多元-02-2020270026

2020270026 王姿文

3/8/2021

Data Description

World Happiness Report

数据来自Kaggle:World Happiness Report (https://www.kaggle.com/unsdsn/world-happiness)，描述不同国家的幸福指数，此处任意挑选2016的数据来绘制简单的探索性资料分析。

```
happy <- read_csv("archive/2016.csv")
happy <- happy %>%
  rename('Happiness_Rank' = 'Happiness Rank',
         'Happiness_Score'='Happiness Score',
         'Lower_Confidence_Interval'='Lower Confidence Interval',
         'Upper_Confidence_Interval'='Upper Confidence Interval',
         'Economy_GDP'='Economy (GDP per Capita)',
         'Health'='Health (Life Expectancy)',
         'Trust_Government_Corruption' = 'Trust (Government Corruption)',
         'Dystopia_Residual'='Dystopia Residual')
dt <- head(happy)
kbl(dt) %>%
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = F, font_size = 7)
```

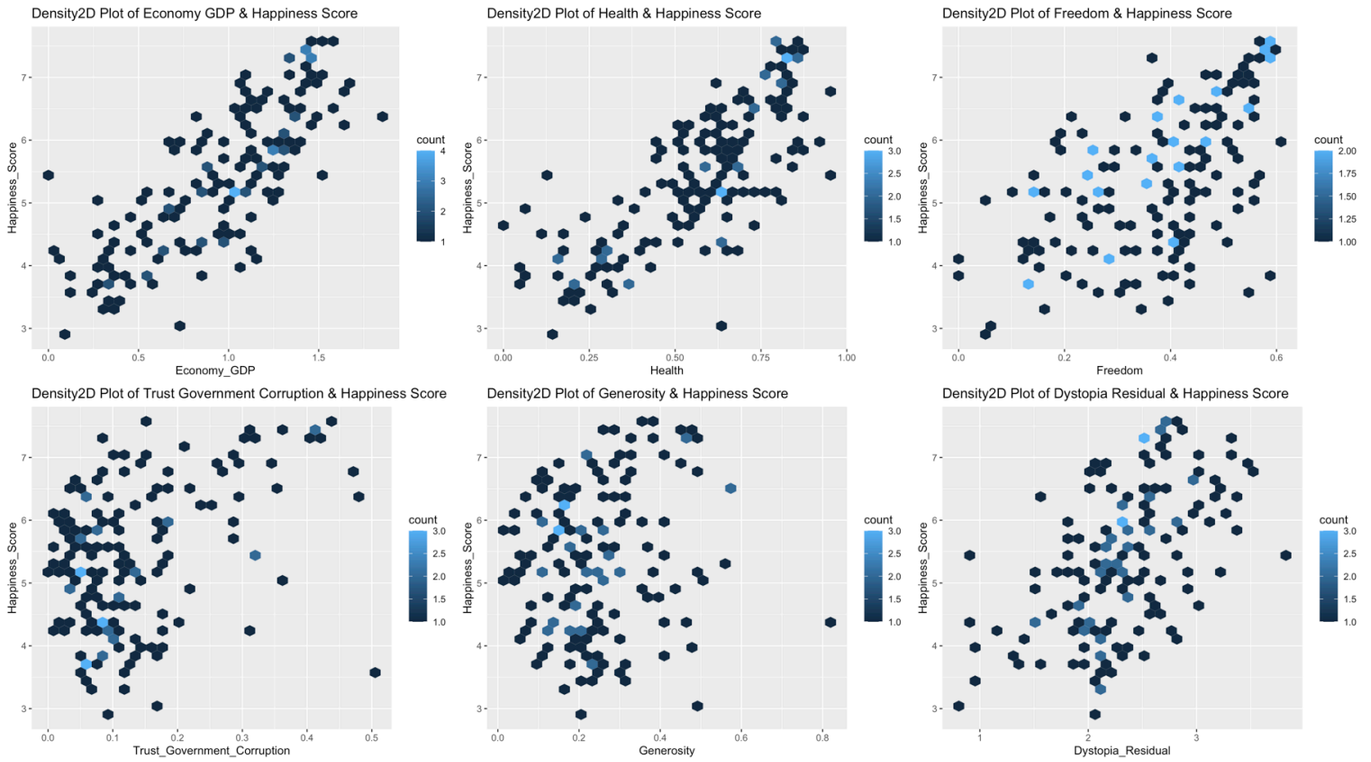
Country	Region	Happiness_Rank	Happiness_Score	Lower_Confidence_Interval	Upper_Confidence_Interval	Economy_GDP	Family	Health	Freedom	Trust_Government_Corruption	Generosity	Dystopia_Residual
Denmark	Western Europe	1	7.526	7.460	7.592	1.44178	1.16374	0.79504	0.57941	0.44453	0.36171	2.73939
Switzerland	Western Europe	2	7.509	7.428	7.590	1.52733	1.14524	0.86303	0.58557	0.41203	0.28083	2.69463
Iceland	Western Europe	3	7.501	7.333	7.669	1.42666	1.18326	0.86733	0.56624	0.14975	0.47678	2.83137
Norway	Western Europe	4	7.498	7.421	7.575	1.57744	1.12690	0.79579	0.59609	0.35776	0.37895	2.66465
Finland	Western Europe	5	7.413	7.351	7.475	1.40598	1.13464	0.81091	0.57104	0.41004	0.25492	2.82596
Canada	North America	6	7.404	7.335	7.473	1.44015	1.09610	0.82760	0.57370	0.31329	0.44834	2.70485

Exploratory Data Analysis(EDA)

蜂窩圖

此图横轴为 `Happiness_Score` ，可以非常明显看出两个变量间是否正相关以及两个变量间的计数大小。 这张图结合了Scatter Plot和Bar Plot能看出的insight，能看出 `Happiness_Score` 、 `Health` 正相关，其余则没有明显相关性。其中 `Freedom` 的计数多寡较为均匀分布。

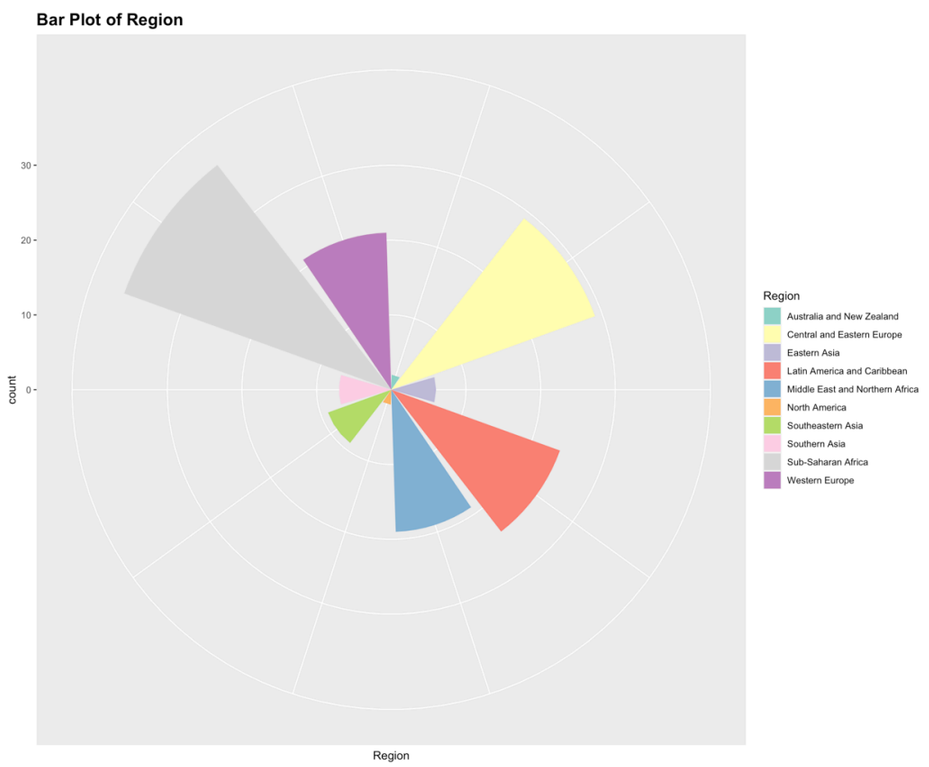
```
hp <- function(x,y,nn,nx,ny){
  ggplot(happy, aes(x, y)) +
    geom_hex() +
    labs(title =paste('Density2D Plot of ', nn,sep=''),
         x=nx,y=ny)
}
p1 <- hp(happy$Economy_GDP,happy$Happiness_Score,'Economy GDP & Happiness Score',
         'Economy_GDP','Happiness_Score')
p2 <- hp(happy$Health,happy$Happiness_Score,'Health & Happiness Score',
         'Health','Happiness_Score')
p3 <- hp(happy$Freedom,happy$Happiness_Score,'Freedom & Happiness Score',
         'Freedom','Happiness_Score')
p4 <- hp(happy$Trust_Government_Corruption,happy$Happiness_Score,'Trust Government Corruption & Happiness Scor
e',
         'Trust_Government_Corruption','Happiness_Score')
p5 <- hp(happy$Generosity,happy$Happiness_Score,'Generosity & Happiness Score',
         'Generosity','Happiness_Score')
p6 <- hp(happy$Dystopia_Residual,happy$Happiness_Score,'Dystopia Residual & Happiness Score',
         'Dystopia_Residual','Happiness_Score')
grid.arrange(p1, p2,p3,p4,p5,p6,nrow = 2)
```



星图

此图横轴为 `Region` ，可看出不同地区的数据收集计数，像是Sub-Saharan Africa的国家纪录较多。

```
ggplot(data=happy, aes(x = Region,fill = as.factor(Region))) +
  geom_bar()+
  theme(plot.title = element_text(size=16, face="bold"),
        axis.text.x = element_blank()) +
  labs(title ="Bar Plot of Region",fill = "Region") +
  coord_polar(theta = "x") +
  scale_fill_brewer(palette="Set3")
```

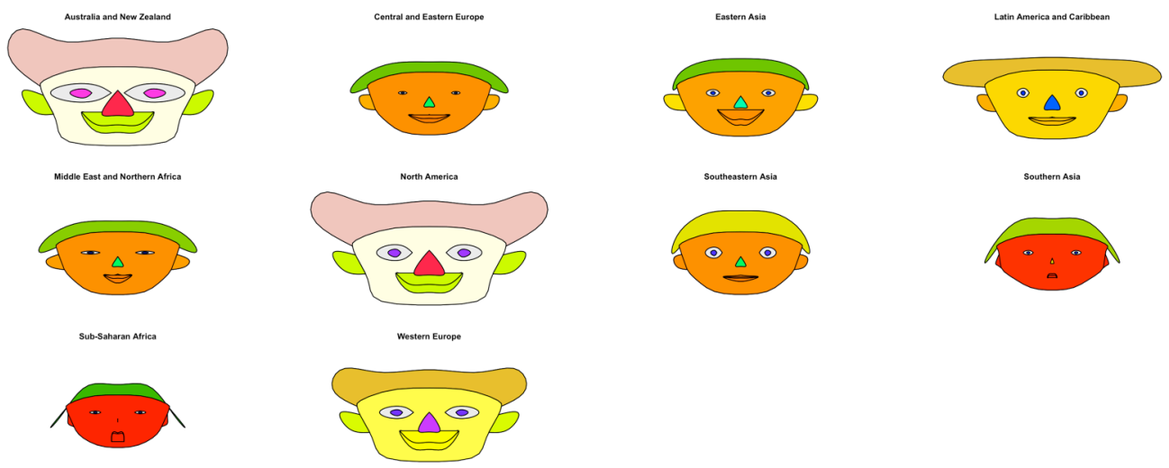


脸谱图

以下为原始的连续型变量根据 `Region` 计算平均值后绘制而成。分别为 `Region` 内的十个地区，而下表可以看出不同五官所代表的变量。
根据整张脸来看相似程度可以视为以下几组： 1. Australia and New Zealand, North America, Western Europe
2. Southern Asia, Southeastern Asia
3. Central and Eastern Europe, Middle East and Northern Africa
4. Eastern Asia, Southeastern Asia
5. Latin America and Caribbean

```
happy_g <- happy %>%
  group_by(Region) %>%
  summarise(Happiness_Score = mean(Happiness_Score),
            Lower_Confidence_Interval = mean(Lower_Confidence_Interval),
            Upper_Confidence_Interval = mean(Upper_Confidence_Interval),
            Economy_GDP = mean(Economy_GDP),
            Family = mean(Family),
            Health = mean(Health),
            Freedom  = mean(Freedom),
            Trust_Government_Corruption = mean(Trust_Government_Corruption),
            Generosity = mean(Generosity),
            Dystopia_Residual = mean(Dystopia_Residual)
  )

aplpack::faces(happy_g[,2:11],labels = happy_g$Region,cex = 1)
```

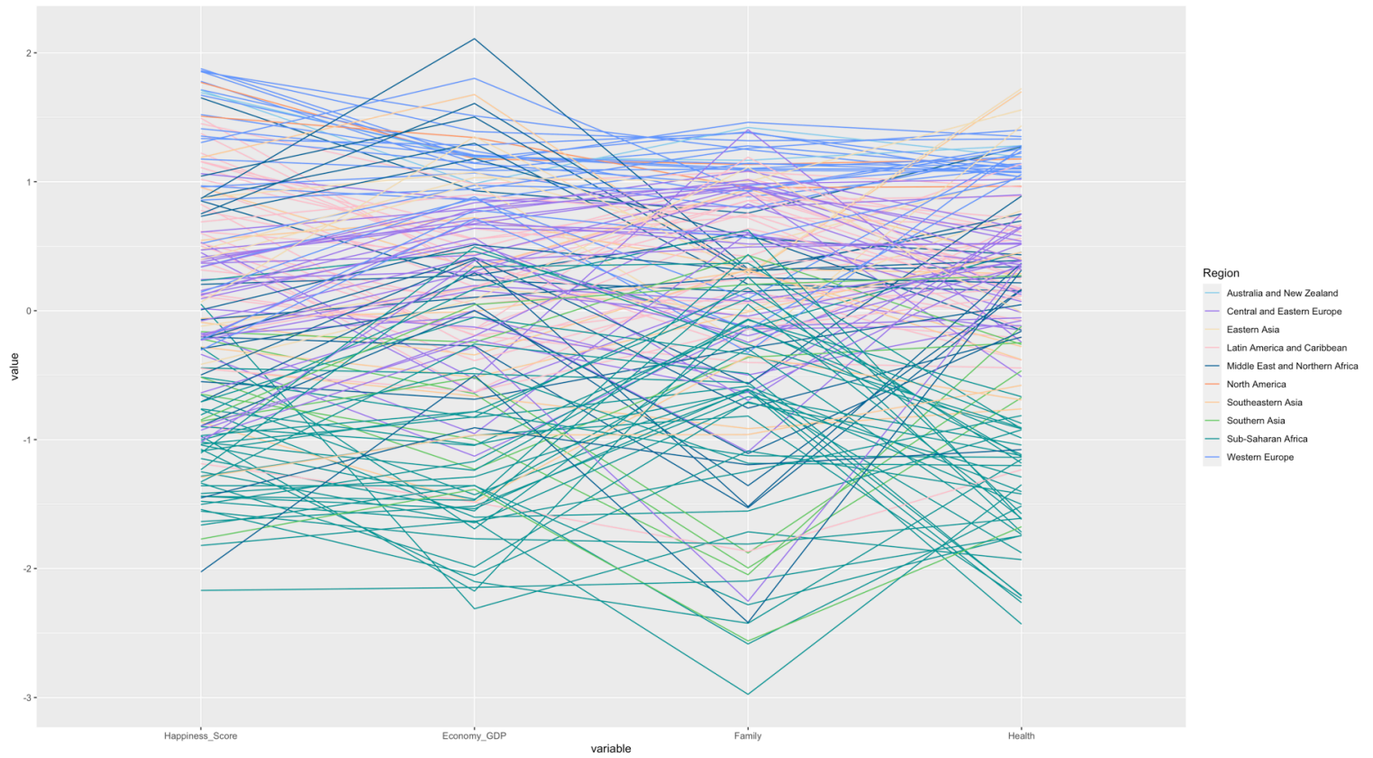



```
## effect of variables:
##  modified item       Var
##  "height of face   " "Happiness_Score"
##  "width of face    " "Lower_Confidence_Interval"
##  "structure of face" "Upper_Confidence_Interval"
##  "height of mouth  " "Economy_GDP"
##  "width of mouth   " "Family"
##  "smiling          " "Health"
##  "height of eyes   " "Freedom"
##  "width of eyes    " "Trust_Government_Corruption"
##  "height of hair   " "Generosity"
##  "width of hair    " "Dystopia_Residual"
##  "style of hair    " "Happiness_Score"
##  "height of nose   " "Lower_Confidence_Interval"
##  "width of nose    " "Upper_Confidence_Interval"
##  "width of ear     " "Economy_GDP"
##  "height of ear    " "Family"
```

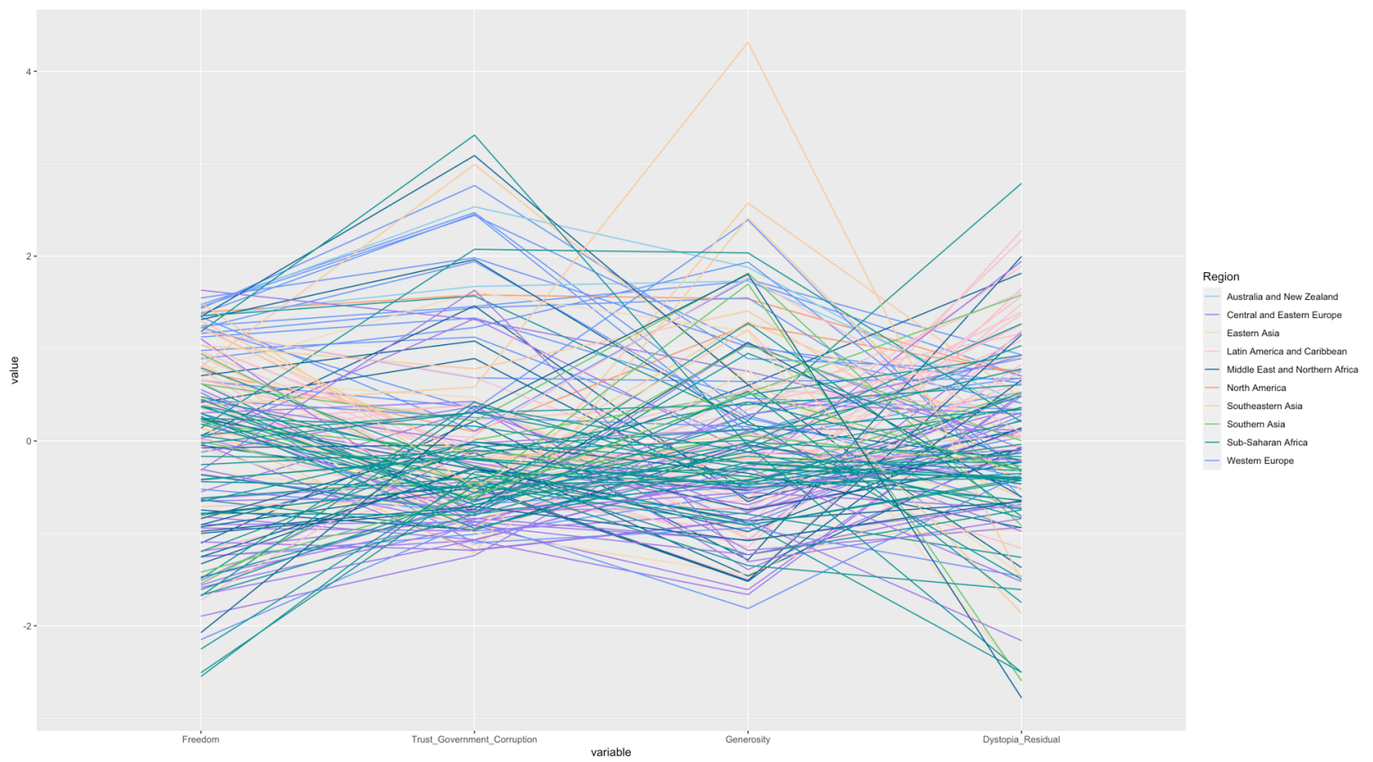
平行坐标图

以下平行坐标图的横轴分别为不同的连续型变量，可以看出因为分组（`Region`）太多，导致平行座标图的成效不太好。

```
ggparcoord(happy, columns = c(4,7,8,9), groupColumn = 2)  +
  scale_color_manual(values=c( "skyblue", "mediumpurple2", "wheat","pink","#006699",
                               "#FF9966","#FFCC99","#66CC66","#009999","#6699FF") )
```




```
ggparcoord(happy, columns = 10:13, groupColumn = 2)  +
  scale_color_manual(values=c( "skyblue", "mediumpurple2", "wheat","pink","#006699",
                               "#FF9966","#FFCC99","#66CC66","#009999","#6699FF") )
```



Conclusion

Exploratory Data Analysis(EDA)是十分重要的分析前置作业，由上述简单几张图，就能得知某些地区的幸福指数较高，也能知道哪些变量和幸福指数的分布状况呈现正相关，同时也能得知不同连续型变量的分布，以评断是否需要做后续的数据转换。后续能根据这些insight来分析并预测数据结果。

其中感到最有趣的图是脸谱图，透过简单的脸谱图，就能以整张脸来分类，且可以知道不同地区间的形状差异是来自于哪个变量，简单一张图可以看出很多细节。