Assignment 1

Introdução

No mundo actual existem cada vez mais dados disponíveis sobre o dia a dia de um cidadão normal, quer seja através da utilização dos mais diversos dispositivos como através da utilização de serviços colectivos em grandes centros urbanos. De forma a tirar partido desta enorme quantidade de dados para chegar a conclusões viáveis em tempo útil, é necessário fazer uso dos métodos disponíveis para redução de variáveis e explicação de variância. Com isto é possível aumentar o rendimento das análises dos dados.
 Neste projecto foi proposto a escolha de um *dataset* com determinadas características e a realização de diversos tipos de análises, univariada, bivariada e multivariada, aos dados obtidos.
 Numa primeira parte será exposto uma breve análise dos dados agrupados por continentes, de forma a dar uma visão mais geral ao leitor, e também de contextualizar com a realidade actual. De seguida, cada variável será analisada com o intuito de se perceber como estão distribuídas. Posteriormente, o alvo de estudo será a relação entre essas mesmas variáveis e a sua influência, principalmente no *Score* de cada país.
 Na segunda e última parte serão abordados os métodos de *Factor Analysis* de forma a avaliar o peso de cada variável nas restantes. Os dados serão normalizados de forma a que os consigamos analisar na mesma escala de grandeza e avaliados sobre a sua adequação. Para terminar, uma *Principal Component Analysis* será realizada e conclusões de como algumas variáveis são mais relevantes do que outras.

Análise de dados

Introdução

Para a realização deste trabalho foi escolhido um dataset que traduz o nível de felicidade, bem como outros indicadores, de diversos países, relativo ao ano de 2019. Uma vez que cada entrada correspondia a um país, decidiu-se agrupar os dados pelo continente ao qual os países pertencem.
 Começando pelo *Score*, este é baseado nas respostas de um questionário sobre a avaliação da qualidade de vida da população. Na questão, conhecida como *Escada de Cantril* é pedido que se imagine uma escada com 10 degraus (0 em baixo e 10 no topo). O décimo degrau corresponde à melhor vida que o questionado poderia ter, e o primeiro, a pior.

Na imagem acima podemos observar a média dos *Scores* dos diversos países pelo continente ao qual pertencem. Rapidamente se observa que continentes onde se encontram os países mais desenvolvidos são os que, em média, têm um *Score* mais elevado. Na imagem abaixo podemos observar a distribuição do *Score* pelos diversos países do mundo.

Existem outros factores que estão fortemente relacionados, mas sem impacto, com o *Score*, tais como o *GDP per capita (PIB per capita), Social support (apoios sociais), Healthy life expectancy (esperança média de vida), Freedom to make*

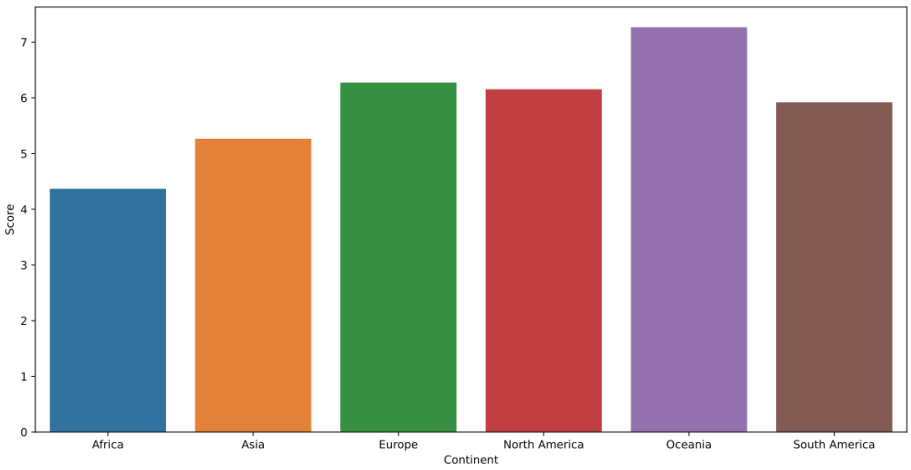


Figure 1: Média de Score por Continente

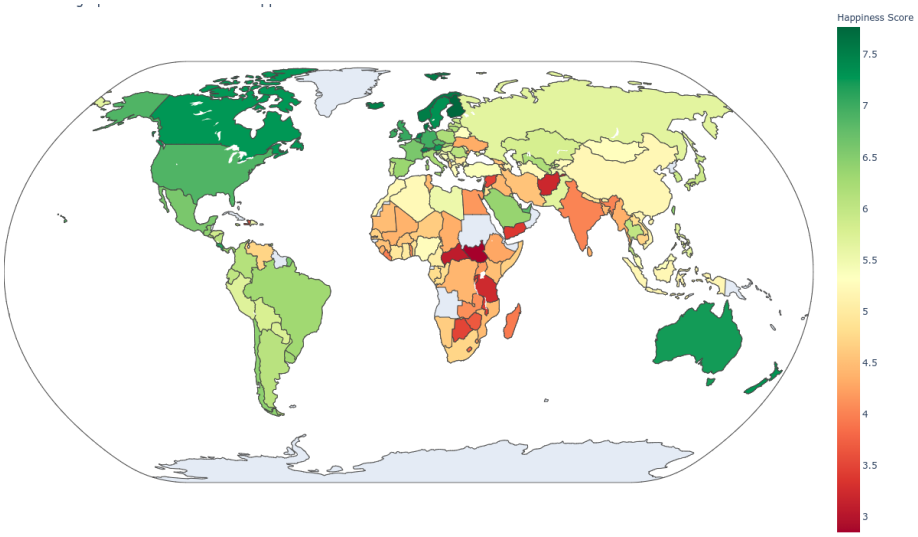


Figure 2: Score dos diversos países

life choices (liberdade), Generosity (generosidade) e Perception of corruption (Corrupção). É importante realçar que todos estes factores são também resultados de questionários, expecto o PIB per capita e a esperança média de vida.
 Uma nota para a variável *Corrupção*. Esta não representa o quão corrupto é um país, mas sim a capacidade da população em detectar/identificar corrupção.

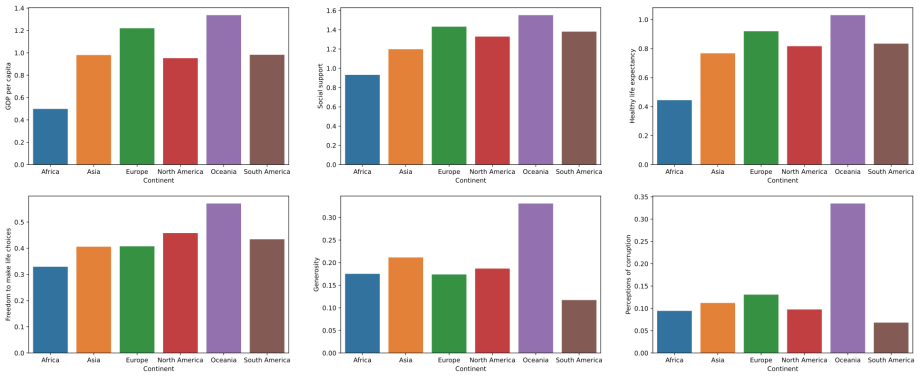


Figure 3: Média dos restantes factores por continente

Como esperado, todos os indicadores seguem o *Score*. Relativamente à generosidade e corrupção, a Oceania apresenta valores de média muito mais elevados que os restantes continentes devido ao facto de a sua amostra é constituída por apenas dois países de elevado índice de desenvolvimento.
 Para concluir, através de uma análise rápida dos dados agrupados por continente, fica retida a ideia que os países considerados desenvolvidos vão obter resultados mais elevados, e que esses mesmo países se encontram, na sua maioria, no hemisfério norte.

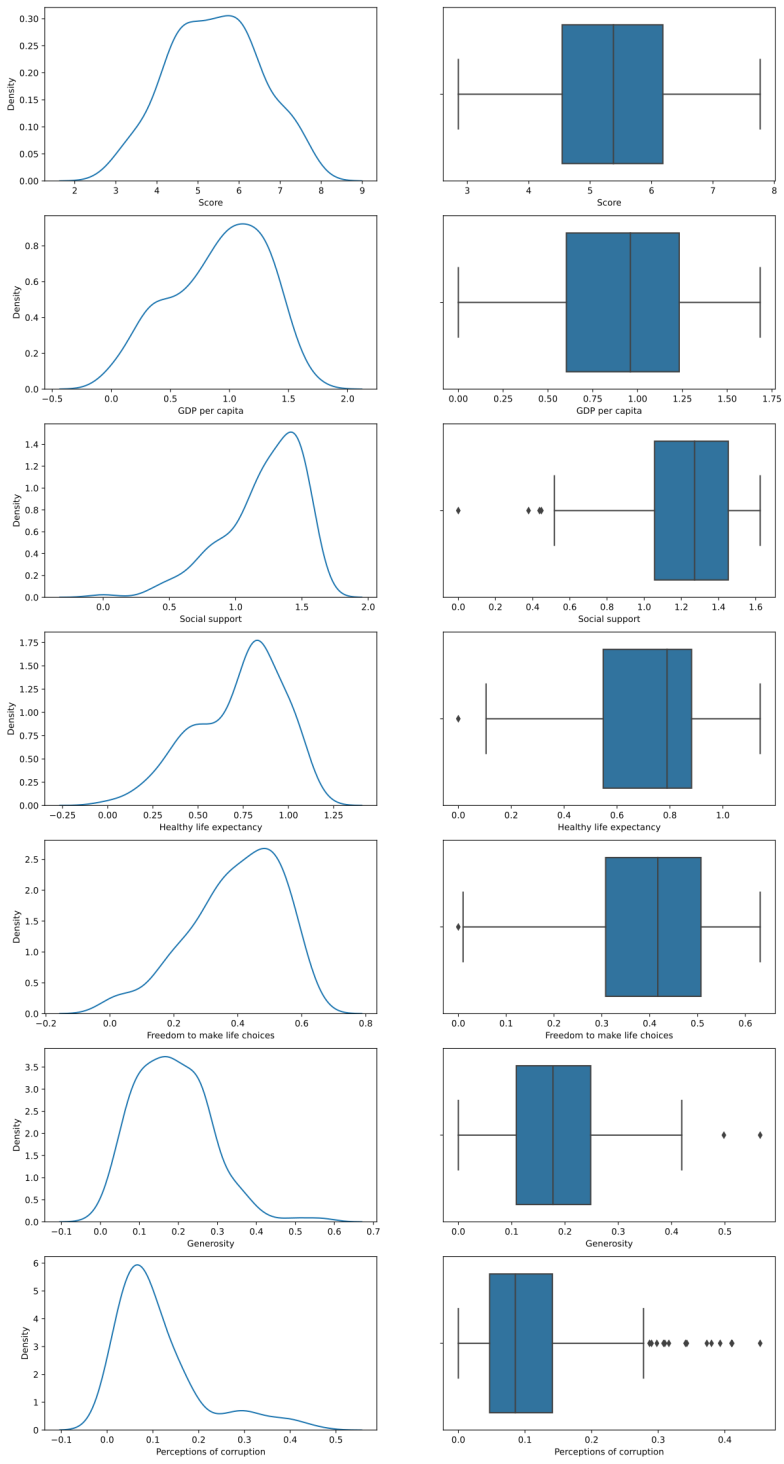
Análise Univariada

Neste capítulo será realizada uma análise estatística a cada uma das variáveis presentes no *dataset*. Analisando a tabela abaixo representada podemos retirar algumas conclusões:

- O número de países representados é de 156;
- A média do *Score* está próximo do meio da escala e o valor máximo é de 7.77 e mínimo de 2.85;
- Metade dos países representados têm um *Score* compreendido entre 4.5 e 6.2;
- Alguns países não têm dados relativos aos factores em causa, pois temos mínimos de 0;
- As respectivas médias e medianas andam próximas, de onde concluímos que a distribuição dos valores será aproximadamente centrada.

Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Gener
count	156	156	156	156	156
mean	5.407	0.905	1.208	0.725	0.392
std	1.113	0.398	0.299	0.242	0.143
min	2.853	0.000	0.000	0.000	0.000
25%	4.544	0.602	1.055	0.547	0.308
50%	5.379	0.960	1.271	0.789	0.417
75%	6.184	1.232	1.452	0.881	0.507
max	7.769	1.684	1.624	1.141	0.631
IQR	1.640	0.629	0.396	0.334	0.199
skew	0.011	-0.385	-1.134	-0.613	-0.685
mad	0.916	0.332	0.236	0.199	0.116
kurt	-0.608	-0.769	1.229	-0.302	-0.068

Analisando o valor de *skewness* concluímos que não temos nenhuma das variáveis totalmente simétricas quanto à sua distribuição. Por outras palavras, a *skewness* traduz a falta de simetria das distribuições. Adicionando a análise de *Kurtosis*, podemos identificar alguns casos em que a presença de *outliers* é bastante provável (*p.e. Perceptions of corruption*).


Density
Score
Score
Density
GDP per capita
GDP per capita
Density
Social support
Social support
Density
Healthy life expectancy
Healthy life expectancy
Density
Freedom to make life choices
Freedom to make life choices
Density
Generosity
Generosity
Density
Perceptions of corruption
Perceptions of corruption

Recorrendo à análise gráfica as conclusões convergem. Todas as distribuições são aproximadamente simétricas, com a presença dos expectáveis *outliers*. Alguns deles correspondem aos valores em falta de alguns países. Contudo, a sua existência é única, e por isso, não significativa.
 Como esperado, na corrupção temos a presença de um número maior de *outliers*, o que vai de encontro ao respectivo valor de *Kurtosis*.
 Concluindo, não existe uma grande disparidade em relação ao *Score* e ao *GDP per capita (PIB per capita)* dentro dos países analisados. Relativamente ao *Social support (apoios sociais)* e à *Healthy life expectancy (esperança média de vida)*, as populações têm uma visão positiva do seu país. O mesmo pode ser dito da *Freedom to make life choices (liberdade)*. Por último, *Generosity (generosidade)* e *Perception of corruption (Corrupção)*, com uma performance menos boa. A corrupção, como foi dito anteriormente, traduz a percepção de corrupção. Logo podemos concluir que a presença de *outliers* corresponde às populações dos países mais desenvolvidos, possivelmente devido ao maior nível educacional das mesmas.

Análise Bivariada

Neste capítulo será analisada a relação entre as variáveis. Os casos mais pertinentes serão quase sempre relativos à relação com o *Score*. No entanto, a análise de outras relações pode ser útil de forma a clarificar, indirectamente, alguns tópicos.

Analisando a linha do *Score* vemos que existe uma correlação positiva com quase todas as variáveis. Contudo, a generosidade practimente não têm influência, e a corrupção apenas se consegue detectar a presença de correlação em valores elevados.
 Algumas correlações positivas esperadas comprovam-se, tal como *PIB per capita* e *Apoio social*, *PIB per capita* e *esperança média de vida*, e por último *Apoio social* e *esperança média de vida*. Nota para o facto de não se conseguir identificar nenhuma correlação negativa.
 Antes de passar para a análise do mapa de correlações, notar que existe uma visível correlação positiva entre a *percepção de corrupção* e *liberdade*, o que de certa forma é expectável, tendo em conta que os países no mundo com maior liberdade são também os que apresentam maior transparência nos seus processos governamentais.

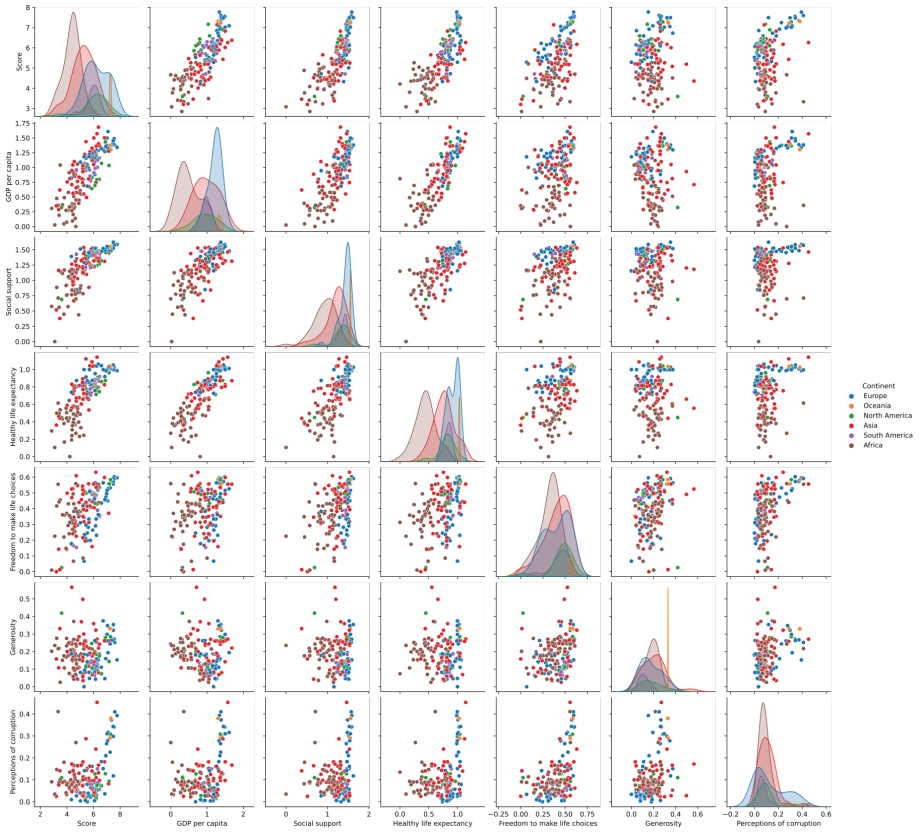
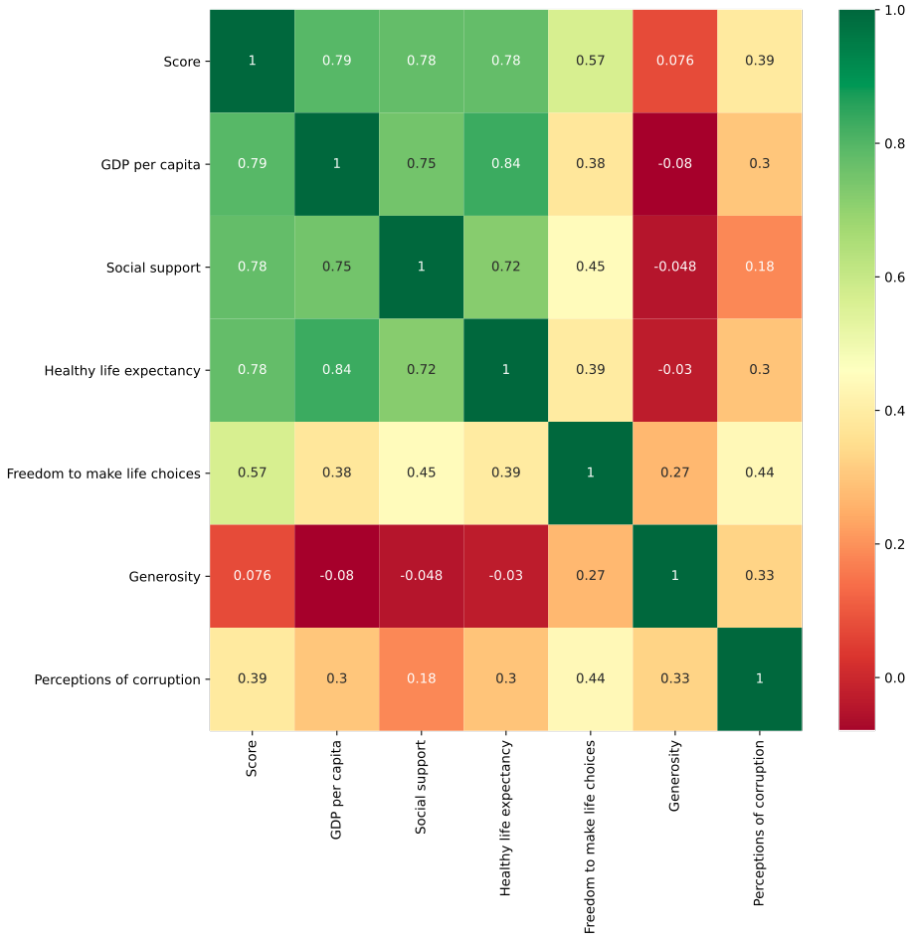

Score
GDP per capita
Social support
Healthy life expectancy
Freedom to make life choices
Generosity
Perceptions of corruption
Continent
Europe
Oceania
North America
Asia
South America
Africa

Figure 4:



Analisando o mapa de correlações vemos que vai de encontro às conclusões obtidas anteriormente. Fortes correlações positivas entre as quatro primeiras variáveis, a muito reduzida presença de correlações negativas (e com valores muito próximos de zero) e a correlação positiva, embora não relevante, entre a *percepção de corrupção* e *liberdade*.

Factor Analysis

Normalização

Uma vez que as variáveis têm diferentes escalas, por exemplo o *Score* varia entre 0 e 10, ao passo que a perceção de corrupção encontra-se entre 0 e 1 (uma vez que se trata de uma média que avalia as resposta a uma pergunta com a possibilidade de responder sim (1) ou não (0)) procedeu-se à normalização dos valores (com remoção da média e variância unitária) em todas as variáveis com exceção da *Country or Region*, visto se tratar de uma variável categórica nominal.

Testes de adequação

Inicialmente foi aplicado o teste de *Bartlett* sobre os dados normalizados, sendo a hipótese nula a matriz de correlações trata-se de uma matriz de identidade, o que indicaria que as variáveis seriam não correlacionadas e portanto, não adequadas para *factor analysis* (FA) [1]. Obteve-se um valor de chi-quadrado de aproximadamente 656 e o nível de significância obtido foi de 0.0 ($5 * 10^{-126}$) indicando a rejeição da hipótese e, portanto, que os dados são adequados ao tipo de análise em questão.
 Foi, também, realizado o teste de adequação *Kaiser-Meyer-Olkin's* (KMO), obtendo-se o valor de aproximadamente 0.84 o que evidencia uma forte adequação à realização de FA. A tabela seguinte exibe os valores de *Measure of Sampling Adequacy* (MSA) para as variáveis em análise. Todos as variáveis possuem um MSA superior a 0.5, tendo sido portanto mantidas para análise [2].

Score	*GDP*	*Social Support*	*Healthy life exp.*	*Freedom*	*Generosity*	*Corruption*
0.855	0.827	0.871	0.862	0.829	0.596	0.752

PCA e análise

Procedeu-se à aplicação da PCA sobre os dados normalizados, tendo-se obtidos os *eigenvalues* apresentados na tabela seguinte:

	Eigenvalue	Fracção de Variância Explicada (%)	Fracção acumulada (%)
1	3.837141	54.46	54.46
2	1.436346	20.39	74.85
3	0.616839	8.76	83.61
4	0.559896	7.95	91.56
5	0.263794	3.74	95.30
6	0.173418	2.46	97.76
7	0.157726	2.24	100.00

Na imagem ilustrada podemos observar a representação gráfica da anterior tabela:

Verifica-se que a primeira e segundas componentes explicam aproximadamente 54% e 20 % da variância, totalizando 74.85% da variância total, sendo que para as componentes seguintes existe uma queda brusca relativamente a estas. Aplicando o critério de Kaiser (selecionar apenas as componentes a que corresponde um

[1] IBM, «IBM Docs», Out. 24, 2014. www.ibm.com/docs/en/spss-statistics/23.0.0 (acedido Abr. 13, 2021)

[2] IBM, «Kaiser-Meyer-Olkin measure for identity correlation matrix», Abr. 16, 2020. https://www.ibm.com/support/pages/kaiser-meyer-olkin-measure-identity-correlation-matrix (acedido Abr. 13, 2021).

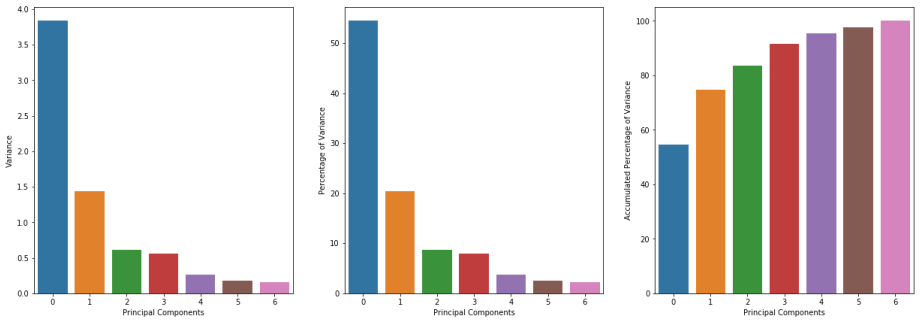


Figure 5:

eigenvalue superior a 1) foram selecionadas as primeiras duas componentes (PC1 e PC2).
 Na tabela seguinte apresentam-se os *loadings* das primeiras 2 componentes, que representam as correlações entre as componentes e as variáveis, valores (absolutos) mais elevados de correlação encontram-se realçados. Estes valores refletem a importância de cada variável nas componentes.

Features	PC1	PC2
Score	`-0.475861`	-0.028371
GDP	`-0.454825`	-0.213377
Healthy life exp	`-0.436582`	-0.207148
Social support	`-0.450150`	-0.177856
Freedom	-0.332201	0.362130
Generosity	-0.048232	`0.693809`
Corruption	-0.246511	`0.516346`

Seguidamente podemos observar os dados dos *loadings* no círculo de correlação. Como se pode visualizar a PC1 está mais correlacionada com as variáveis *Score, GDP, Social support and Healthy life exp.*, ao passo que a PC2 está mais ligada às 3 restantes variáveis.
 Foi aplicada FA com duas componentes tendo sido utilizada uma rotação (de forma a melhorar a interpretação) do tipo *varimax* e o método de eixo principal para realizar a extração, tendo-se obtido as *loadings, communalities* e variância específica representadas na seguinte tabela:

	Factor 1	Factor 2	*Communalities*	Variância específica
Score	0.886962	0.278875	0.864474	0.14
GDP	0.922187	0.056855	0.853661	0.15
Healthy life exp.	0.886130	0.051952	0.787925	0.21
Social support	0.899390	0.093791	0.817701	0.18
Freedom	0.466562	0.624670	0.607894	0.39
Generosity	-0.188509	0.812598	0.695852	0.30
Corruption	0.247258	0.742319	0.612174	0.39

Pela análise da tabela podemos concluir que o Factor 1 está fortemente relacionado com as primeiras 4 variáveis ao passo que o Factor 2 encontra-se mais correlacionado com as restantes variáveis. Em termos de *communalities* verifica-se que uma fração significativa das variáveis é explicada pelo factores presentes. Na seguinte está representado o círculo de correlação das *loadings* obtidas.

Na próxima figura encontra-se representado o gráfico dos indivíduos (países) quando a eles é aplicada a transformação imposta pelo modelo determinado

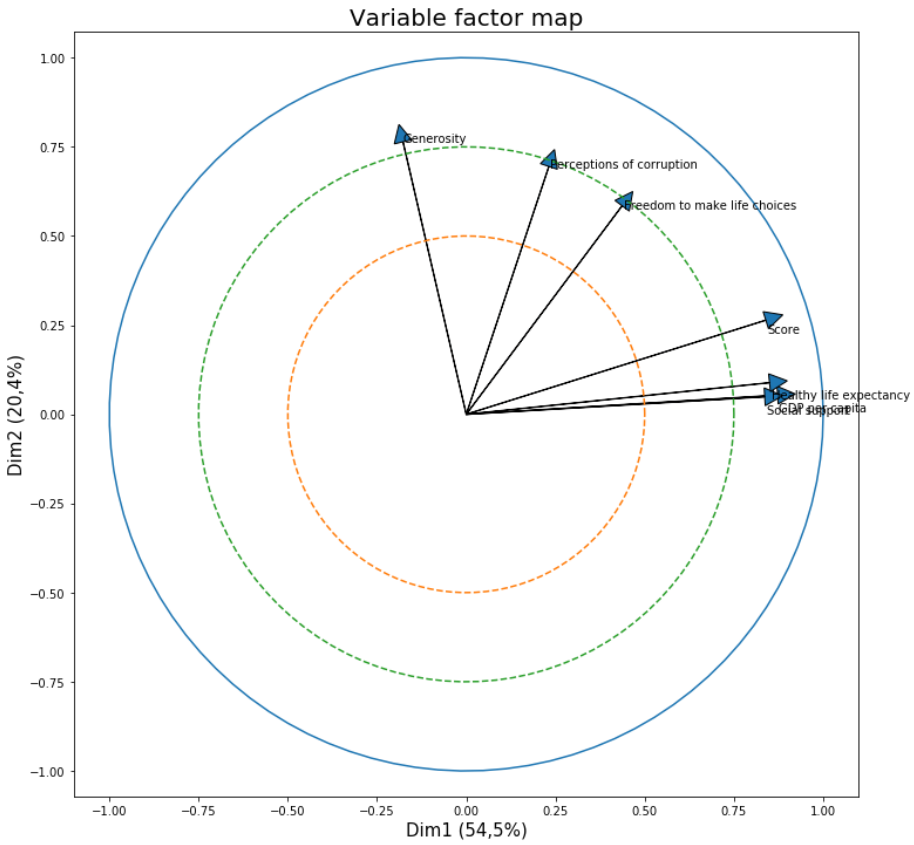

Variable factor map
Generosity
Perceptions of corruption
Freedom to make life choices
Score
Healthy life expectancy
GDP per capita
Social support
Dim1 (54,5%)
Dim2 (20,4%)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Figure 6:

(uma vez que a correlação destas variáveis é mais forte com a componente 1. No caso dos países africanos verifica-se que maioritariamente ocupam as posições do gráfico mais à esquerda (indicando maus parâmetros em termos de *score*, esperança média de vida saudável, GDP e suporte social).

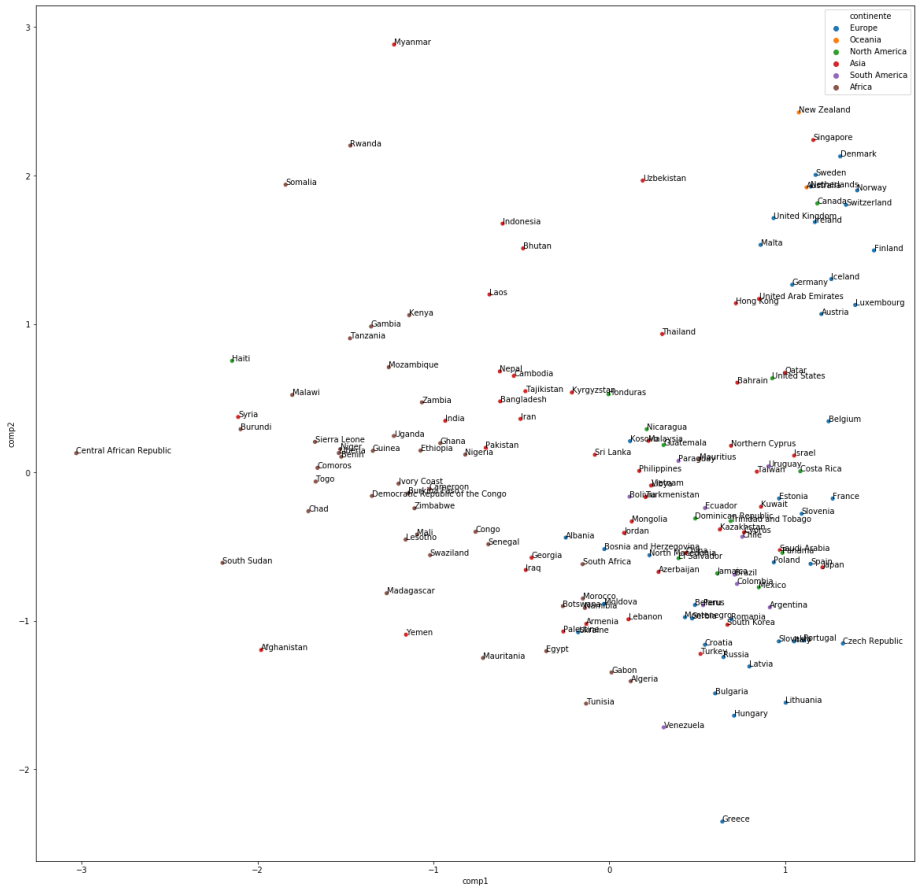


Figure 7:

Conclusão

Com a realização deste trabalho é possível retirar a conclusão de que os métodos estatísticos abordados são de uma enorme importância para a escolha de variáveis. Embora o *dataset* escolhido não fosse muito extenso, foi possível analisar a importância de cada uma das variáveis e o peso das mesmas.
 Encontraram-se algumas dificuldades, nomeadamente na procura de suporte para alguns métodos na linguagem *Python*, pois ainda não é tão abrangente neste capítulo como o *R*.
 Relativamente aos dados, podemos concluir que existem variáveis mais importantes para que a resposta da população sobre o

índice de felicidade seja mais elevada, tais como o *GDP* ou o *Social support*. Ambas seriam provavelmente dedutíveis sem esta análise. Contudo, outras variáveis que geralmente consideraríamos relevantes, concluiu-se que não o são.

Referências