11
Entrepreneurship Cost	-0.76	0.18	0.14
Birth Rate	-0.78	0.36	0.25
Automotive Mortality	-0.84	0.03	0.21
Child Mortality	-0.93	0.07	0.18
Trust Government (Corruption)	0.29	0.74	0.33
Freedom	0.5	0.7	-0.14
Obesity	0.57	-0.14	0.6
Air Pollution	-0.51	-0.3	0.54

Corr 1.0 0.5 0.0 -0.5 -1.0

PCA Interpretation

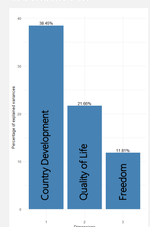
The below presented results of PCA show the quality of representation (cos2) of the countries and variables on PC1 and PC2. Finally, this dimensionality reduction has a clustering impact on Happiness Score of countries. This trend starts from left middle to upper right of the Biplot.

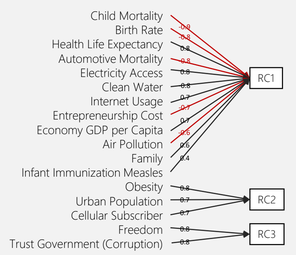




Factor Analysis (FA)

FA Scree Plot



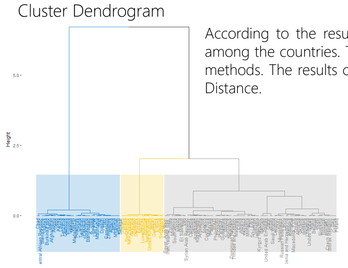


The results of FA show how the assignment of the variables to the components changes. After rotating the PC's, the second and the third dimensions explain more variance and the loadings become more clear in terms of interpretation.

Cluster Analysis

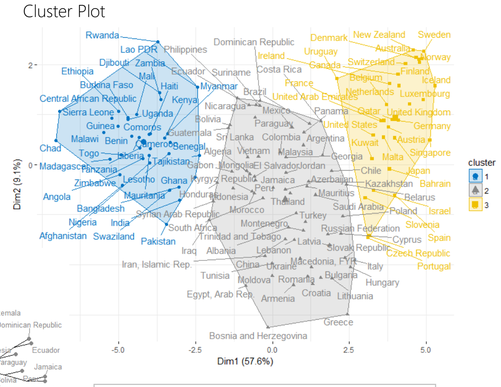
Hierarchical Clustering on Principal Components

Cluster Dendrogram

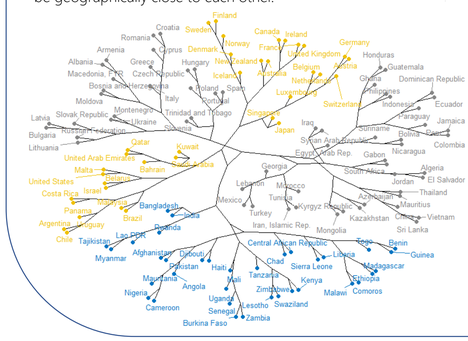


According to the results of the Dendrogram, 3 clusters seem to make the most accurate division among the countries. To proof the optimal number of clusters, we also perform Elbow and Silhouette methods. The results confirm number of clusters equal to 3. The distance measure used is Euclidean Distance.

Cluster Plot



The clusters show a geographical pattern among the grouped countries. Looking at the Phylogenic Tree, we can see that countries in the same branches tend to be geographically close to each other.



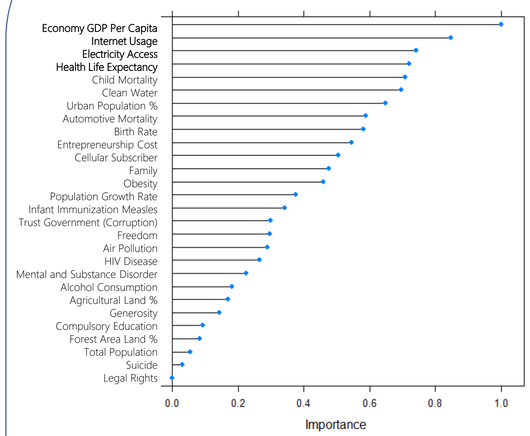
Africa
- lower Country Development
- mid to higher Freedom & Trust in Government

Western Europe, North America, Australia
- higher Country Development
- mid to higher Freedom & Trust in Government

Asia, South America, Eastern Europe
- mid Country Development
- lower to mid Freedom & Trust in Government

Predictive Analysis

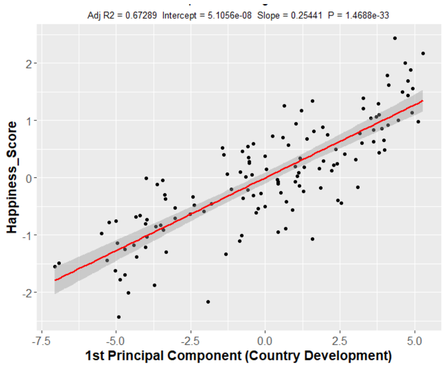
Variable Importance on Happiness Score



Performing Recursive Feature Elimination (RFE) to obtain Variable Importance provides a ranking of the variables. It shows their level of explanation in regard to the target variable Happiness Score. The first 4 variables predict happiness with R^2 = 70%. One can conclude, that the higher the GDP, Internet Usage, Electricity Access and Health Life Expectancy, the happier the people living in a country.

Economy · Infrastructure · Health

Simple Linear Regression with 1 Principal Component



Multiple Linear Regression with 2 Principal Components

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \qquad i = 1,2,\dots,132$$

R-squared: 75%
Significance: PC1 and PC2 are highly significant on Happiness Score.

Assumptions of the model:
- ✓ Linear in parameters
- ✓ No perfect collinearity
- ✓ Zero Conditional Mean
- ✓ Homoscedasticity
- ✓ No correlation of the errors
- ✓ Normality

Discussion

United Nations assigns a Happiness Scoring to countries:

Factors related to this "Happiness Score" given:

- Variables such as Economy, Health and Infrastructure do have a considerable higher impact on this Happiness than variables like Suicide, Population Size and Compulsory Education
- Geographical pattern in Happiness around the world

By taking into account the linear combination of a set of variables, we can enhance interpretability:

- Visual inspection of components to derive trends and clusters
- Predictive model proves its efficiency on Happiness

But... In the end, happiness is something you can not measure accurately.

What is Happiness for you?

NOVA IMS – Descriptive Analytics, Spring Semester 2018
Dahmane Sheikh · Daniel Cooper · Maria Lechleitner