多元01-2020270026

2020270026 王姿文

2021/03/01

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。 以下五种图采用课程授课所提及的五种图，绘制方式使用 `ggplot2` 。

```
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)
```

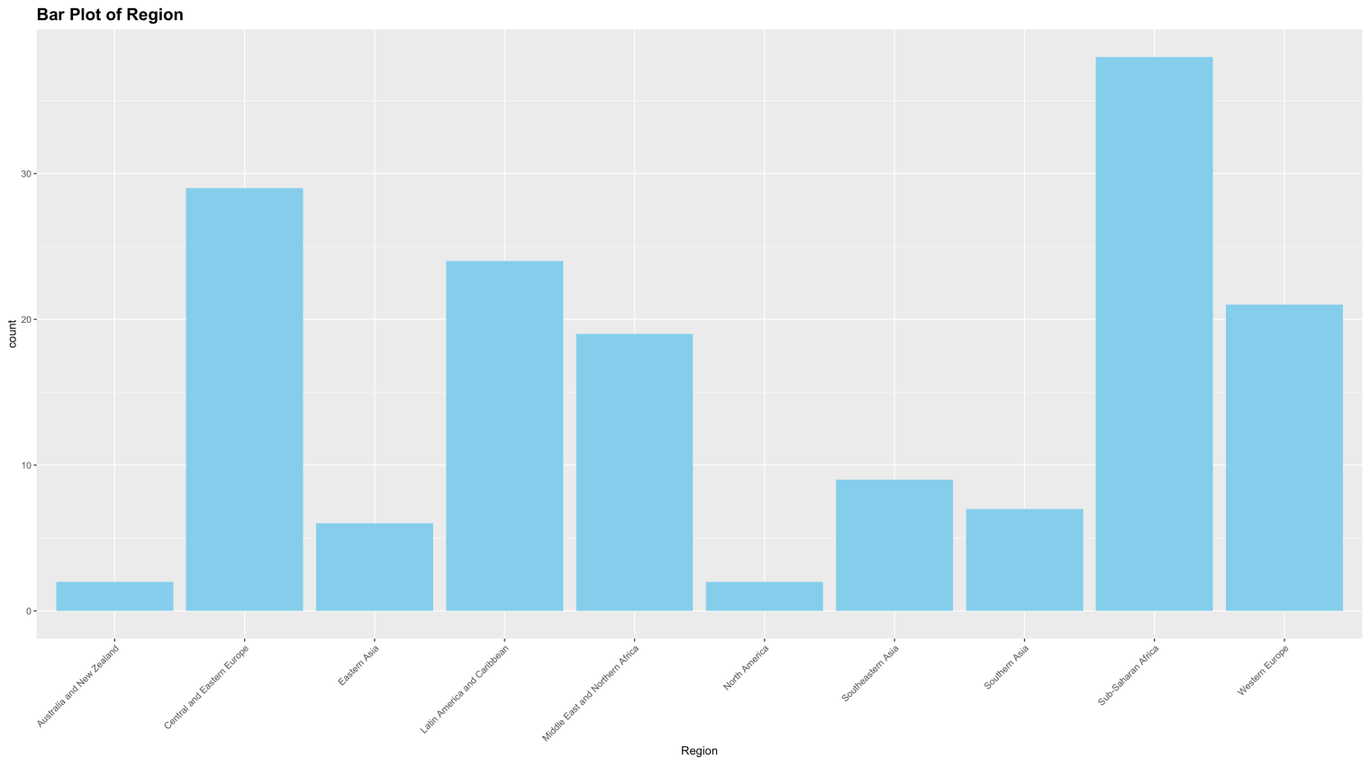
Country	Region	Happiness_Rank	Happiness_Score	Lower_Confidence_Interval	Upper_Confidence_Interval	Economy_GDP	Family	Health	Freedom	Trust_Government_Corruption	Generosity	Dystopia_Residual
Denmark	Western Europe	1	7.526	7.460	7.592	1.44178	1.16374	0.79504	0.57941	0.44453	0.36171	2.73939
Switzerland	Western 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)

Bar Plot

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

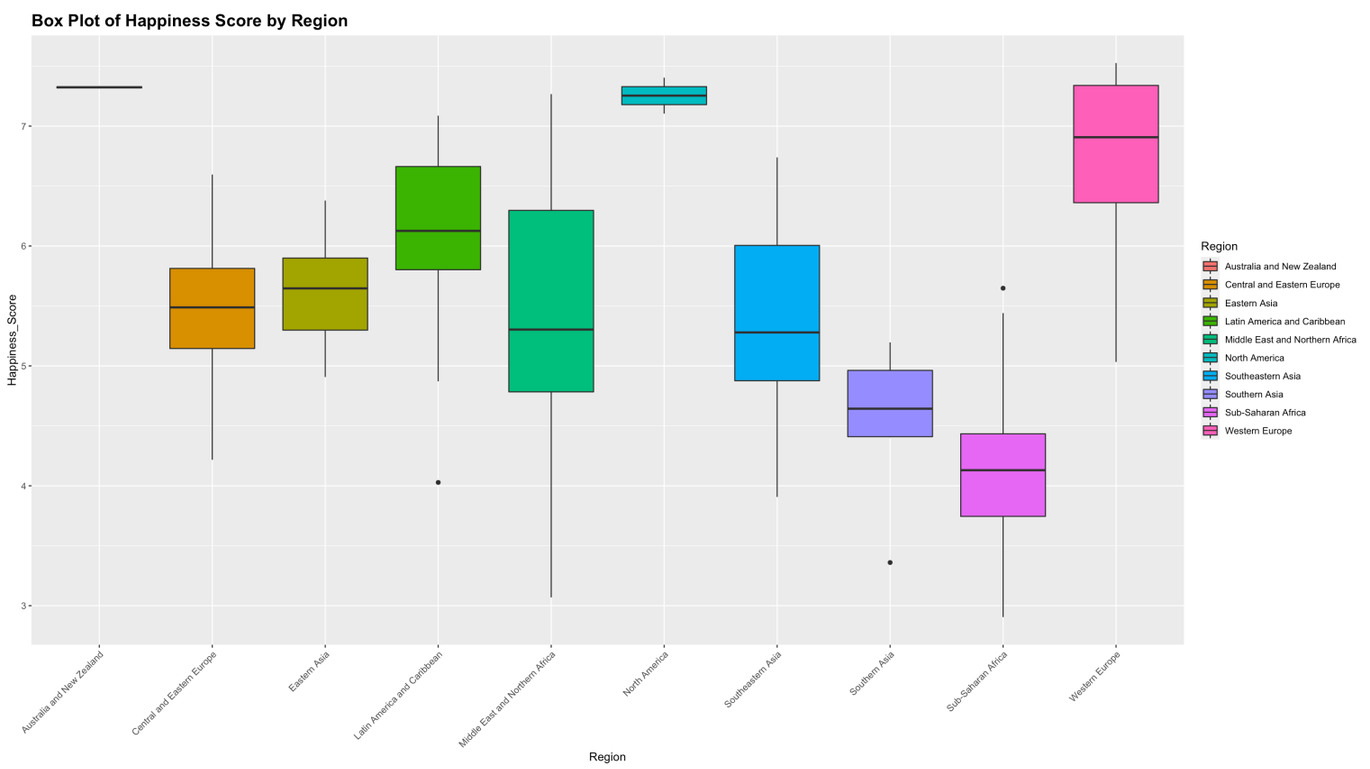
```
ggplot(data=happy, aes(x = Region)) +
  geom_bar(fill = 'skyblue')+
  theme(plot.title = element_text(size=16, face="bold"),
        axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title ="Bar Plot of Region")
```



Box Plot

此图横轴为 `Region` 、纵轴为 `Happiness_Score` ，可看出，西欧、北美、澳洲地区的幸福指数较高。

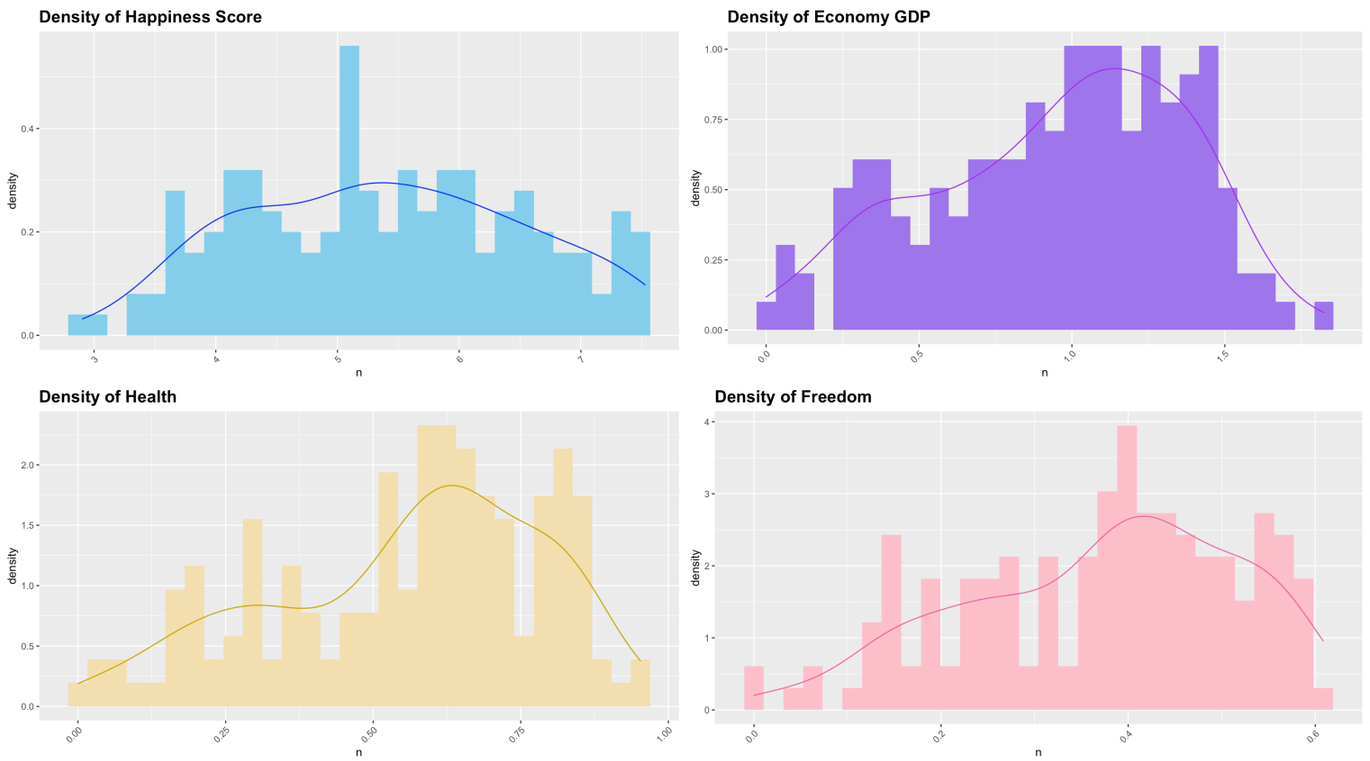
```
ggplot(happy,aes(Region,Happiness_Score,fill = Region)) +
  geom_boxplot() +
  theme(plot.title = element_text(size=16, face="bold"),
        axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title ="Box Plot of Happiness Score by Region")
```



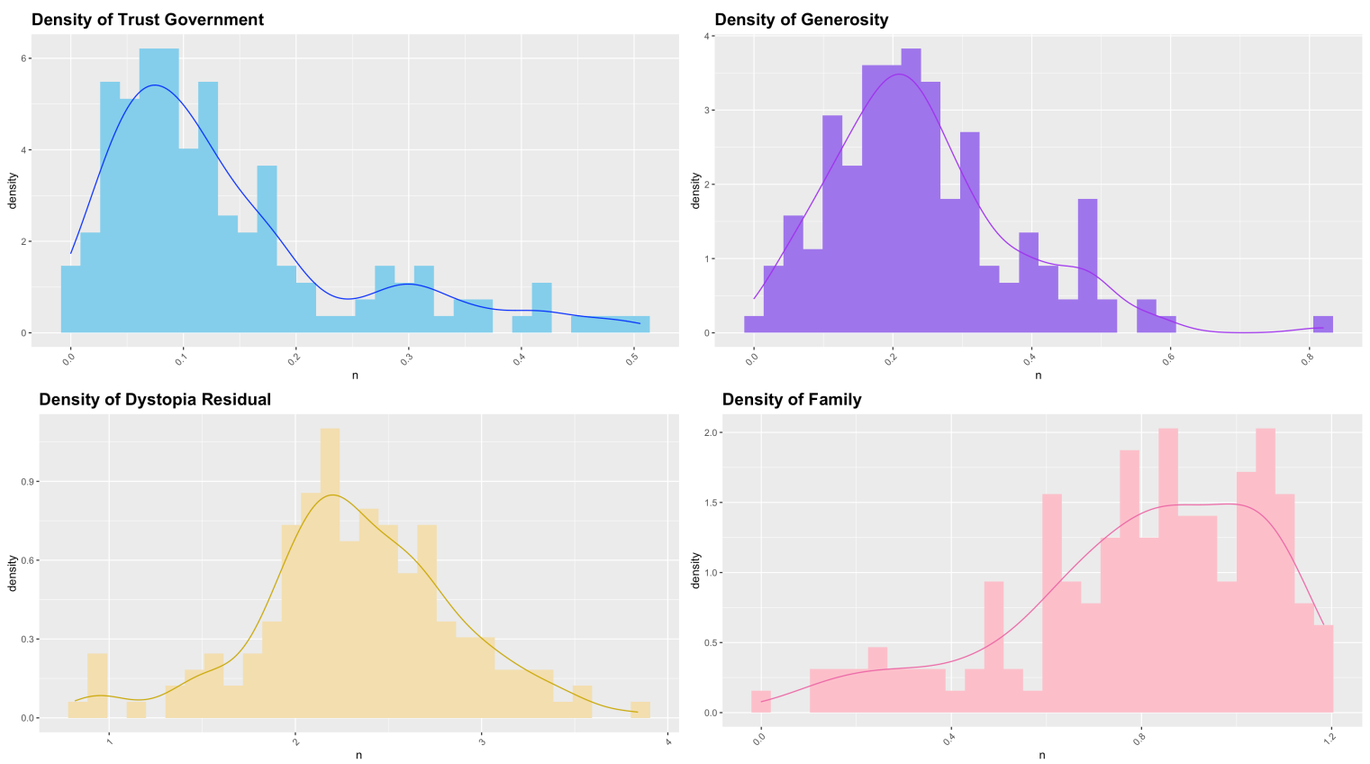
Density Plot

以下为密度图，可以看出 `Happiness_Score` 、 `Dystopia_Residual` 趋向正态分布， `Trust_Government_Corruption` 、 `Generosity` 略微右偏，其余的连续型变量则略微左偏。

```
dp <- function(n,hc,dc,nn){
  ggplot(happy,aes(x = n,y = ..density..)) +
           geom_histogram(fill = hc) +
           geom_density(color = dc) +
           theme(plot.title = element_text(size=16, face="bold"),
                 axis.text.x = element_text(angle = 45, hjust = 1)) +
           labs(title =paste('Density of ', nn,
                             sep=''))

}

p1 <- dp(happy$Happiness_Score,'skyblue','blue','Happiness Score')
p2 <- dp(happy$Economy_GDP,'mediumpurple2','purple','Economy GDP')
p3 <- dp(happy$Health,'wheat','gold3','Health')
p4 <- dp(happy$Freedom,'pink','hotpink2','Freedom')
p5 <- dp(happy$Trust_Government_Corruption,'skyblue','blue','Trust Government')
p6 <- dp(happy$Generosity,'mediumpurple2','purple','Generosity')
p7 <- dp(happy$Dystopia_Residual,'wheat','gold3','Dystopia Residual')
p8 <- dp(happy$Family,'pink','hotpink2','Family')

grid.arrange(p1, p2, p3,p4,nrow = 2)
```




```
grid.arrange(p5, p6, p7,p8,nrow = 2)
```

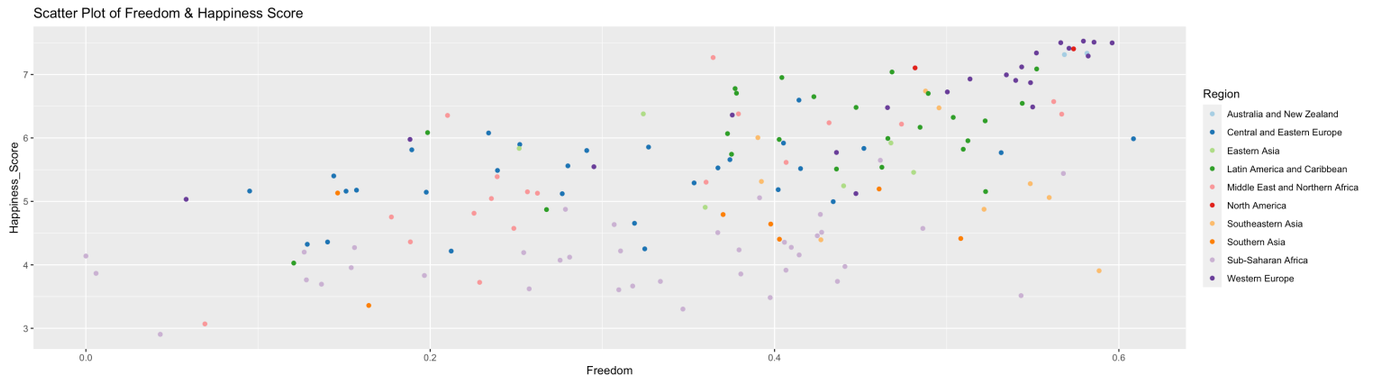


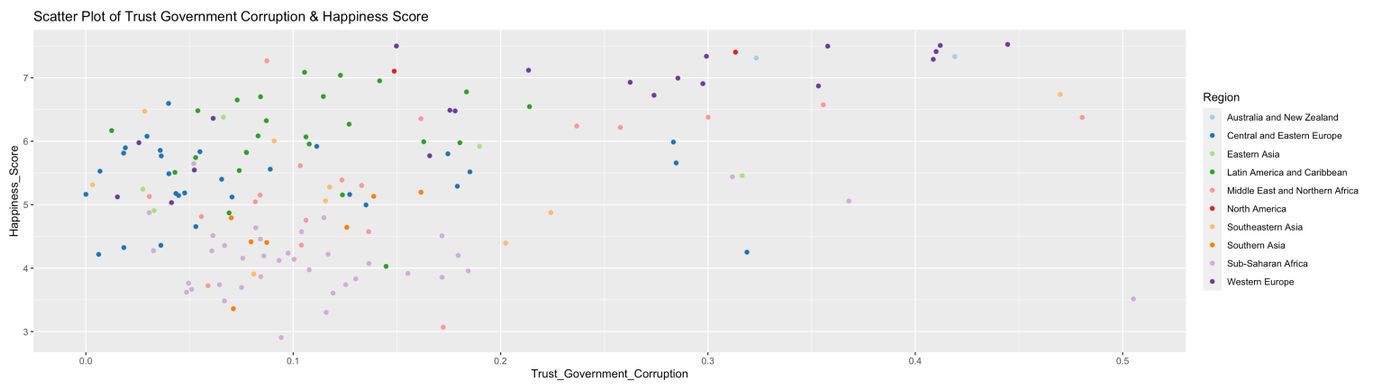
Scatter Plot

以下散布图的纵轴都是 `Happiness_Score` 。
可以看出 `Dystopia_Residual` 、 `Generosity` 、 `Trust_Government_Corruption` 三者和 `Happiness_Score` 的线性正相关较不明显。

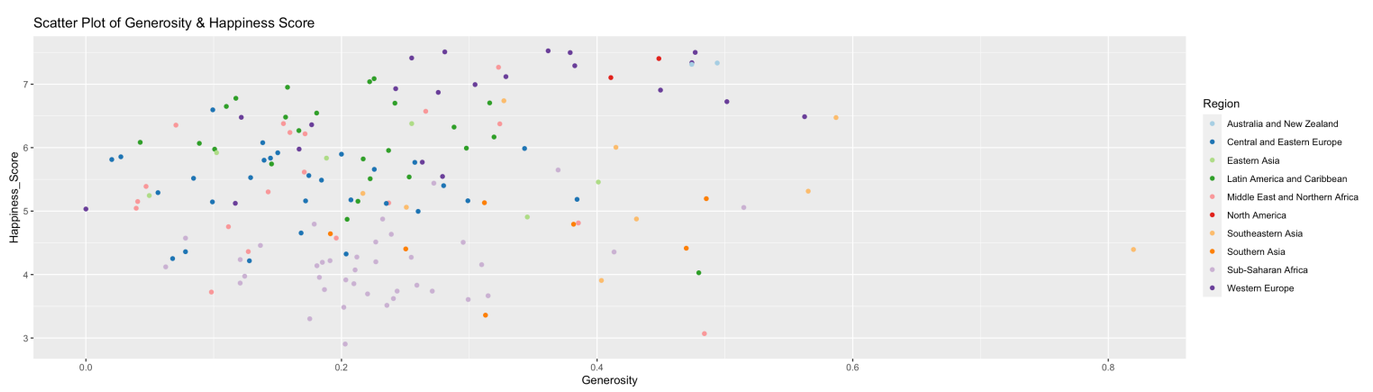
```
sp <- function(x,y,nn,nx,ny){
  ggplot(data=happy, aes(x=x, y=y,group = Region,color = Region))+
  geom_point() +
  scale_color_brewer(palette="Paired") +
  labs(title =paste('Scatter Plot of ', nn,sep=''),
       x=nx,y=ny)}

p1 <- sp(happy$Economy_GDP,happy$Happiness_Score,'Economy GDP & Happiness Score',
         'Economy_GDP','Happiness_Score')
p2 <- sp(happy$Health,happy$Happiness_Score,'Health & Happiness Score',
         'Health','Happiness_Score')
p3 <- sp(happy$Freedom,happy$Happiness_Score,'Freedom & Happiness Score',
         'Freedom','Happiness_Score')
p4 <- sp(happy$Trust_Government_Corruption,happy$Happiness_Score,'Trust Government Corruption & Happiness Scor
e',
         'Trust_Government_Corruption','Happiness_Score')
p5 <- sp(happy$Generosity,happy$Happiness_Score,'Generosity & Happiness Score',
         'Generosity','Happiness_Score')
p6 <- sp(happy$Dystopia_Residual,happy$Happiness_Score,'Dystopia Residual & Happiness Score',
         'Dystopia_Residual','Happiness_Score')
grid.arrange(p1, p2,nrow = 2)
```

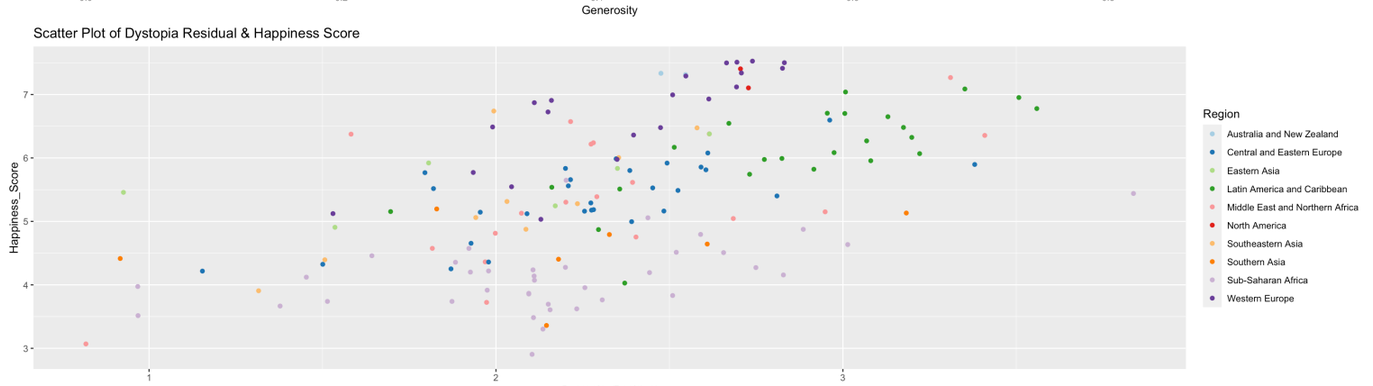

```
grid.arrange(p3, p4,nrow = 2)
```






```
grid.arrange(p5, p6,nrow = 2)
```

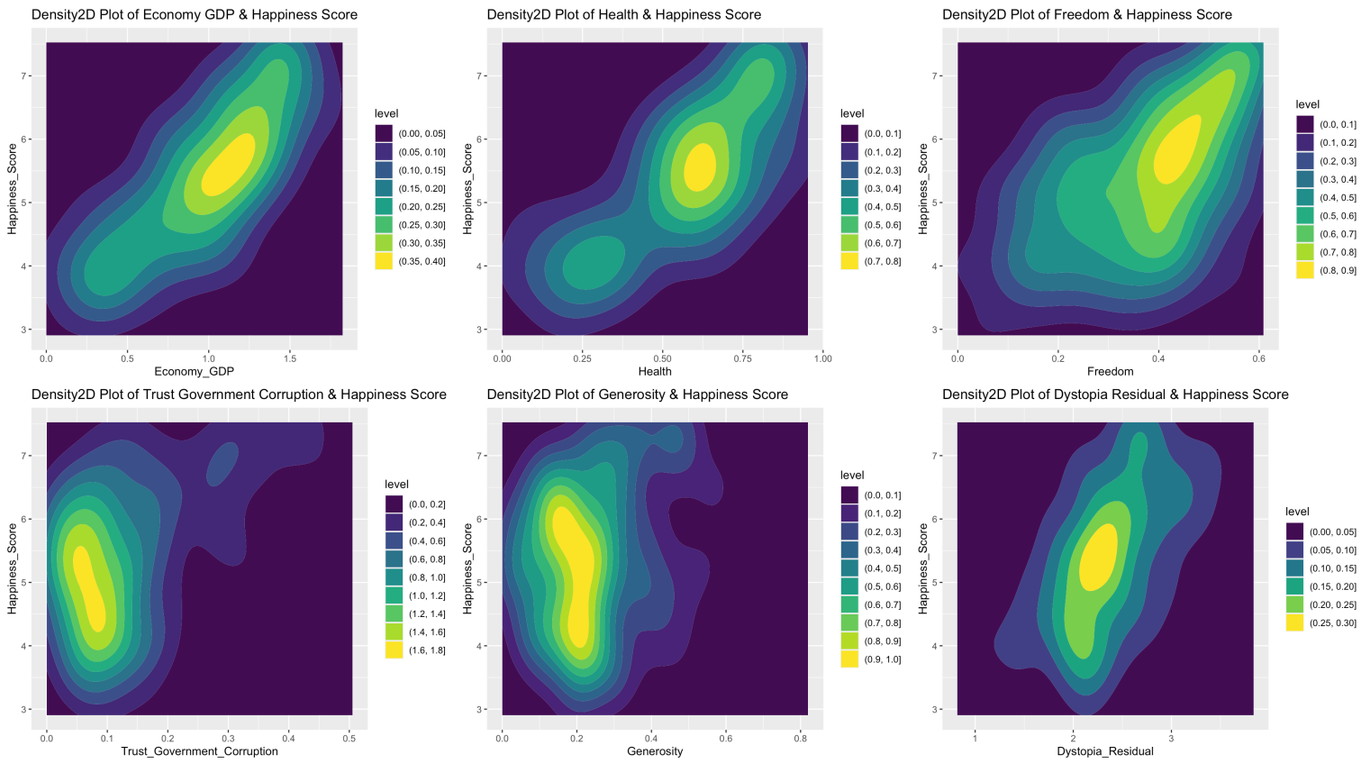




Density2D Plot

以下二元密度图的纵轴都是 `Happiness_Score` 。
可以非常明显看出分布密度以及两个变量间是否正相关。 这张图结合了Scatter Plot和Density Plot能看出的insight，一样能看出 `Dystopia_Residual` 、 `Generosity` 、 `Trust_Government_Corruption` 三者和 `Happiness_Score` 的线性正相关较不明显； `Happiness_Score` 、 `Dystopia_Residual` 趋向正态分布， `Trust_Government_Corruption` 、 `Generosity` 略微右偏，其余的连续型变量则略微左偏。

```
hp <- function(x,y,nn,nx,ny){
ggplot(happy, aes(x, y)) +
  geom_density_2d_filled() +
    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)
```



Conclusion

Exploratory Data Analysis(EDA)是十分重要的分析前置作业，例如说由上述简单几张图，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。