# 基于 R 语言的衡量性别差异的研究分析

姓名 于沁涵 学号：1910227

## 一、背景介绍

一个婴儿出生时，医生总是会跟大人们说的一句话是：“恭喜啊，是个男孩/女该。”很早之前，人们通过一些特征，将这个大群体分为了两类——男人和女人。不可否认的是，他们之间有很多不同的地方。从生理结构到社会责任、兴趣爱好等，男女之间都有着比较明显的不同。在做同一个决定的时候，不同性别的人往往也会得到不同的建议。例如，取名字的时候，女生都会起带文雅字号的名字，例如萍、怡、静；而男生往往会包含雄、国、成等字眼。或者，做了同一件事的人会因为性别不同而得到不同的评价。比如在有些人看来，女博士是“第三种人”，且大多相貌平庸、性情古怪。身居高位或工作繁忙的女性，会被请教“如何平衡家庭与工作”。

甚至，人们对待一些事的时候往往会带上性别的字眼，将性别划分的很清晰，甚至人们对待一些事情的时候会因为性别产生不公。例如，有些公司录取人的时候，会以女性是否有结婚或者备孕的打算而限制录取，而男性则不会有这样的限制；再比如，当女性因为家暴而进入互联网的舆论中时，大众的第一反应往往是女性在婚姻中是否做到了淑德、自爱，而不是如何维权，如何保护自己。

因此，男女之间的差异不能就这样被忽视。但是，这个差别具体有多大呢？从封建社会开始，男主外女主内的观念就已经根深蒂固，男性被赋予责任、有能力的标志，女性被赋予多才、贤惠的标志。但是如今的社会已经慢慢在改变这种趋势，男女平等，女性独立的主题不断的被当代人们强调，但是仍有一些因素让男女之间的差异始终不可避免，这些差异或来自于本身的生理因素，抑或是来自于历史发展而来的长期观念。

面对这种复杂问题，我们要建立一套统一的视角和评价标准，使参与讨论的各方使用同一种语言对话。本案例以衡量性别差异问题为例，基于 R 语言，为大家展示一个优秀指标体系的范例，以及如何通过它所产生的数据了解和评价现实状况，以解决衡量性别差异进展的问题。

## 二、数据来源和说明

本文数据包含两部分：第一部分数据为 2006~2017 年，GGGR 提供的各国与全球性别差异指标体系分项得分和性别差异得分；这是一个典型的、同时包含了截面与时间因素的面板数据；第二部分数据为各国的基本属性，如所处地区、主要宗教、主要官方语言、收入水平、人口规模等。这部分数据，由 2017 年 GGGR 及互联网获得，是典型的截面数据。

## 三、描述性分析

### （一）指标体系介绍

我们首先介绍指标体系的概念。《全球性别差异报告》（简称 GGGR）是一份展示男女间在经济地位、学习机会、政治参与及卫生福利四个范畴中的差距的报告，由世界经济论坛发布，首次报告于 2006 年，覆盖全球 150 个国家和地区，各期报告使用统一的指标定义及计算口径以便于衡量各国和全球在弥合性别差异方面取得的进展或倒退。

GGGR 的指标体系由指标、分项得分和性别差异得分三层逐级加权汇总而成。性别差异得分首先被细化为经济参与度与机会（Economic Participation and Opportunity）、教育程度（Educational Attainment）、健康与生存（Health and Survival）、政治赋权（Political Empowerment）四个分项。每个分项下又下辖若干指标，四个分项共计细化为 14 个指标。

从衡量性别差异这一目标出发，这个指标体系在设计时遵循了以下 2 个原则：

1、衡量差异，而非绝对发展水平。

2、衡量结果，而不考虑相关的原因。

通过这个设计，GGGR 得以以一个非常简洁的指标体系来反映性别差异这个异常复杂问题的全貌。

GGGR 从各个数据源收集好数据，计算出每个国家的各项指标取值后，对其进行逐级汇总：首先，将每个国家的指标取值按照固定权重进行加权平均，得到每个国家的分项得分；然后，计算四个分项得分的算数平均数，得到各国性别差异得分；接下来，以人口为权重，对各国的分项得分进行加权平均，得到全球的分项得分，再取全球分项得分的算数平均数为全球性别差异得分。

### （二）数据加工

接着我们对数据进行分析。将数据读入后，我们对数据进行了预处理加工。对于面板数据，将指标名称进行划分，将 Global Index 命名为“综合指数”，其他命名为“分项指数”。

对于截面数据，将主要宗教进行划分，其中“天主教”、“基督教”、“东

正教”均统一为“基督教”，其他宗教名称保留不变；将主要语言进行划分，除“英语”、“西班牙语”、“阿拉伯语”、“法语”、“葡萄牙语”保留外，其余均命名为“其他”。

此外，为了更好的进行数据可视化，我们为面板数据新增一列，显示各项指标名称的中文名称。其中，英文变量名和中文变量名的对应关系如下：

(1)Political Empowerment：政治赋权

(2)Health and Survival：健康与生存

(3)Education Attainment：教育程度

(4)Economic Participation and Opportunity：经济-参与度与机会

(5)Global Index：综合指数

下面，我们进行绘图，将数据可视化。

### （三）全球性别得分差异

首先，我们对因变量得分进行分析。我们绘制直方图观察 2017 年全球得分情况，如图 1。

从图中可以得到，综合指数的得分在 70%左右，这意味着在全球范围内，女性与男性的总体差异已经被消灭了近 70%。但四个分项得分差异较大，其中，得分最高的是健康与共存，得分最低的是政治赋权，四个分项得分从高到低的排序为健康与生存、教育程度、经济-参与度机会、政治赋权。这意味着在教育程度和健康与生存上，几乎 90%的性别差异被消灭了；但是在政治赋权上，性别差异只被消灭了 25%，这意味着与男性已达到的水平相比，女性只取得了不到 25%的进展。因此，在最贴近人们生活的两方面健康与生存和教育程度，男女之间的差异已经基本平衡，但是在政治、经济方面，女性的地位仍和男性之间存在差异。

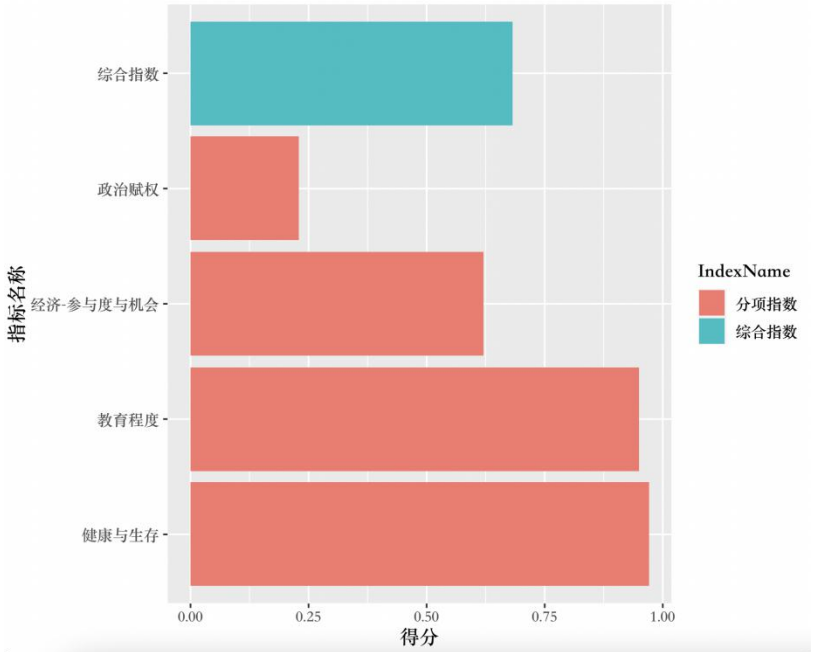


图 1 2017 年全国性别差异得分

### （四）各国的得分分布

下面，我们对各国的得分进行分析。我们绘制箱线图观察 2017 年各国得分分布情况，如图 1。

从图中可以得到，综合指数的得分在 70%左右，这意味着在全球范围内，女性与男性的总体差异已经被消灭了近 70%。但四个分项中，得分最高的是教育程度，得分最低的是政治赋权，四个分项得分从高到低的排序为教育程度、健康与生存、经济-参与度机会、政治赋权。

因此，在教育程度和健康与生存上，近 90%的性别差异被消灭了，男女差异可以认为基本消除；但是在政治赋权上，性别的差异仍有一定的差距。同时，对于不同国家，在教育程度、健康与生存两个进展最好的分项中，各国的进度相差也相对较小，而在经济和政治两个进展较慢的分项中，各国的情况差异相对较大。例如，在经济-参与度与机会这个分项中，布隆迪位列第一，消除了 85%左右的性别差异，而叙利亚仅消除了 25%的性别差异。

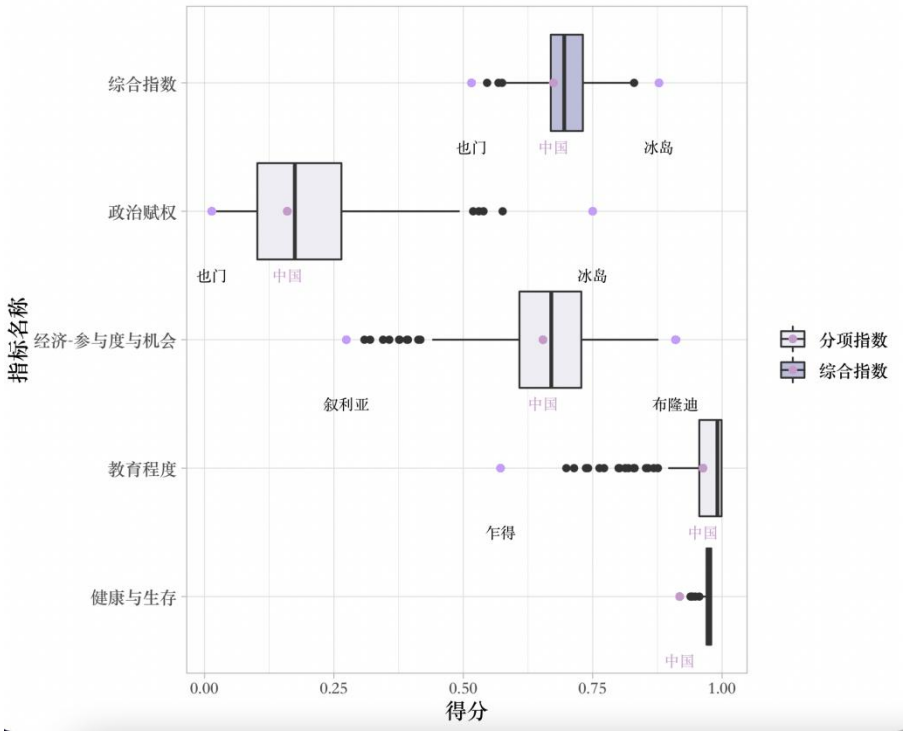


图 2 2017 年各国得分分布箱线图

再次将视线聚焦于中国。2017 年中国的性别差异得分在 144 个参评国家中排名第 100，四个分项指标也全部处于各国得分中位数之下，尤其是健康与生存一项排名倒数第一，这是由于中国的出生人口性别比失衡造成的。

大量统计数据表明，在不受干扰的情况下，人类婴儿的性别比例基本相同，通常在 102 至 107 之间，即出生 100 个女孩，男孩数量为 102 至 107 个。然而在 1982 年以前，中国新生儿的性别比基本保持在 1.07。1982 年以后，新生儿的性别比显着增加，男多于女的情况出现了一个小幅增加，到了 90 后、00 后时代，男多于女的状况，更为严重。造成这种现象的根本原因在于自古以来“重男轻女”

的观念以及现代社会丁克观念、独生子女政策的影响，目前，这个问题仍没有被很好的解决。因此，中国也需要在未来聚焦于这个问题，寻找一条合适的解决道路。

因此，我们得到结论：中国在消除性别差异上的进展并不理想，在个别方面表现亟待提高。

### （五）各国综合指数得分与收入水平的关系

下面，我们对探究性别差异得分与各影响因素之间的关系。首先，我们探究各国综合指数得分与收入水平的关系。我们绘制箱线图得到结果如图 3。

从图中可以得到，低收入的得分最低，高收入和中上收入的得分最高，这说明随着收入降低，国家的性别差异呈现逐渐拉大的趋势。即国家的富裕程度与性别平等之间存在正向的相关关系。

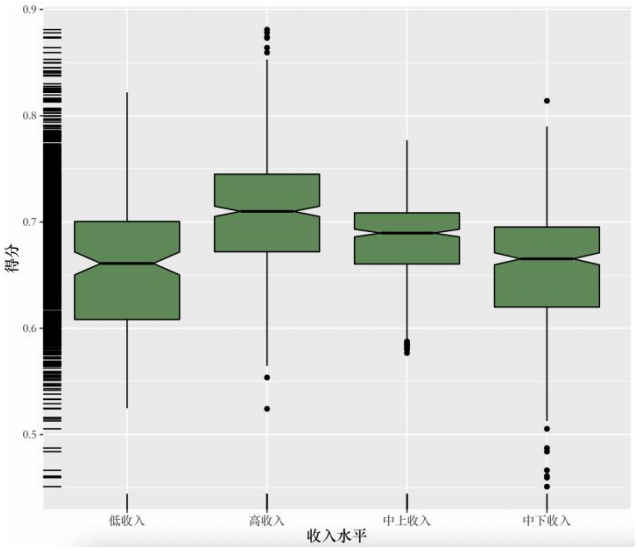


图 3 各国综合指数得分与宗教信仰的关系箱线图

### （五）各国综合指数得分与宗教信仰的关系

下面，我们探究各国综合指数得分与宗教信仰的关系。我们绘制箱线图得到结果如图 4。

从图中可以得到，信仰伊斯兰教的国家的性别差异得分最低，而信仰基督教的国家性别差异得分最高，其中信仰不同宗教的国家的性别得分差异从高到低排名为基督教、佛教、其他、伊斯兰教。这说明宗教信仰与国家消除性别差异的程度之间也存在关联。

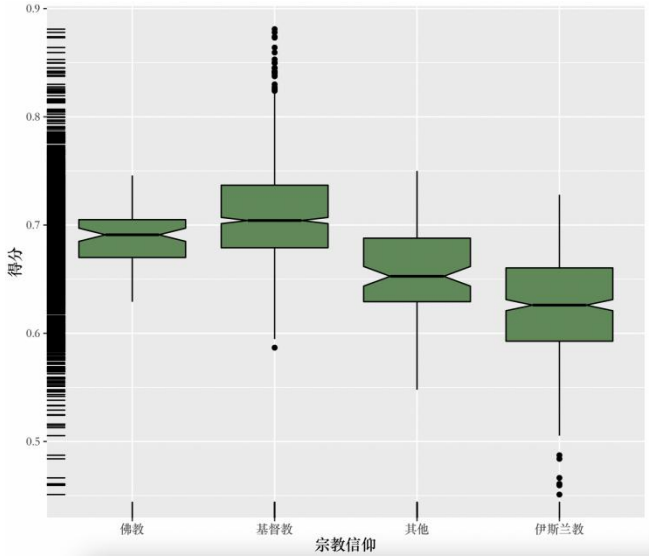


图 4 各国综合指数得分与宗教信仰的关系箱线图

同时，收入水平与宗教信仰之间也有一定的联系。资料显示，2017 年参评国家中共有高收入国家 48 个，其中信仰伊斯兰教的国家 6 个，得分水平在 48 个国家中分别名列第 38 名、第 42~46 名；而排名前 4 位的北欧四国，所信仰的主要宗教皆为基督教。这也正好与高收入水平得分分散程度大相对应。

接着，我们对各分项的宗教信仰进行更深入一步的探究。

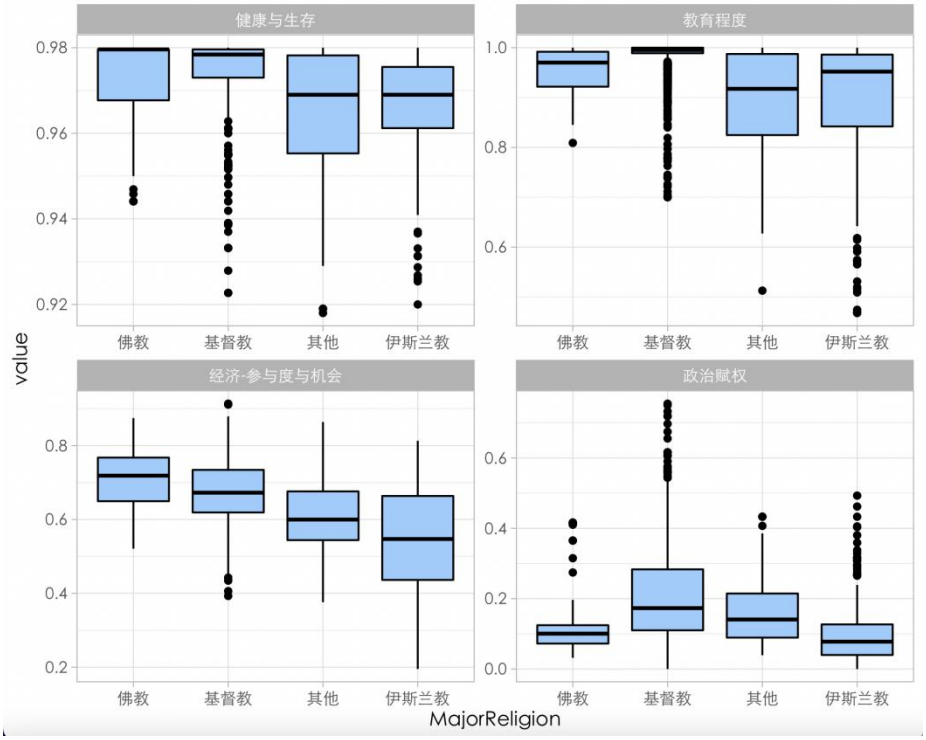


图 5 2017 年按主要宗教划分的各国各领域性别差异

由图 5 得到，信仰基督教国家在各分项中的表现都位列前茅，说明信仰基督教国家在消除性别差异的发展上展现最好；信仰佛教的国家在四个分项中，只有政治赋权方面表现相对保守，其他三个分项中也展现很好；而信仰伊斯兰教的国

家在四个分项中得分普遍偏低。这说明宗教信仰对性别差异的影响已经渗透进人们的生活，不容忽视。

### （六）各国综合指数得分变化

最后，我们利用散点图来描述，各国综合指数得分的变化，取 2006 年和 2017 年都参选的国家绘制散点图，得到结果如图 6。

由图 6 可以得到在 2006～2017 年间，大部分国家在推进性别平等方面取得了进步。无论是在性别平等程度总体较高的欧洲，还是消除性别差异情况不特别理想的非洲和亚洲，都有国家取得了非常明显的进步。但也有十几个国家，性别差异反而在进一步拉大，而这些国家主要分布在亚洲。

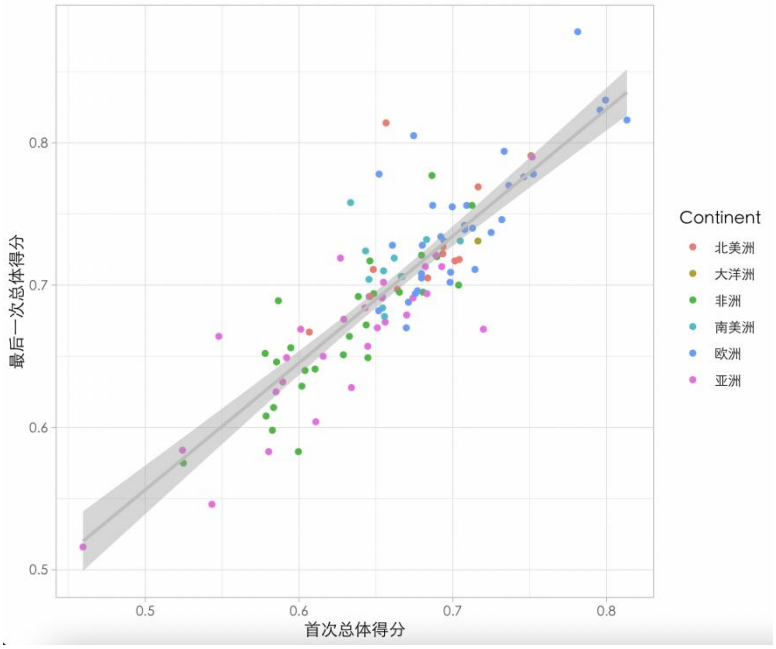


图 6 各国首次总体得分和最后一次总体得分

## 七、结论

通过上述结果，我们基于 R 语言对性别差异问题进行研究，探索了在 GGGR 的指标体系下，全球及分项指数得分的分布、各国指数得分与收入、宗教信仰的关系以及得分变化，得到全球、各国的消除性别差异进展情况。我们遵循由全貌至细部、由核心至影响因素、由截面状况至变化趋势的思路，逐步掌握了性别差异问题的影响因素和变化趋势，得到一些结论，并可以为现实中致力于消除性别差异的国家和人民提供一些合理建议。

其中，我们建议女性在政治、经济方面积极努力，加大女性的参与度，此外，也要注意收入水平和宗教信仰可能带来的性别差异的影响。但可观的是通过分析我们得到各国在消除性别差异的进展上都取得了不错的进步，因此，我们对待这个问题可以有坚决的信心，努力实现性别平等的社会。

## 附录 1　R 语言程序代码

```
###加载包
library(sqldf)
library(ggplot2)
library(gridExtra)
library(dplyr)
###Task 1　读入数据，了解数据概况
data <- read.csv('/Users/yuqinhan1229/Desktop/gggr.csv')
cntr <- read.csv('/Users/yuqinhan1229/Desktop/country.csv')
summary(data)
summary(cntr)
###Task 2　加工数据
####对 data 中 IndexName 进行划分
for (i in 1:8170){
  if(data[i,3]=='Global Index'){
    data[i,3]= '综合指数'
  }
  else{
    data[i,3]= '分项指数'
   }
}

####对 cntr 中 MajorReligion 进行划分
for (i in 1:150){
  if((cntr[i,7]=='天主教')|(cntr[i,7]=='基督教')|(cntr[i,7]=='东正教')){
    cntr[i,7]= '基督教'
  }
}

####对 cntr 中 MajorLanguage 进行划分
for (i in 1:150){
  if((cntr[i,4]!='英语')&(cntr[i,4]!='西班牙语')&(cntr[i,4]!='阿拉伯语')&(cntr[i,4]!='
法语')&(cntr[i,4]!='葡萄牙语')){
    cntr[i,4]= '其他'
  }
}
```

```
####Task 3　加工数据，生成新变量 IndexNameChs
data1 <- read.csv('/Users/yuqinhan1229/Desktop/gggr.csv')
attach(data1)
data$IndexNameChs[IndexName=='Political Empowerment']='政治赋权'
data$IndexNameChs[IndexName=='Health and Survival']='健康与生存'
data$IndexNameChs[IndexName=='Education Attainment']='教育程度'
data$IndexNameChs[IndexName=='Economic Participation and Opportunity']='经济-
参与度与机会'
data$IndexNameChs[IndexName=='Global Index']='综合指数'
detach(data1)


###Task 4 绘图，查看 2017 年全球得分情况
p1<-ggplot(data=data[data$country=='World',], mapping=aes(x = value, y =
IndexNameChs,fill=IndexName))+xlab("得分")+ylab("指标名称")+
   geom_bar(stat="identity",position=position_dodge(0.75))+
   theme(text=element_text(family="Songti SC",size=12,face = "bold"))
p1


###Task 5 绘图，查看 2017 年报告中各国的得分分布
ggplot(data[(data$country!="World")&(data$year=="2017"),],aes(x=value,y=IndexN
ameChs,fill=IndexName))+geom_boxplot()+labs(fill="",x="得分",y="指标名称
")+scale_fill_brewer(palette = "Purples")+
   geom_point(x=0.516,y=5,color="#CC99FF")+annotate('text',family="Songti
SC",x=0.516,y=4.5,label=expression("~也门~"),size=3)+
   geom_point(x=0.878,y=5,color="#CC99FF")+annotate('text',family="Songti
SC",x=0.878,y=4.5,label=expression("~冰岛~"),size=3)+
   geom_point(x=0.674,y=5,color="#CC99CC")+annotate('text',family="Songti
SC",x=0.674,y=4.5,label=expression("~中国~"),size=3,color='#CC99CC')+
   geom_point(x=0.014,y=4,color="#CC99FF")+annotate('text',family="Songti
SC",x=0.014,y=3.5,label=expression("~也门~"),size=3)+
   geom_point(x=0.75,y=4,color="#CC99FF")+annotate('text',family="Songti
SC",x=0.75,y=3.5,label=expression("~冰岛~"),size=3)+
```

```
   geom_point(x=0.16,y=4,color="#CC99CC")+annotate('text',family="Songti
SC",x=0.16,y=3.5,label=expression("~中国~"),size=3,color='#CC99CC')+
   geom_point(x=0.2740,y=3,color="#CC99FF")+annotate('text',family="Songti
SC",x=0.2740,y=2.5,label=expression("~叙利亚~"),size=3)+
   geom_point(x=0.9100,y=3,color="#CC99FF")+annotate('text',family="Songti
SC",x=0.9100,y=2.5,label=expression("~布隆迪~"),size=3)+
   geom_point(x=0.654,y=3,color="#CC99CC")+annotate('text',family="Songti
SC",x=0.654,y=2.5,label=expression("~中国~"),size=3,color='#CC99CC')+
   geom_point(x=0.572,y=2,color="#CC99FF")+annotate('text',family="Songti
SC",x=0.572,y=1.5,label=expression("~乍得~"),size=3)+
   geom_point(x=0.963,y=2,color="#CC99CC")+annotate('text',family="Songti
SC",x=0.963,y=1.5,label=expression("~中国~"),size=3,color='#CC99CC')+
   geom_point(x=0.918,y=1,color="#CC99CC")+annotate('text',family="Songti
SC",x=0.918,y=0.5,label=expression("~中国~"),size=3,color='#CC99CC')+
   theme(text=element_text(family="Songti SC",size=12,face = "bold"))



###根据国家名称将该数据与 cntr 合并
df<- merge(cntr,data, by = 'country')
summary(df)



###绘制分组箱线图，查看 2017 年各国综合指数得分与宗教信仰的关系
df1<-df[df$IndexName=='综合指数',]
###收入水平性别差异
ggplot(df1, aes(x=IncomeLevel,y=value))+xlab("收入水平")+ylab("得分
")+ylim(c(0.5,1))+
   geom_boxplot(fill="palegreen4",color="black", notch=TRUE)+
   geom_rug(color="black")+
   theme(text=element_text(family="Songti SC",size=12,face = "bold"))

###宗教信仰性别差异
df2<-df[df$IndexName!='综合指数',]
ggplot(df1, aes(x=IncomeLevel,y=value))+xlab("宗教信仰")+ylab("得分
")+ylim(c(0.5,1))+
```

```
  geom_boxplot(fill="palegreen4",color="black", notch=TRUE)+
  geom_rug(color="black")+
  theme(text=element_text(family="Songti SC",size=12,face = "bold"))

###各国按主要宗教划分的分面性别差异
options(repr.plot.width=6, repr.plot.height=6)
ggplot(data=df2, aes(x=MajorReligion,y=value)) +
  geom_boxplot(fill="#99CCFF",color="black") +
  facet_wrap(~IndexNameChs, nrow=2,scales="free")



###查看各国综合指数得分的变化情况
df3<- df[df$year==2006,]
df4<- df[df$year==2017,]
df3 <- df3[-c(16,19)]
df4 <- df4[-c(16,19)]
###挑选数据、合并数据
df_new <- inner_join(df3, df4,by = c("country", "ChineseCountryName", "Continent",
"MajorLanguage",
                                     "Density", "LandArea", "MajorReligion",
"District", "IncomeLevel", "GDP",
                                     "GDPpercapita", "population",
"PopulationGrowthRate", "PopulationSexRation",
                                     "HumanCapitalIndex", "IndexName",
"IndexNameChs"))
df_new <- df_new[df_new$IndexName=='综合指数',]
###画散点图，拟合直线
ggplot(df_new, aes(x=value.x, y=value.y,
                           color=Continent)) + geom_point()+
  geom_smooth(method="lm", color="grey", linetype=1) +
  labs(x="首次总体得分", y="最后一次总体得分")
```