世界各国幸福水平差异分析及影响因素探究

吕康乐 2021012572

2023 年 6 月 25 日

摘要：

今年十四届全国人大一次会议上曾强调“必须以满足人民日益增长的美好生活需要为出发点和落脚点”，究竟什么因素会更增加人们的幸福感？本文基于多元统计分析方法，对2015-2022 年世界上 150 多个国家的幸福相关指标进行分析。利用因子分析对影响幸福的指标进行降维和解读，利用回归分析对幸福指数进行预测，利用聚类方法对不同国家进行聚类。通过本文的研究，可以增进对幸福的理解，对个人选择和国家发展都有一定的意义。

目录

一、研究问题与背景描述

“幸福是我们一切行为的终极目标，我们为此所做的所有事情其实都是手段。”从古至今，每个人都在用各种方式追寻幸福、守护幸福，工作收入、生活压力、身体健康、社会地位等众多因素都影响着一个人的幸福程度。究竟什么因素会更增加人们的幸福感？是亲情、自由、金钱、社会地位……？联合国也一直关注于幸福问题，从 2012 年起开始发布《全球幸福指数报告》，每年一期，在世界范围内得到了政府、机构组织、社会团体等的认可。报告基于人均国内生产总值（GDP）、健康预期寿命、生活水平、国民内心幸福感、人生抉择自由、社会清廉程度以及慷慨程度等多方面因素进行研究并得出结果。

那么，这些因素可以概括为哪几个方面，哪个更加重要？近年来我们的幸福指数是在逐年提高吗？哪些国家或地区整体幸福感排名较高？幸福水平高的国家在地理位置、经济发展、政治体制等方面有哪些共同特点？

二、数据介绍

2.1 数据集描述

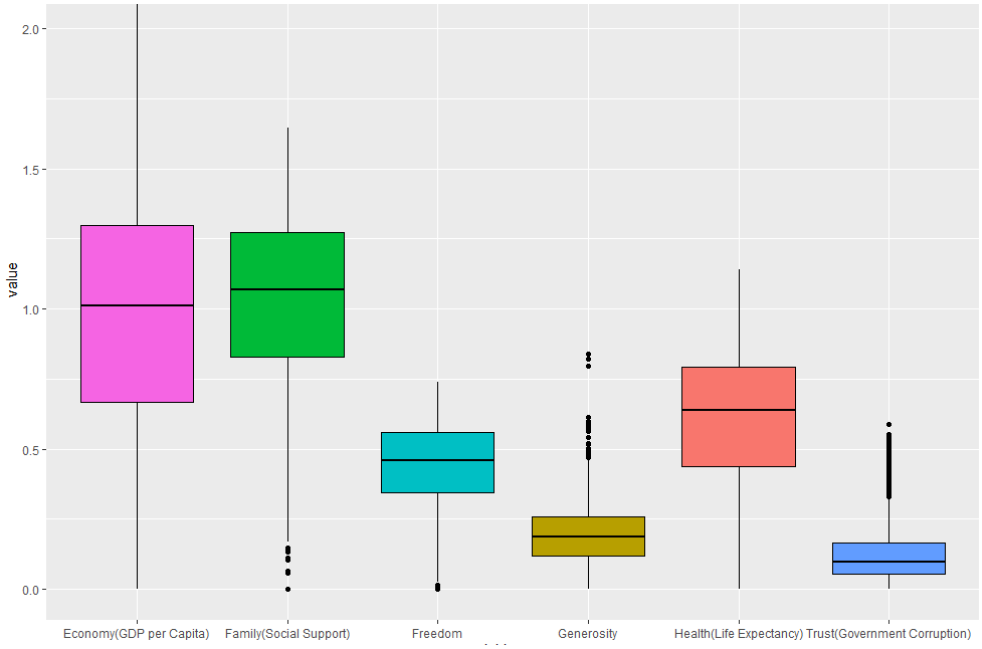
本研究所使用的数据集来源于课堂提供的 happiness 数据集，也可在网站 https://www.kaggle.com/datasets/shivkumarganesh/world-happiness-report-20152022 上找到。数据集为 2015-2022 历年 150 多个国家的幸福程度和相关指标得分，第一列为幸福指数排名，2、3 两列为国家和地区，第 4 列为幸福指数，5~10 列为影响幸福的 6 个因素得分，包括经济生产、社会支持、预期寿命、自由、没有腐败和慷慨程度，最后一列是年份。其中第 4 列的幸福得分来自盖洛普世界民意调查，要求民众在[0，10]区间打分，0 分代表最差，10 分代表最好；5~10 列数据来源于联合国对各个方面的评估。

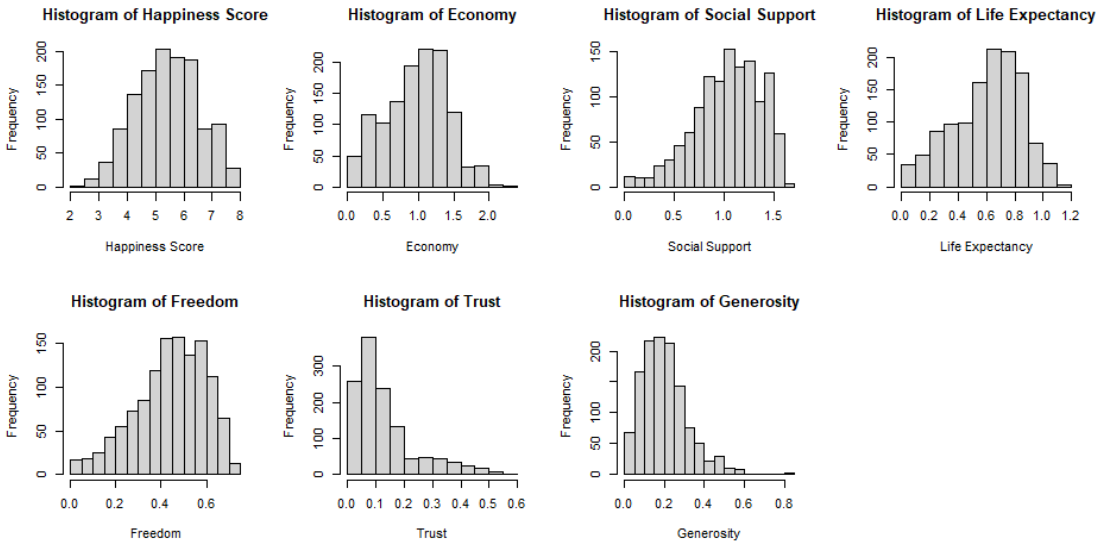
2.2 单变量描述

首先我们关注各个数值变量的基本分布情况，如下表和下图所示：

名称	含义	最小值	中位数	平均值	最大值
Happiness Score	幸福指数	2.404	5.410	5.429	7.842
Economy(GDP per Capita)	经济生产	0.000	1.012	0.975	2.209
Family(Social Support)	社会支持	0.000	1.069	1.033	1.644
Health(Life Expectancy)	预期寿命	0.000	0.639	0.608	1.141
Freedom	自由	0.000	0.459	0.441	0.740

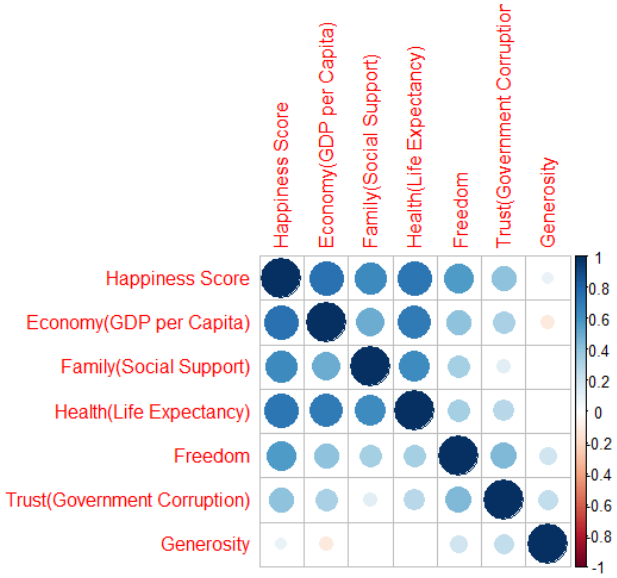
Trust(Government Corruption)	无腐败	0.000	0.096	0.131	0.587
Generosity	慷慨	0.000	0.187	0.202	0.838
Year	年份	2015	2018	2018	2022





可以看到，影响幸福的六个因素得分的数值大小基本一致，而幸福指数要比这些因素得分大很多。从箱型图中看出部分变量有一些离群值，但考虑到各国之间差异可能是巨大的，我们暂不考虑离群值的情况，即认为所有数据是正常的。此外，这 6 个因素的方差也各有不同，因此后续分析要尽量减少依赖于同方差假设的内容。从直方图中，幸福指数、经济得分和预期寿命近似于正态分布，而其余变量略有一些左偏或右偏。

2.3 相关性描述



画出各数值变量间的相关性图示，幸福指数与除了慷慨之外的五个变量都有一定的正线性相关关系，其中和经济、寿命两个因素得分的相关程度最高，这也与我们的直观认知较为吻合。各个变量之间也有相关性，经济、社会支持、寿命三者两两的相关程度都较高，而另外三个因素信任、慷慨和自由基本互不相关，和其他变量的相关性也都较弱。这对后续因子分析降维的结果提供了参考。

三、方法

在后续研究中，主要采用如下方法：

1. 首先进行因子分析，思考影响幸福水平的 6 个因素能否进一步总结；

2. 接下来进行回归分析，得到幸福指数和 6 个因素之间的回归表达式，探究各个变量的显著性；同时以年份作为因子，看幸福水平是否在随时间变化；

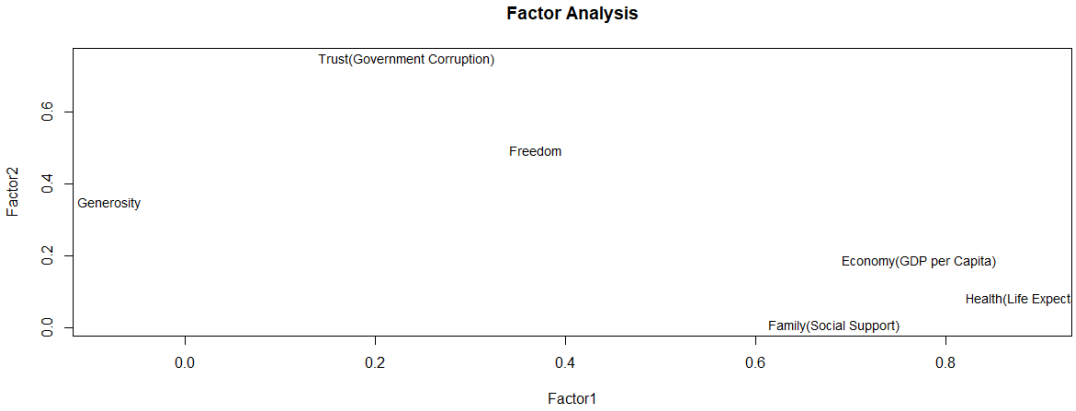
3. 最后进行聚类分析，得到哪些国家可以归为一类，通过探索同一类的国家有什么共同特征，可以进一步增进对幸福的理解。

四、结果

4.1 因子分析：寻找影响幸福水平的潜在因子

我们首先考虑求得可解释这 6 个影响幸福水平因素的潜在因子，基于数据的相关系数矩阵对数据进行因子分析建模。我们先尝试提取 2 个公共因子，将主成分法和极大似然法的结果进行对比，可以发现因子含义与得分基本是一致的，说明 2 个公共因子较为合适。而若

提取 3 个因子，会发现两种方法的结果具有较大差别，不具有可解释性，因此 2 个公共因子是较为理想的结果。而对比 2 个因子下的主成分法和极大似然法，可以发现主成分法每个变量的特殊方差基本都比极大似然法大较多，这说明主成分法可能由于假设较强而低估了特殊方差，不如极大似然法更可靠，因此下面展示的均为极大似然法结果。为了获得更好的可解释性，采用 varimax 方法对因子进行旋转。将各变量因子上的载荷可视化，如下图所示：



变量 Economy(GDP per Capita)、Family(Social Support)和 Health(Life Expectancy)在因子 1 上载荷较为显著，变量 Freedom、Trust(Government Corruption)和 Generosity 在因子 2 上载荷较为显著。因此，我们可以把因子 1 解释为基础生活水平，更多与物质、金钱相关，反映了人们基本的生存条件与社会福利政策；因子 2 可以解释为自我实现水平，更多的是思想上、精神上的幸福感，包括自由选择、慈善行为和社会信任等，反映了人们的自我实现需求。因子具体的载荷和方差的解释比例如下所示：(空缺代表数值很接近于 0)

变量	Factor1	Factor2	共同方差	独特方差
经济生产	0.772	0.185	0.730	0.370
社会支持	0.683		0.466	0.534
预期寿命	0.893		0.805	0.195
自由	0.369	0.492	0.379	0.621
政府信任、无腐败	0.233	0.746	0.610	0.390
慷慨		0.347	0.127	0.873
SS loadings	2.057	0.960		
Proportion Var	0.343	0.160		
Cumulative Var	0.343	0.503		

因子 1 解释了占比 0.343 的方差，高于因子 2 的解释比例 0.160，因此我们可以认为因子 1 在解释方差上比因子 2 更重要。这表明物质上的充裕相对精神上的富足更加重要，物质生活得到基本保障是获得幸福感的基础。当然，只有物质和金钱也是不行的，在此基础上需要再提高精神上的获得感。

4.2 回归分析：探究各个变量对幸福的具体影响

利用因子分析，我们得到了对幸福水平影响因素的深层解读。这里我们建立线性回归的模型，探究各个变量对幸福指数的具体影响。通过多重线性回归，得到的最优模型包括了全部 6 个变量：

$$Happiness\ Score = 2.146 + 0.904Economy + 0.738Family + 1.151Health + 1.567Freedom + 0.821Trust + 0.699Generosity$$

此时 6 个变量全部非常显著（p 值均远小于 0.001）。该包含所有变量的模型拥有最大的 $R^2 = 0.7428$，也可以验证其有最大的$adjusted\ R^2 = 0.7415$，从多个角度来看都是最优的。$R^2 = 0.7428$说明线性回归式可以解释 74.28%的方差，占比较高，说明拟合效果较好。从残差分析结果，可见模型假设都满足。上式一方面验证了六个因素都会影响人们的幸福感，另一方面对一个国家的幸福指数提供了预测方法。

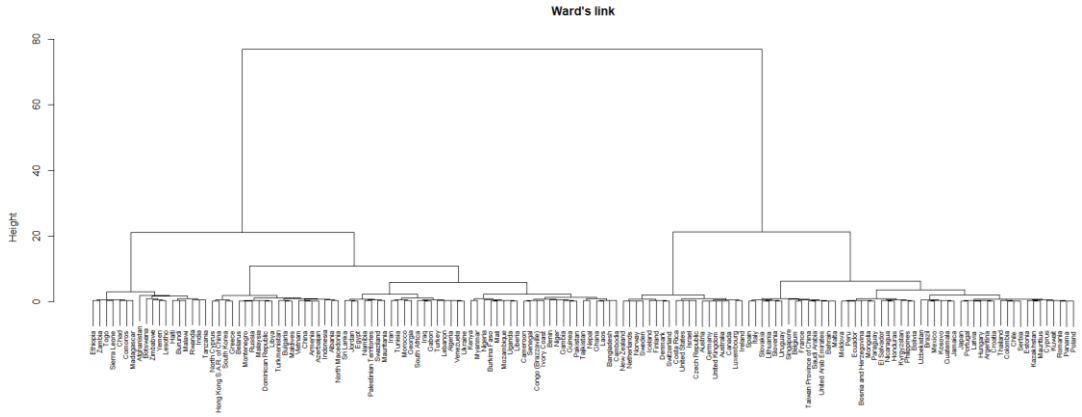
此外，我们还关注于幸福指数是否在变化，人们是否越来越幸福。如果以年份作为因子进行方差分析，p 值远小于 0.001，说明每年的幸福指数是不同的。具体的，全球历年的均值如下表所示：

2015	2016	2017	2018	2019	2020	2021	2022
5.376	5.382	5.354	5.367	5.407	5.473	5.533	5.554

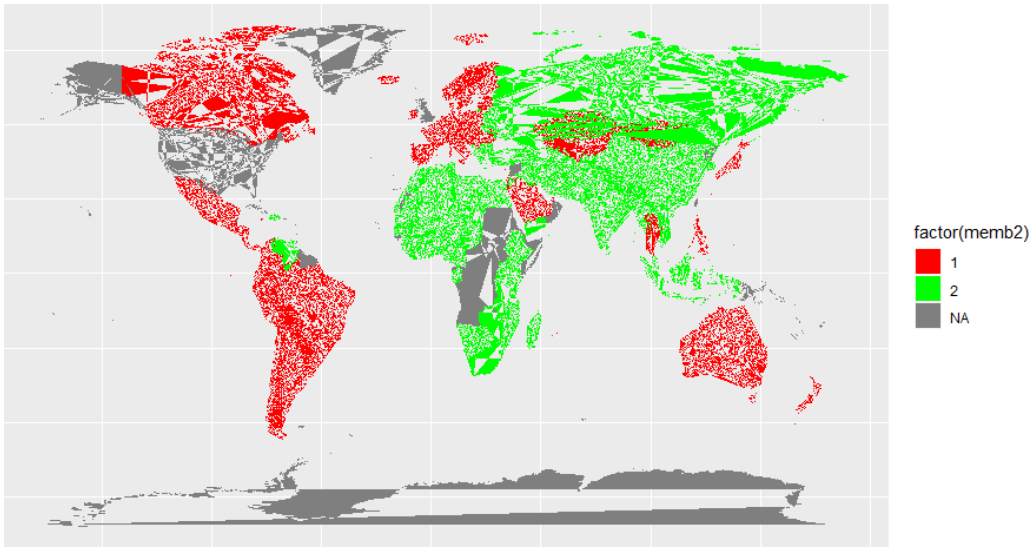
因此全球平均幸福水平总体上先下降后上升，近五年都是在上升的。重点关注中国的变化，是处于升—降—升的的变化规律，19 和 20 年下降，21 年又回升。在具体的 6 个影响因素上，社会支持保障和预期寿命两方面的得分发展缓慢，从 2020 年起下降较多，2022 年甚至与 2015 年持平或不如 2015，这可能部分归因于新冠疫情的冲击。此外，注意到中国在 150 多个国家中一直在 70-90 名之间徘徊，且社会保障和慷慨程度两方面仍低于全球均值较多，说明我们在提高国民幸福感方面还有很多需要努力。

4.3 聚类分析：获得各国幸福水平的类别信息

在聚类分析部分，由于每年的整体排名近似，且多年数据堆在一起不便聚类，因此我们只对 2021 年数据进行聚类（2022 有较多 0 值，不便进行）。以下主要尝试了两种办法：层次聚类法和 k-中心点聚类法。在层次聚类法中，采用的是欧式距离、Ward’s linkage，对含幸福指数在内的 7 种连续变量进行层次聚类，得到树状图如下：



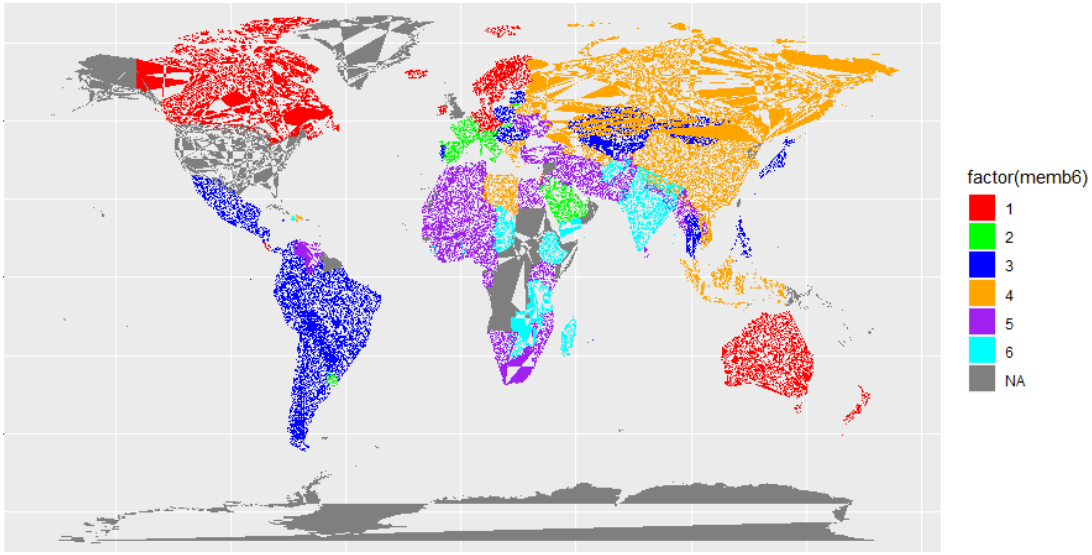
直观看上去，首先可以把所有国家分成两大类；为了使聚类结果能够更好地展示出来，我将这些国家在地图上标记，依据类别信息给不同国家标注对应的颜色，如下图所示：



结合上图和这些国家对应的幸福感排名，可以认为这种聚类的最直接含义是分为幸福度较高（红色表示）和幸福度较低（绿色表示）的两部分。同一类国家更多地趋向于位于同一地区，幸福感较高的国家主要分布在北美洲、南美洲、大洋洲和欧洲，另一类则集中分布于亚洲和非洲。大部分发达国家（如欧洲、北美洲、大洋洲各国、日本）幸福指数高，发展中国家幸福指数整体低一些，这印证了经济变量对幸福指数的重要意义。南美洲虽然有较多发展中国家，但其生活大部分都较为幸福，这可能与政府社会保障较好、社会和谐等因素有关。

使用 k-中心点聚类法、取 k=2 得到的结果与此较为接近，仅小部分国家不同，说明这样分类有一定的合理性，我们后续就只利用层次聚类法并展示相应结果。

依据树状图，考虑将各国聚为 6 类，结果如下图所示：



聚类结果仍然是有很强的地域特征，且基本还是依据幸福水平划分的。北美洲、大洋洲和北欧各国的居民生活幸福感最高，南欧国家其次，接着是南美洲各国及少许亚洲国家，排在第四梯队的是亚洲大部分国家，最后基本都是非洲国家。这样的结果也符合我们的基本认知。

五、讨论与改进

本研究通过因子分析、回归分析和聚类分析，探究影响幸福的主要因素，并分析了世界各国幸福水平差异，对个人选择和国家发展都有一定的意义。

1. 根据因子分析，我们把影响幸福的主要因素划分为两类，一类是物质财富、基本生活保障因素，另一类是自由、信任等精神层面的因素，且物质因子比精神因子更加重要。

2. 根据回归分析，我们得到了幸福指数关于 6 个因素的线性回归表达式，这 6 个变量对于幸福感都有重要的意义，依据回归方程可以预测一个国家的幸福水平。关注于年份对幸福的影响，可以发现全球和中国的幸福感总体上是上升态势，中国的幸福感还落后于较多国家，主要是体现在社会保障和慷慨程度两方面。

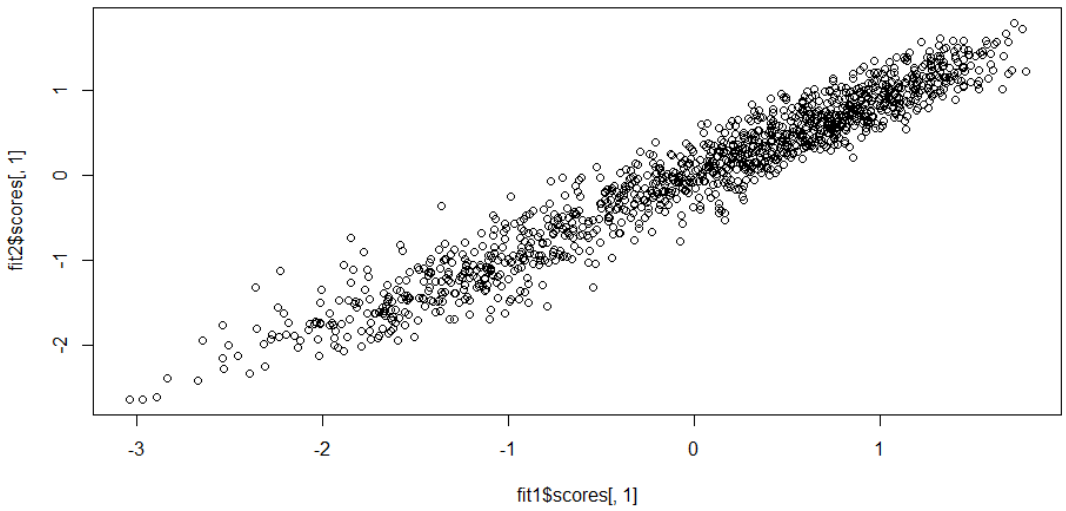
3. 根据聚类分析，全球幸福感的排名表现出地域特征，基本按照北美洲大洋洲和北欧、南欧、南美洲、亚洲、非洲幸福指数从高到低。发达国家往往幸福感高于发展中国家，南美洲虽不够发达但仍有较高的幸福程度。

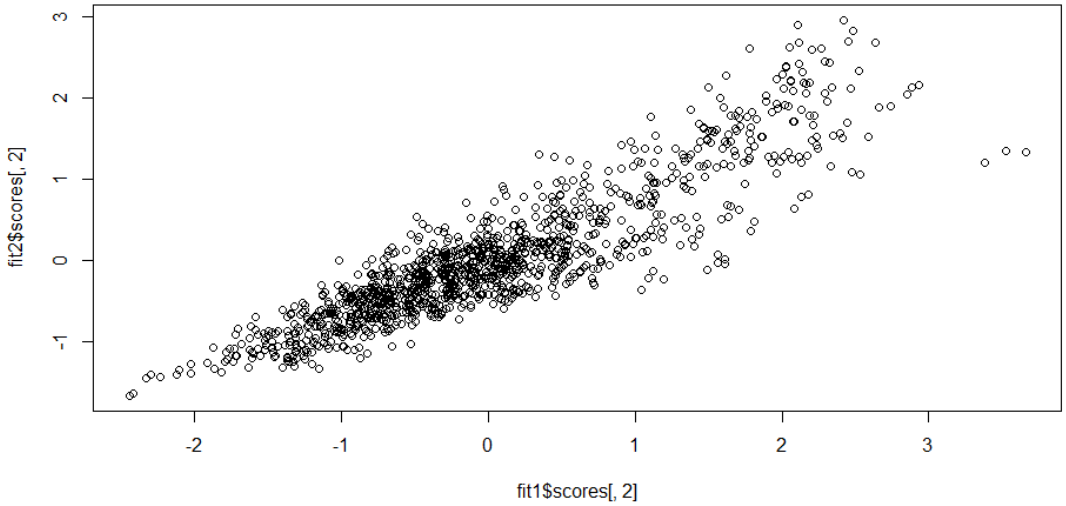
本研究的一个不足之处在于没能很好的可视化。尤其是在聚类分析部分，高维数据难以被低维较好地可视化出来，如果仅用两个变量作为横纵坐标难以看出分类的效果。由于时间紧张，可以在未来尝试 t-SNE、MDS 多维标度法等较好的降维可视化方式。此外，最后部分进行世界地图标注，由于对 R 语言还不熟悉，图片还不够完美，有些地方出现空白缺失，仅能满足看出大体情况。

此外，另一个问题是数据分析时没有考虑异常值问题，这可能与前面的箱型图有所违背。研究也没能充分考虑数据的分布。如直方图所示，部分变量并非正态分布，但在因子分析和回归中最好需要正态假设。鉴于回归的 4 个诊断图并没有明显问题，因此我直接忽略了这点。如果能考虑到数据的分布，并做一些变换，文章会更加严谨。

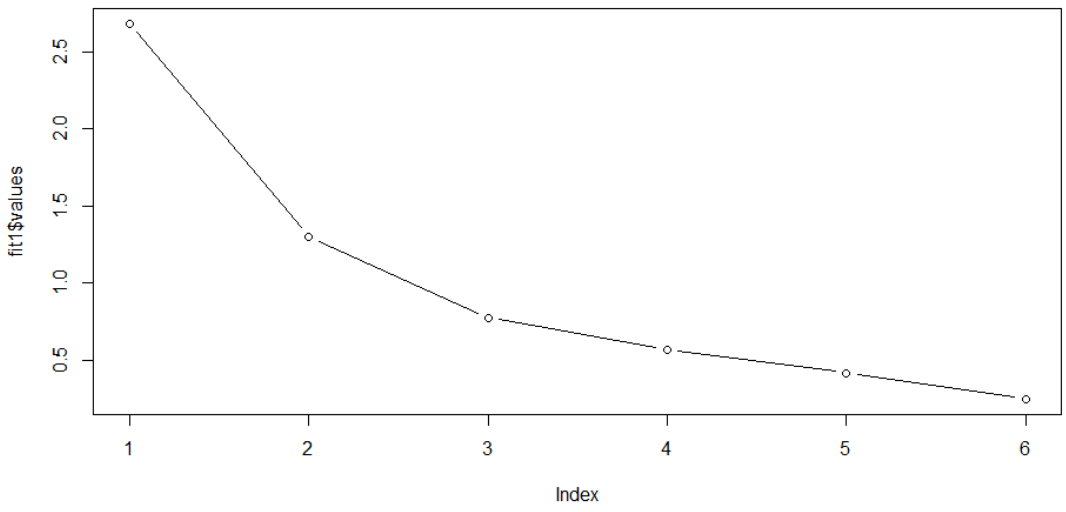
六、附录

因子分析检验取 2 个因子是否合适：

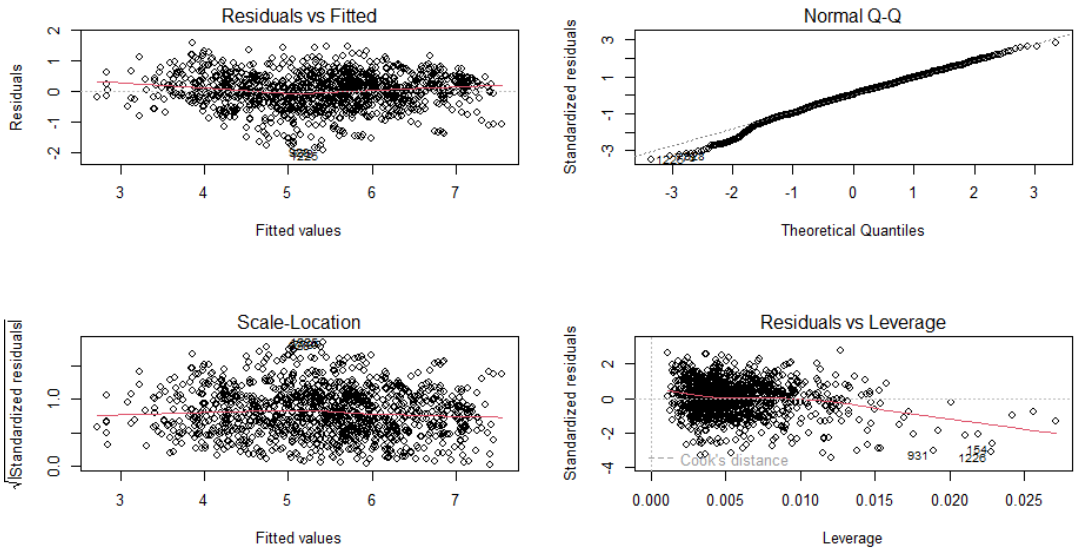




因子分析各因子解释方差图：（也印证了选取 2 个因子是合适的）



回归中的 4 张诊断图：



R 代码：

```
# 导入必要的库

library(dplyr)

library(tidyr)

library(ggplot2)

library(corrplot)

library(psych)

library(cluster)

library(maps)

library(mapdata)


a = file.choose() # choose 'happiness.csv'

a1 = read.csv(a)

a1 = a1[,-1]

## 查看数据的基本信息

str(a1)

## 数据清洗

colnames(a1) <- c("Happiness Rank", "Country", "Region", "Happiness Score", "Economy(GDP per Capita)", "Family(Social Support)", "Health(Life Expectancy)", "Freedom", "Trust(Government Corruption)", "Generosity", "Year")
```

```
a1[,4:10] <- lapply(a1[,4:10], function(x) as.numeric(gsub(",", ".", x)))
str(a1)
summary(a1)
## 计算每列的缺失值数量
missing_values <- sapply(a1, function(x) sum(is.na(x)))
print(missing_values)

# EDA
## 计算所有数值变量的相关性
correlation_matrix <- cor(a1[,-1] %>% select_if(is.numeric))
correlation_matrix <- round(correlation_matrix, 3)
print(correlation_matrix)
corrplot(correlation_matrix, method = "circle")
## 每个变量的箱型图和直方图
a2 = a1[5:10]
color_variable <- c("red", "green", "blue", "orange", "purple", "cyan")
long_data <- gather(a2, key = "variable_name", value = "value")
long_data$color_variable = rep(color_variable, each = 1229)
p <- ggplot(long_data, aes(x = variable_name, y = value, fill = color_variable))
p + geom_boxplot(color = "black")
par(mfrow=c(2,4))
hist(a1$`Happiness Score`, xlab="Happiness Score", main="Histogram of Happiness Score")
hist(a1$`Economy(GDP per Capita)`, xlab="Economy", main="Histogram of Economy")
hist(a1$`Family(Social Support)`, xlab="Social Support", main="Histogram of Social Support")
hist(a1$`Health(Life Expectancy)`, xlab="Life Expectancy", main="Histogram of Life Expectancy")
hist(a1$Freedom, xlab="Freedom", main="Histogram of Freedom")
hist(a1$`Trust(Government Corruption)`, xlab="Trust", main="Histogram of Trust")
hist(a1$Generosity, xlab="Generosity", main="Histogram of Generosity")
par(mfrow=c(1,1))

# FA
## PC method
fit1 <- principal(a2, nfactors=2, rotate="varimax", scores=T, method='regression')
```

```
fit1 # print results
plot(fit1$values,type="b") # scree plot
plot(fit1$loadings)
plot(fit1$loadings,type="n") # set up plot
text(fit1$loadings,labels=names(a2),cex=0.9) # add variable names
## MLE method
fit2 <- factanal(a2, factors=2, rotation="varimax", scores='regression')
fit2
plot(fit2$loadings)
plot(fit2$loadings,type="n",main = "Factor Analysis") # set up plot
text(fit2$loadings,labels=names(a2),cex=0.9) # add variable names
## use factor scores for checking
plot(fit1$scores[,1],fit2$scores[,1]) # help check m
plot(fit1$scores[,2],fit2$scores[,2])
### when m = 3, we can get that the meaning of factor scores isn't clear.
### Hence 2 factors is good.

# regression
reg = lm(a1$`Happiness Score`~., a2)
summary(reg)
par(mfrow=c(2,2))
plot(reg)
par(mfrow=c(1,1))

# take year as factor
a1$Year = as.factor(a1$Year)
reg.year = lm(a1$`Happiness Score`~0+a1$Year)
summary(reg.year)
anova(reg.year)
tapply(a1$`Happiness Score`, a1$Year, mean)
a1[a1$Country == "China", -3]

# clustering using data 2021
a3 = a1[a1$Year == 2021, c(2, 4:10)]
```

```
head(a3)
## hierachical
res1 <- hclust(dist(a3[, -1]), method = "ward.D")
plot(res1, labels=a3$Country, main = 'Ward\'s link', cex=0.7)
memb2 <- cutree(res1, k=2) # 2 classes
memb2
plot(fit2$scores[,1],fit2$scores[,2],col=memb2,xlab="factor score 1",ylab="factor score 2")
## k-medoids
res2 <- pam(a3[, -1], k=2)
plot(fit2$scores[,1],fit2$scores[,2],col=res2$clustering,xlab="factor score 1",ylab="factor score 2")
## compare
memb2
res2$clustering
a3$Country[which(res2$clustering==1)]
a3$Country[which(res2$clustering==2)]
## draw map
world <- map_data("world")
df = cbind(a3, data.frame(memb2))
world_map <- merge(world, df, by.x = "region", by.y = "Country", all.x = TRUE) #  合并数据框
world_map <- ggplot(world_map, aes(x = long, y = lat, group = group)) +
   geom_polygon(aes(fill = factor(memb2)))
world_map + scale_fill_manual(values = c("#FF0000", "#00FF00", "#0000FF", "#FFFF00")) #  添加颜色
## 6 classes
memb6 <- cutree(res1, k=6)
memb6
df1 = cbind(a3, data.frame(memb6))
world_map1 <- merge(world, df1, by.x = "region", by.y = "Country", all.x = TRUE) #  合并数据框
world_map1 <- ggplot(world_map1, aes(x = long, y = lat, group = group)) +
   geom_polygon(aes(fill = factor(memb6)))
world_map1 + scale_fill_manual(values = c("red", "green", "blue", "orange", "purple", "cyan")) #添加颜色
```