

# R Visualizing: 从入门到放弃

## 画出条形图



清华大学  
Tsinghua University

吴温泉

清华大学社会科学学院政治学系



## 课程结构

- I. 为什么要选择R
- II. ggplot绘图语法
- III. 色彩原理
- IV. 基本图表类型
- V. 单变量分析
- VI. 双变量分析










## 课程结构

- I. 为什么要选择R
- II. ggplot绘图语法
- III. 色彩原理
- IV. 基本图表类型
- V. 单变量分析
- VI. 双变量分析

# 为什么要选择R

- 开源免费、功能强大

LOGO	名称	开源	付费	技能要求
	Excel <sup>1</sup>	否	是	界面操作
	Origin <sup>2</sup>	否	是	界面操作
	SigmPlot <sup>3</sup>	否	是	界面操作
	GraphPad Prism <sup>4</sup>	否	是	界面操作
	MATLAB <sup>5</sup>	否	是	编程
	Python <sup>6</sup>	是	否	编程
	R <sup>7</sup>	是	否	编程

# 为什么要选择R

- 上手极快，使用简单

```
import matplotlib.pyplot as plt

labels = ['G1', 'G2', 'G3', 'G4', 'G5']
men_means = [20, 35, 30, 35, 27]
women_means = [25, 32, 34, 20, 25]
men_std = [2, 3, 4, 1, 2]
women_std = [3, 5, 2, 3, 3]
width = 0.35      # the width of the bars: can also be len(x) sequence

fig, ax = plt.subplots()

ax.bar(labels, men_means, width, yerr=men_std, label='Men')
ax.bar(labels, women_means, width, yerr=women_std, bottom=men_means,
        label='Women')

ax.set_ylabel('Scores')
ax.set_title('Scores by group and gender')
ax.legend()

plt.show()
```

Matplotlib

“坠（最）痛苦的就是把鸭嘴笔这个水  
弄到里面，描图的时候一下子就...然后  
就用刀片去刮，这个是描图坠（最）痛  
苦的”

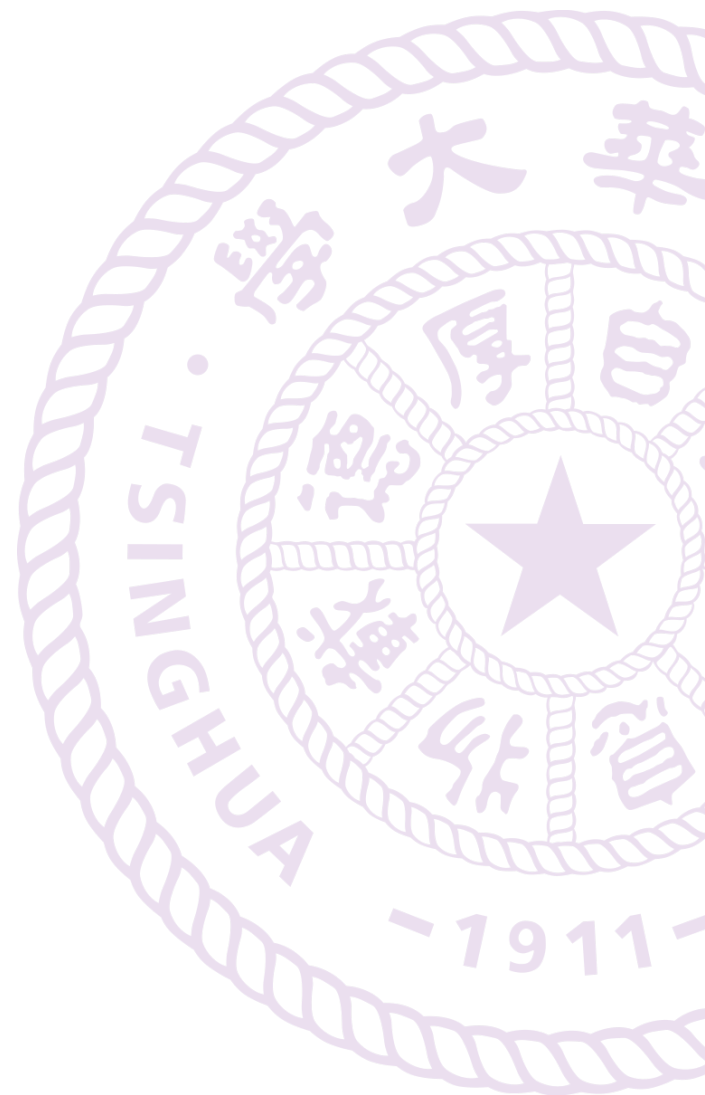
——江泽民

# 为什么要选择R

- 上手极快，使用简单

```
11  
12 ggplot(data = <data>), mapping = aes(<mappings>)) +  
13   geom_xxx() +  
14   scale_xxx() +  
15   coord_xxx() +  
16   facet_xxx() +  
17   theme_xxx()
```

ggplot

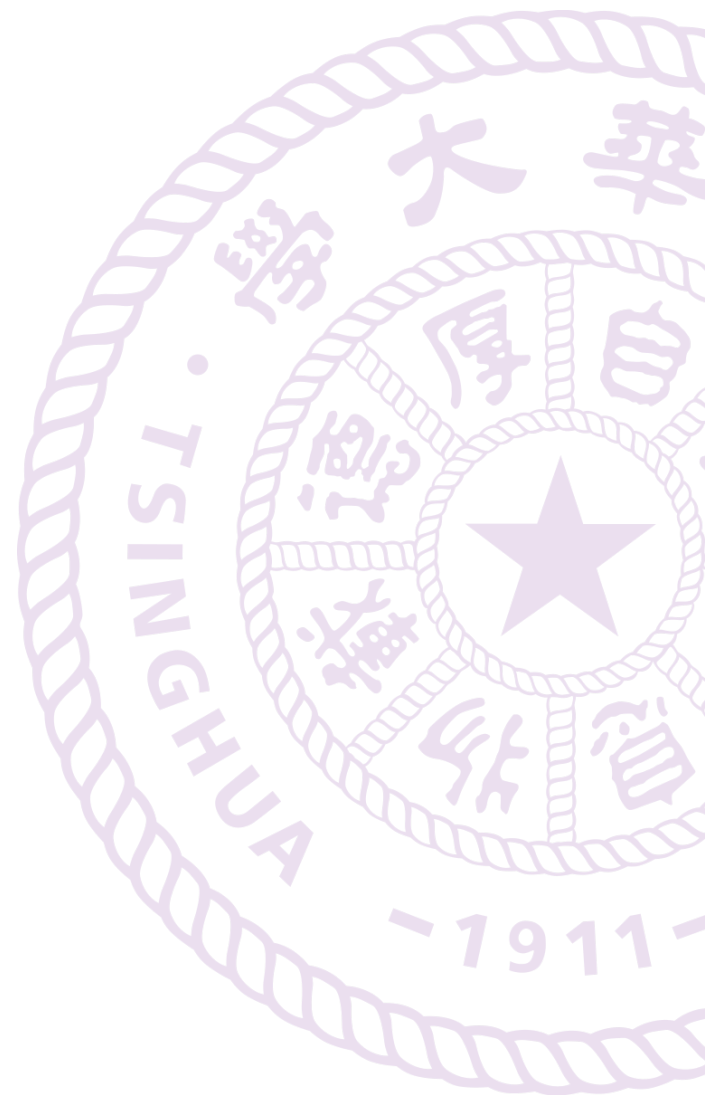


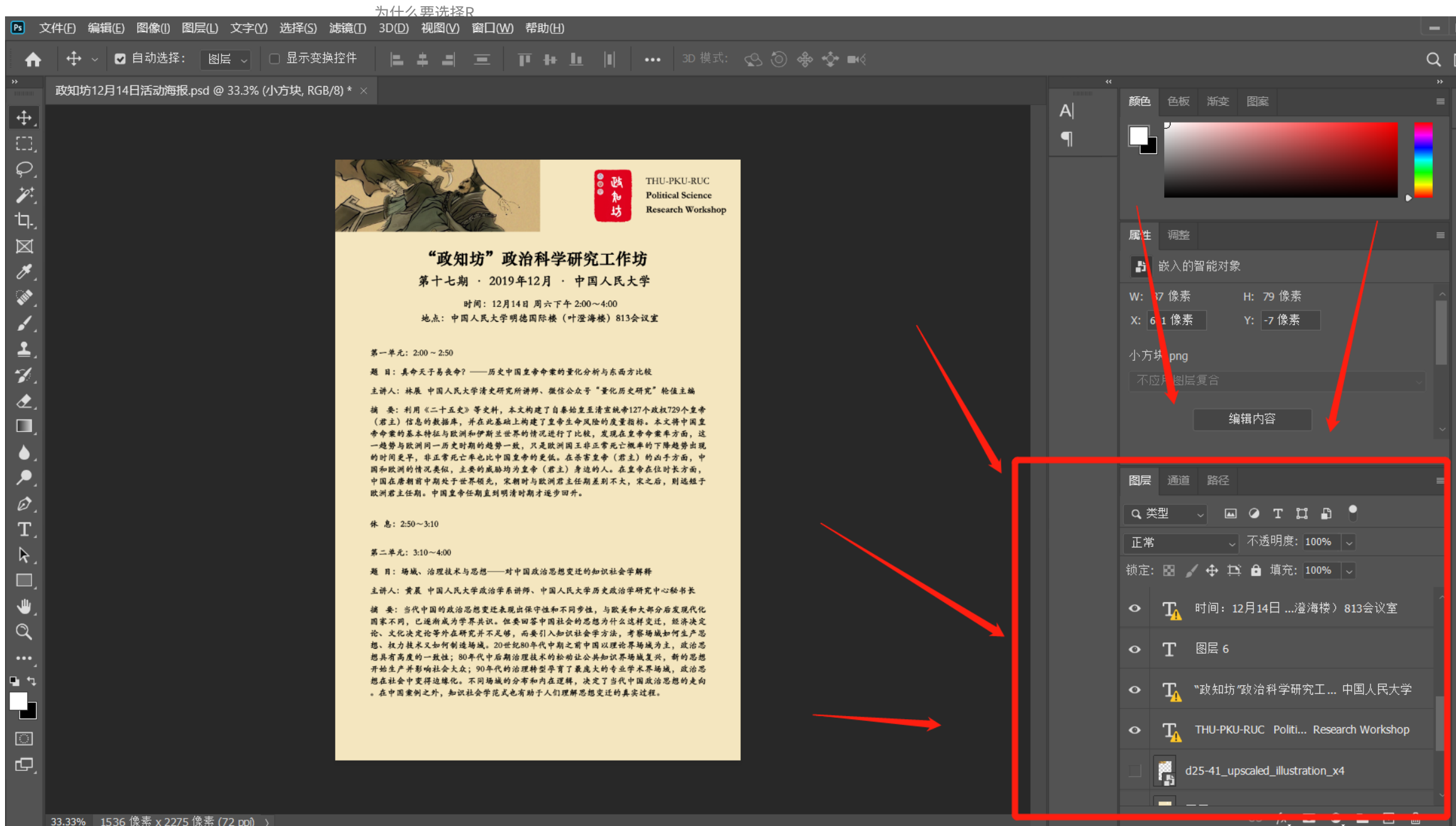
# 为什么要选择R

## ggplot (Grammar of Graphic, 绘图语法)

- 图层的设计方式, 通过 “+” 叠加
- 将数据和图形细节分离
- 图形美观

```
12 ggplot(data = <data>), mapping = aes(<mappings>)) +  
13   geom_xxx() +  
14   geom_xxx() +  
15   geom_xxx() +  
16   scale_xxx() +  
17   scale_xxx() +  
18   coord_xxx() +  
19   facet_xxx() +  
20   theme_xxx()
```









## 课程结构

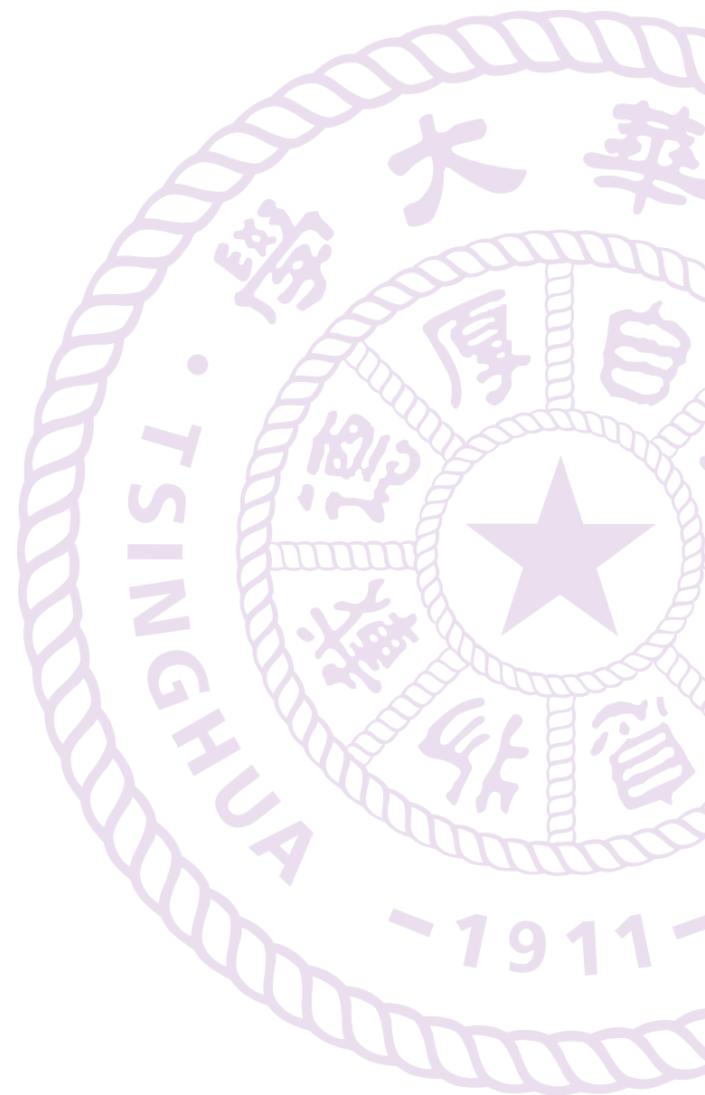
- I. 为什么要选择R
- II. ggplot绘图语法**
- III. 色彩原理
- IV. 基本图表类型
- V. 单变量分析
- VI. 双变量分析

# ggplot绘图语法

## ggplot八股文的要素

- Data 数据
- Aesthetics 美学映射
- Geometry objects 几何对象
- Theme adjustments 其他元素

```
11  
12 ggplot(data = <data>), aes(<mappings>)) +  
13   geom_xxx() +  
14   scale_xxx() +  
15   coord_xxx() +  
16   facet_xxx() +  
17   theme_xxx()
```



# ggplot绘图语法

## 极大提升coding节奏感的五个快捷键

- <-           Alt +-
- Run           Ctrl + Enter
- Tab           Tab
- Note          Ctrl+Shift+C
- Help          ?<Function>

*#在R里面打问号并不会激怒函数*



## Data 数据

- 首先，输入并run：  
view(mtcars)
- 其次，帅气的输入并run：  
library(tidyverse)

为什么要选择R  
ggplot绘图语法  
色彩原理  
基本图表类型  
单变量分析  
双变量分析

# ggplot绘图语法

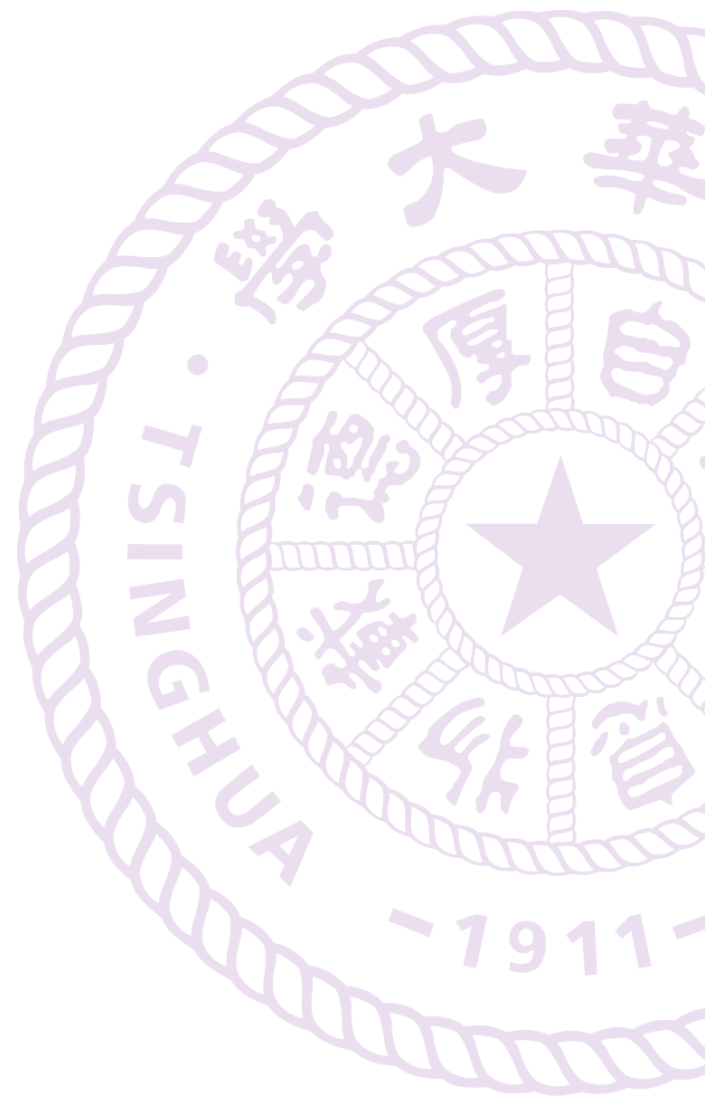
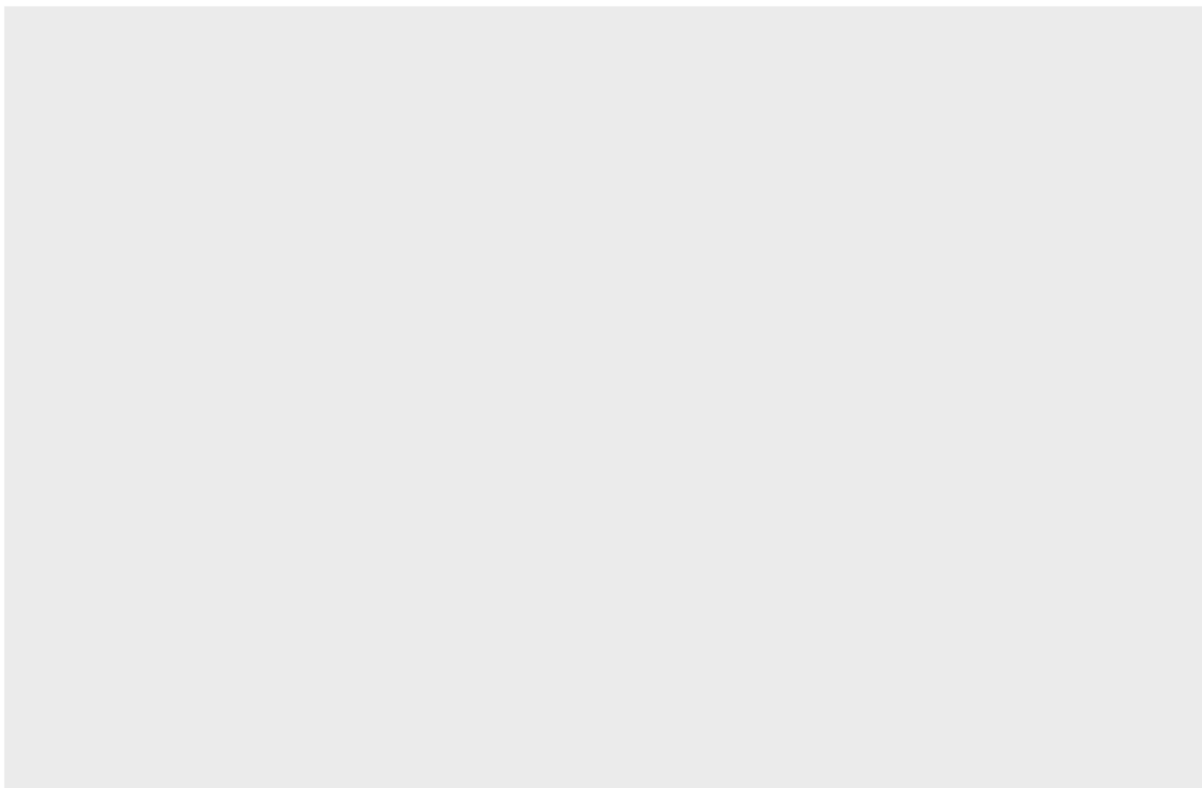
	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb	type
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	Automatic	4	4	Automobiles
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	Automatic	4	4	Automobiles
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	Automatic	4	1	Automobiles
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	Manual	3	1	Automobiles
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	Manual	3	2	Automobiles
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	Manual	3	1	Automobiles
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	Manual	3	4	Automobiles
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	Manual	4	2	Automobiles
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	Manual	4	2	Automobiles
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	Manual	4	4	Automobiles
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	Manual	4	4	Automobiles
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	Manual	3	3	Automobiles
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	Manual	3	3	Automobiles
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	Manual	3	3	Automobiles

为什么要选择R  
ggplot绘图语法  
色彩原理  
基本图表类型  
单变量分析  
双变量分析

# ggplot绘图语法

## Data

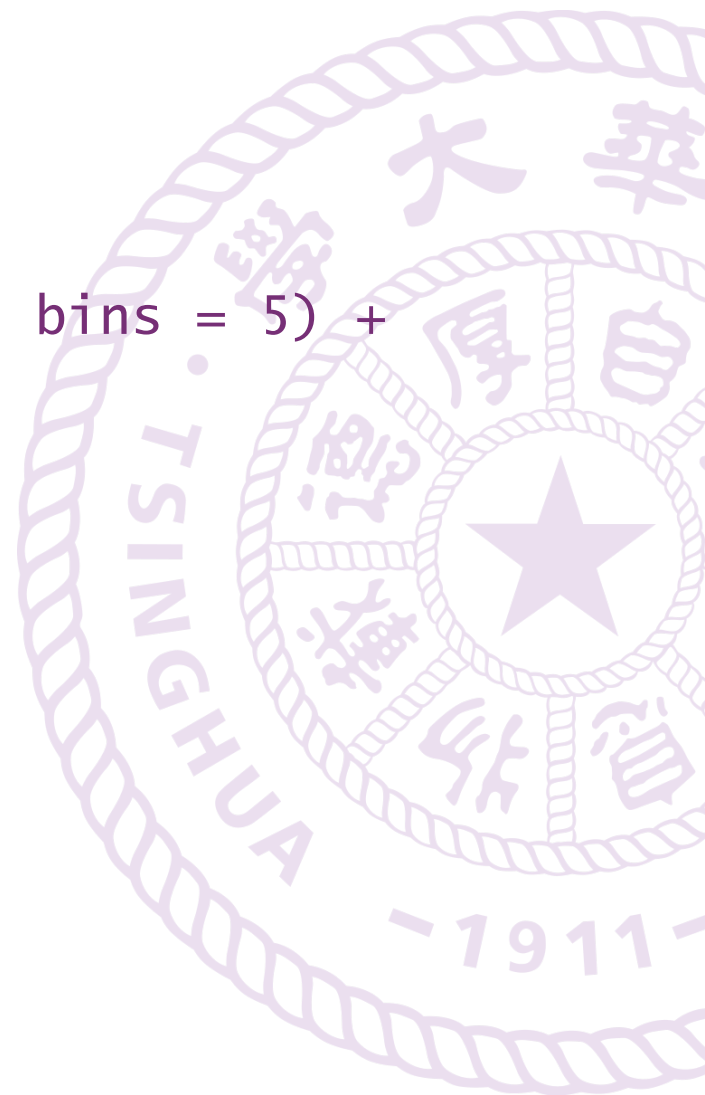
```
ggplot(data = mtcars)
```



# ggplot绘图语法

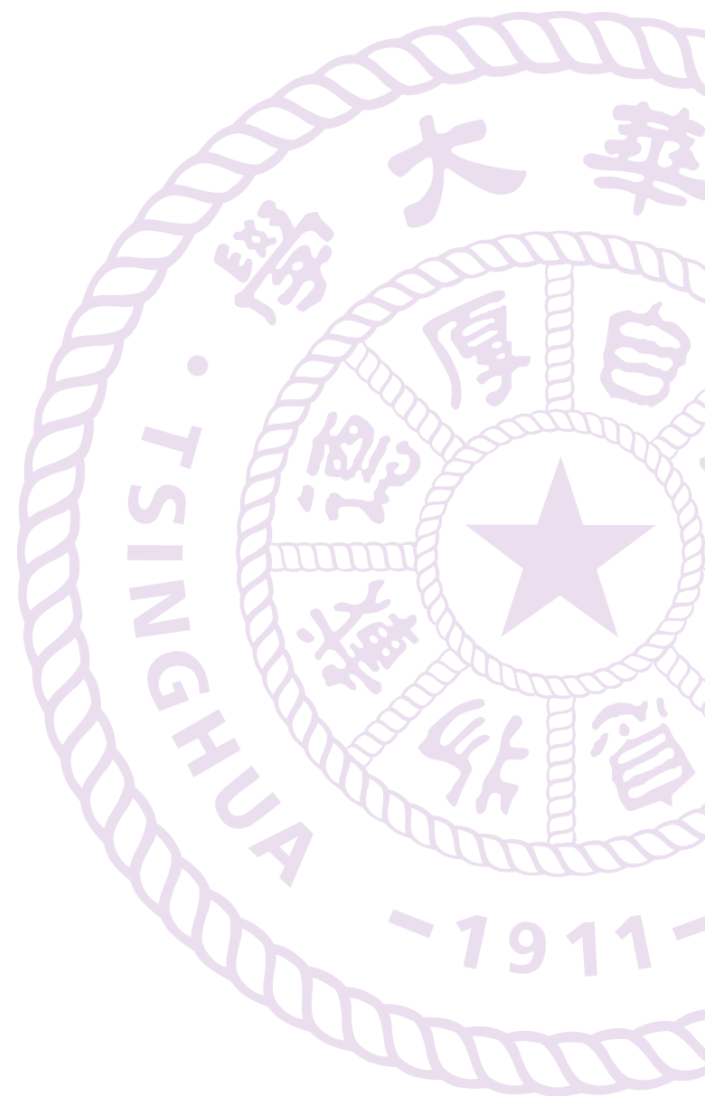
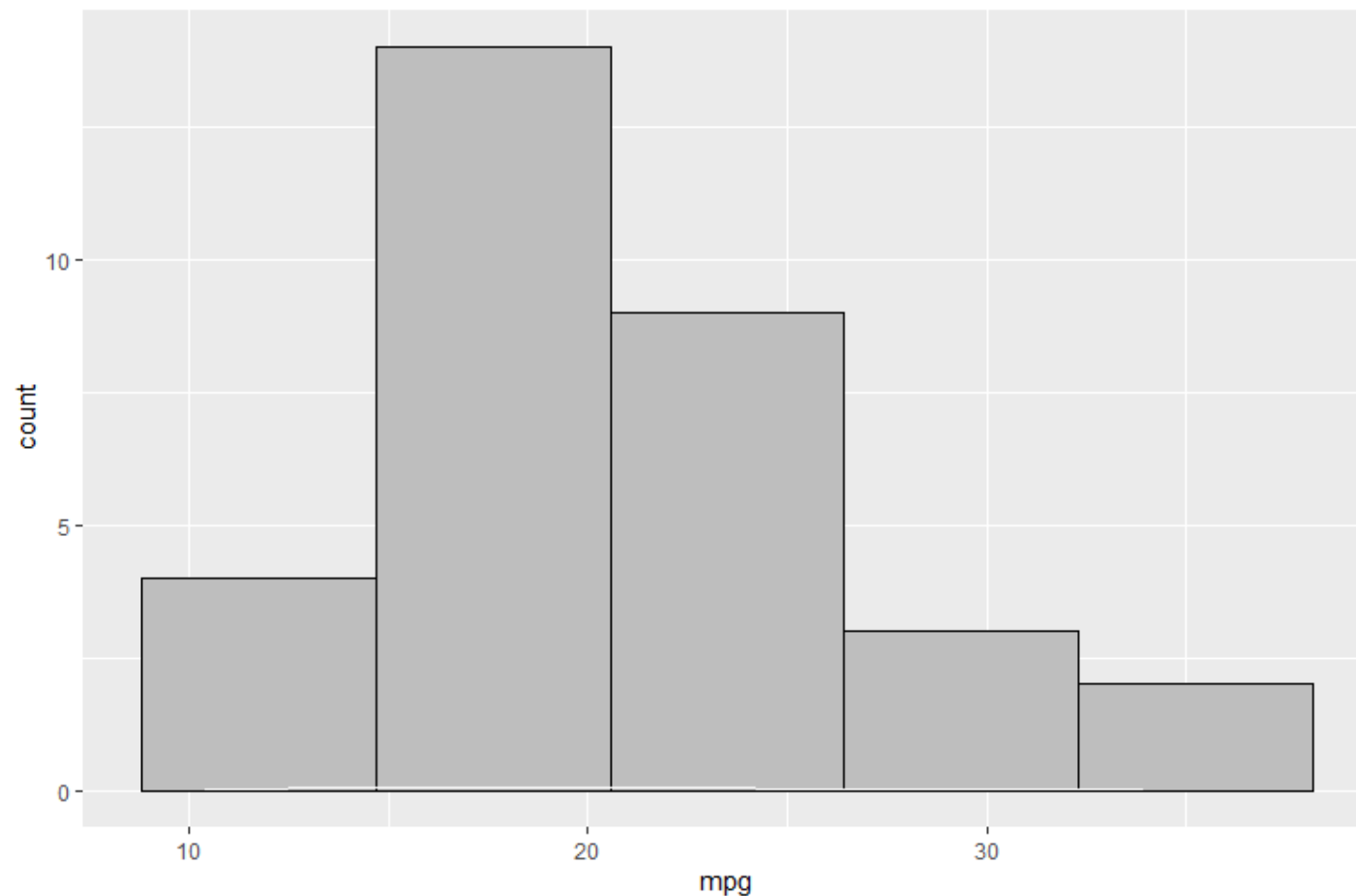
## Data + Aesthetics

```
ggplot(data = mtcars, aes(x = mpg)) +  
  geom_histogram(fill = 'grey', color = 'black', bins = 5) +  
  geom_density(color = 'red')
```



# ggplot绘图语法

## Data + Aesthetics

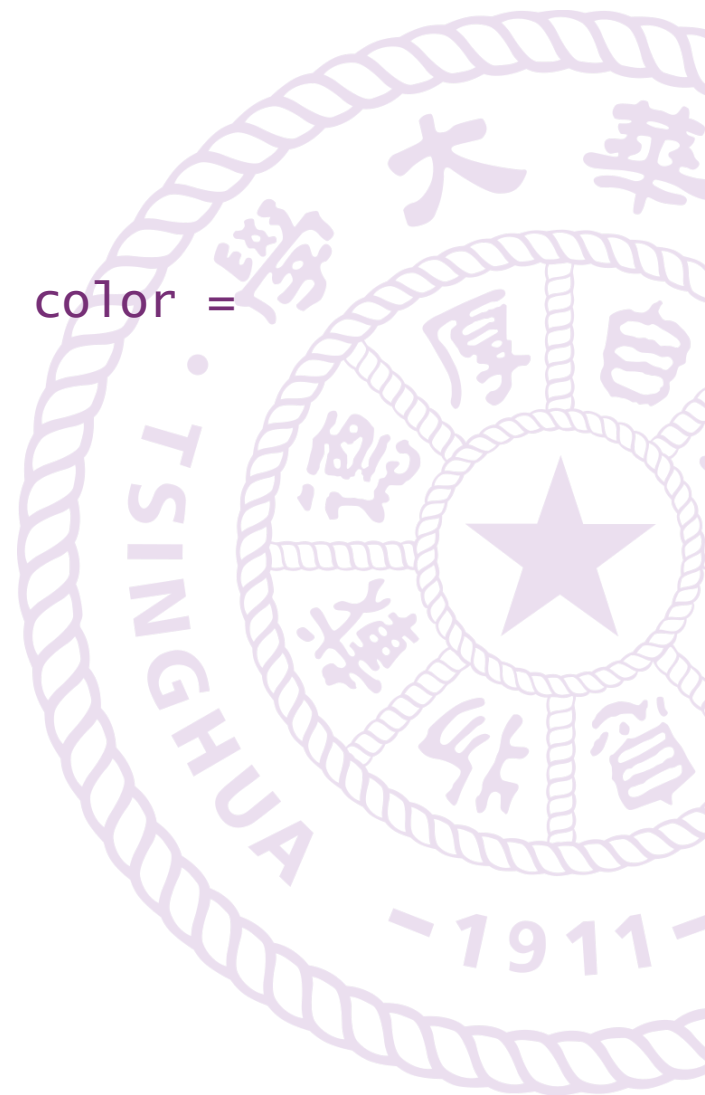


# ggplot绘图语法

## Data + Aesthetics + Geometrics

```
ggplot(data = mtcars, aes(x = mpg)) +  
  geom_histogram(aes(y = ..density..), fill = 'grey', color =  
'black', bins = 5) +  
  geom_density(color = 'red', lwd = 1)
```

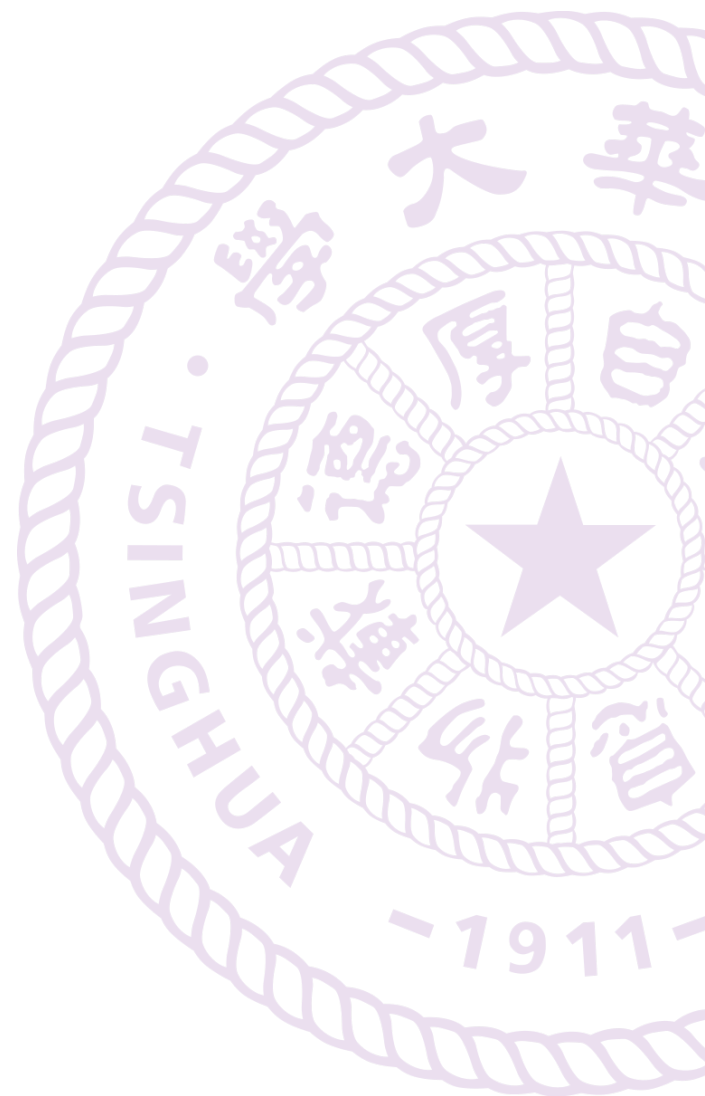
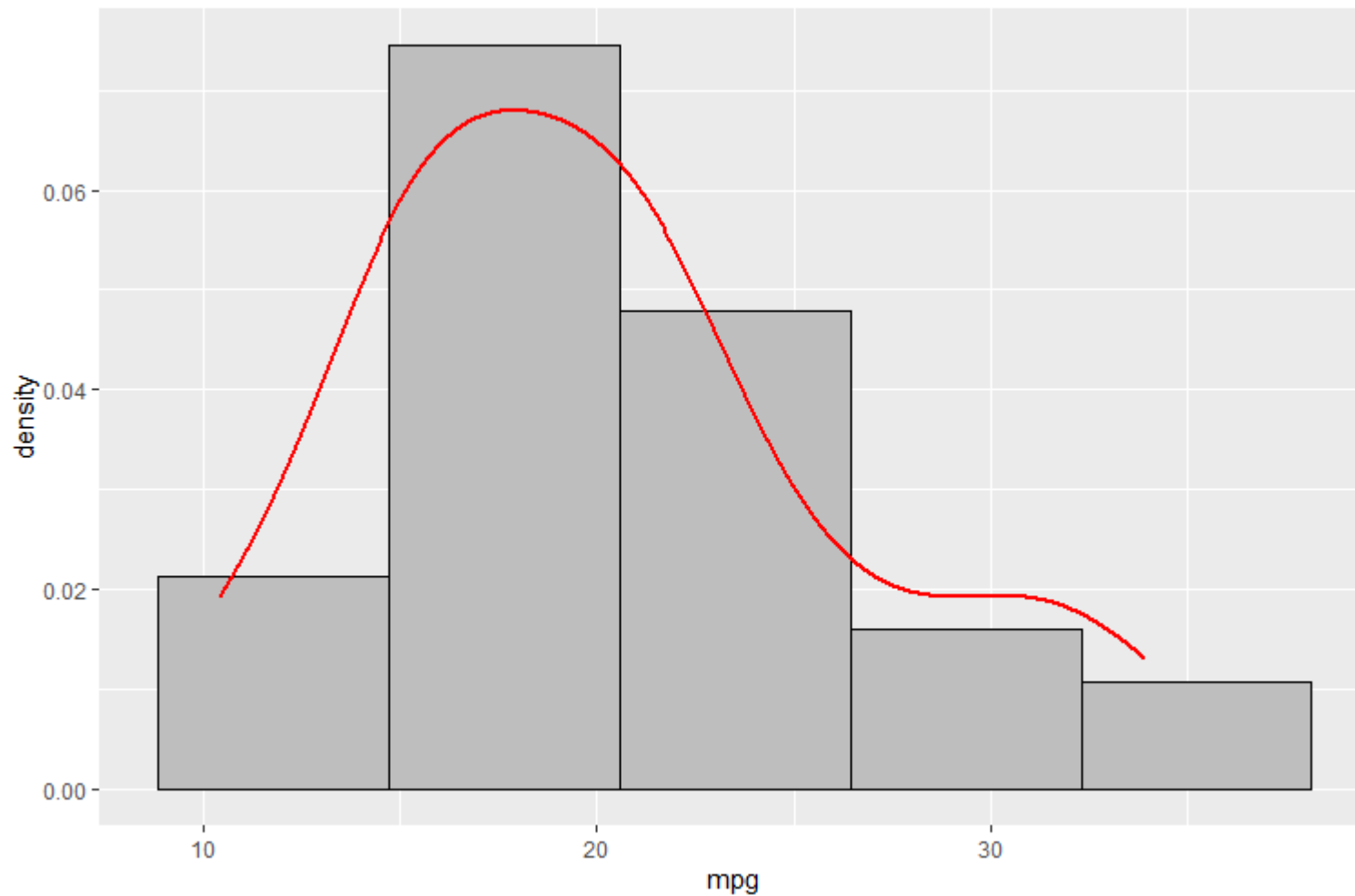
设置比例





# ggplot绘图语法

## Data + Aesthetics + Geometrics



# ggplot绘图语法

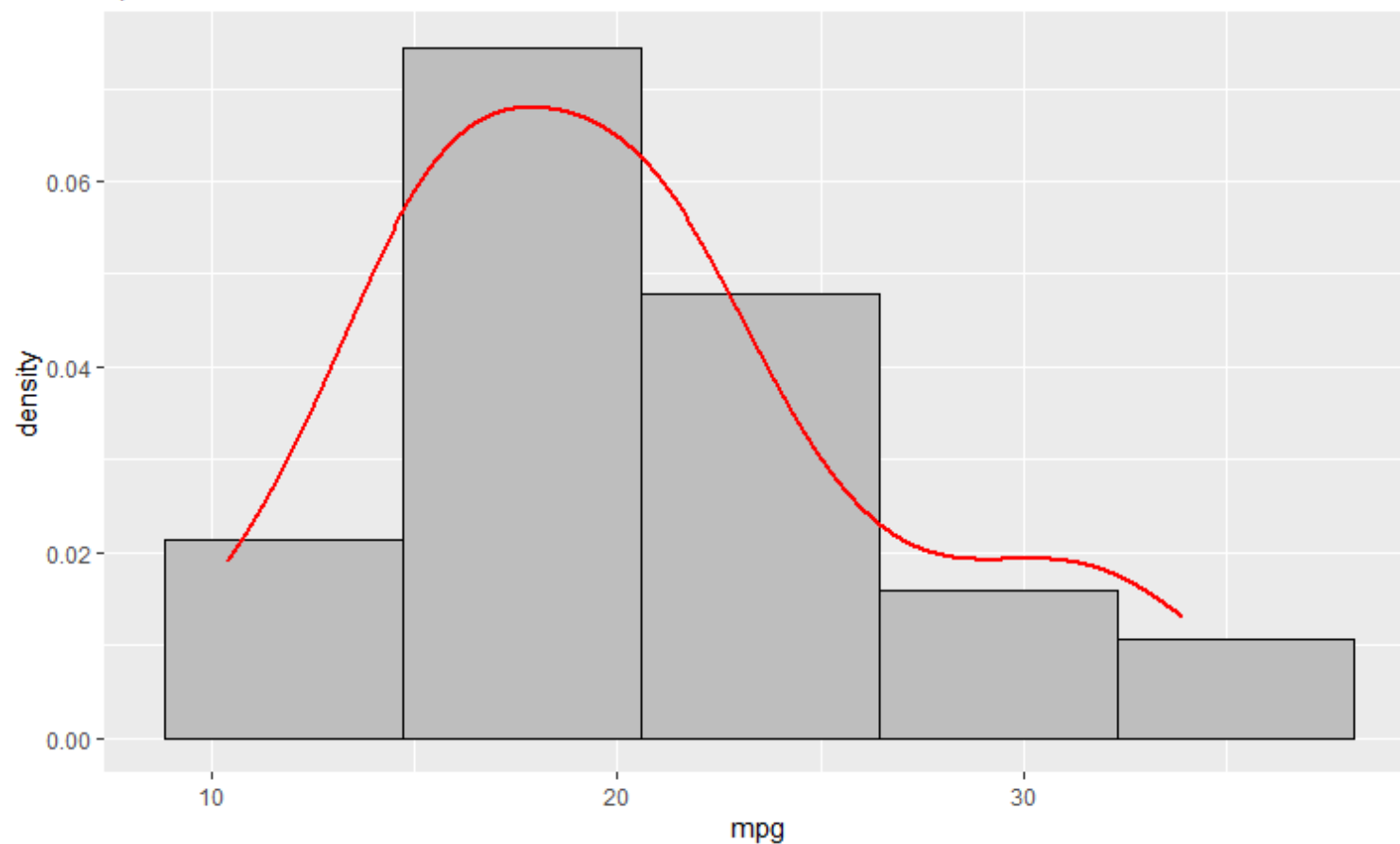
## Data + Aesthetics + Geometrics + Others

```
ggplot(data = mtcars, aes(x = mpg)) +  
  geom_histogram(aes(y = ..density..), fill = 'grey', color = 'black',  
bins = 5) +  
  geom_density(color = 'red', lwd = 1) +  
  ggtitle('Histogram with Imposed Density Curve \n (Miles Per Gallon)')
```

# ggplot绘图语法

## Data + Aesthetics + Geometrics + Others

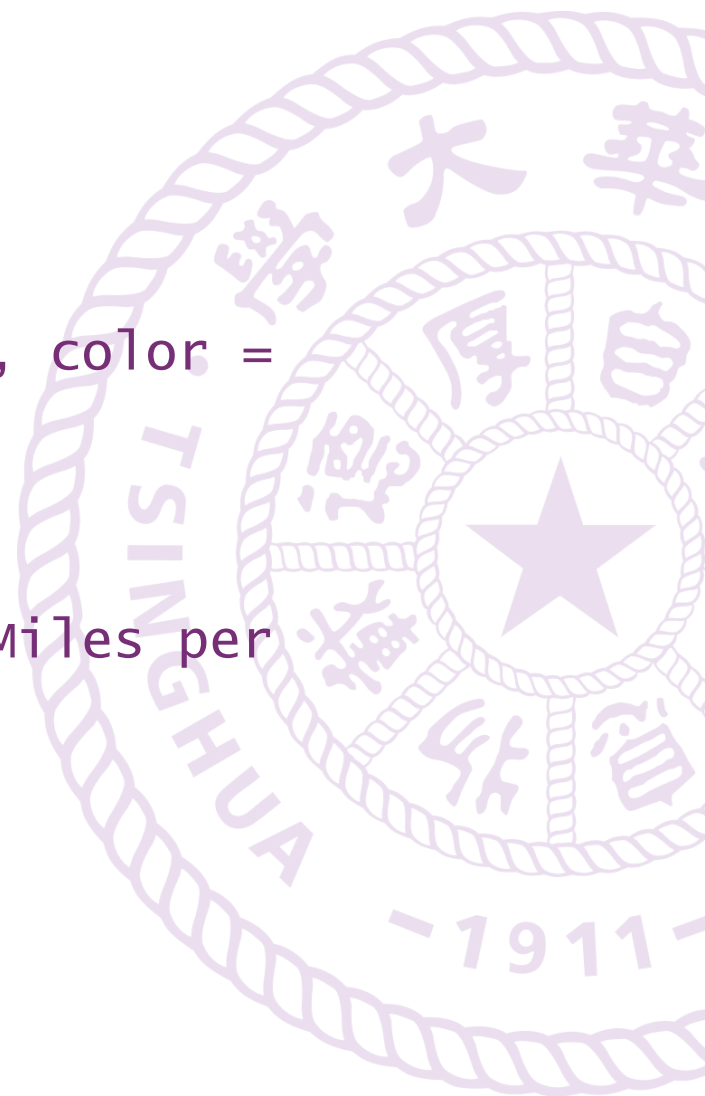
Histogram with Imposed Density Curve  
(Miles Per Gallon)



# ggplot绘图语法

## Data + Aesthetics + Geometrics + Others

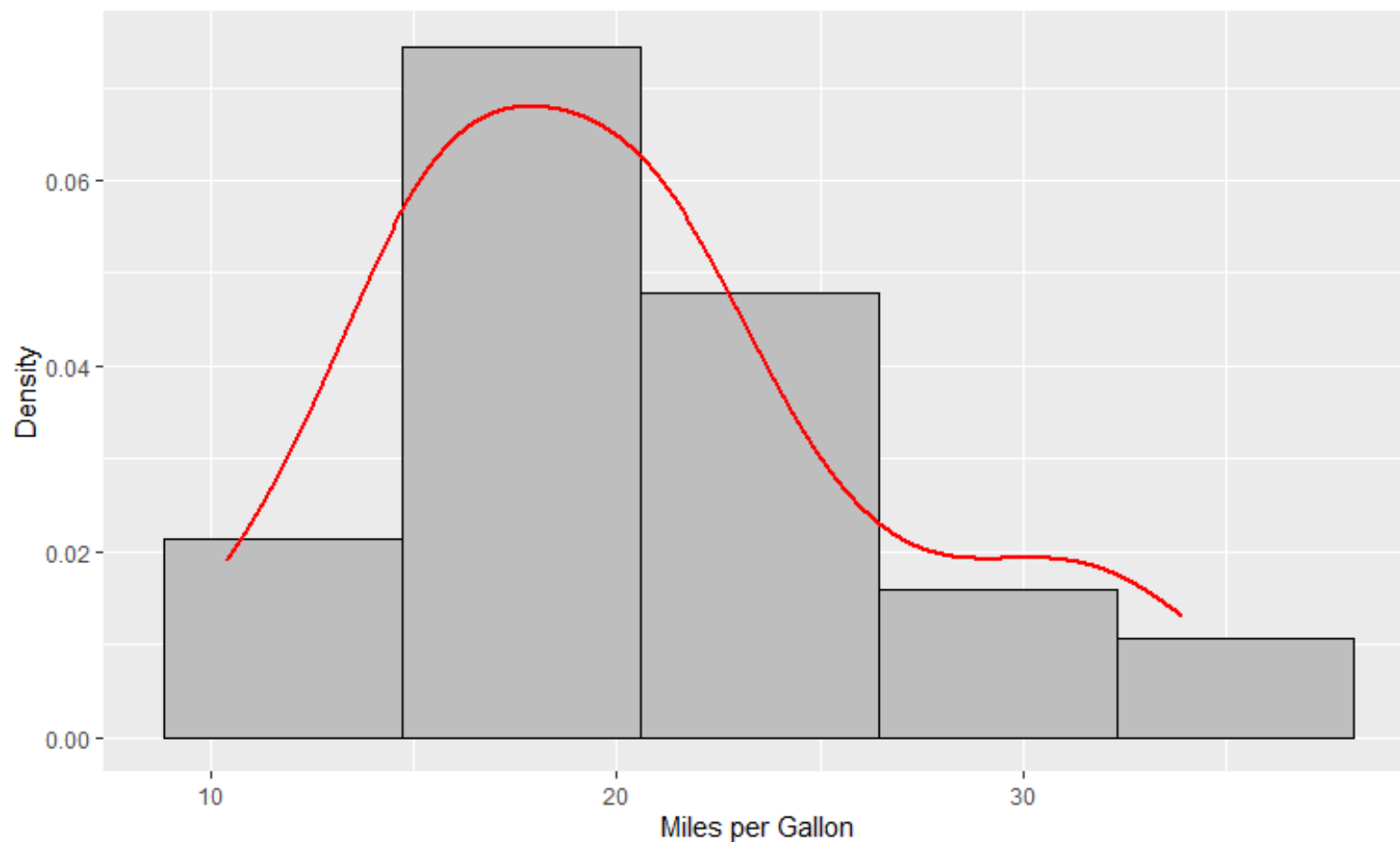
```
ggplot(data = mtcars, aes(x = mpg)) +  
  geom_histogram(aes(y = ..density..), fill = 'grey', color =  
'black', bins = 5) +  
  geom_density(color = 'red', lwd = 1) +  
  ggtitle('Histogram with Imposed Density Curve \n (Miles per  
Gallon)') +  
  ylab('Density') +  
  xlab('Miles per Gallon') +  
  theme(plot.title = element_text(hjust = 0.5))
```



# ggplot绘图语法

## Data + Aesthetics + Geometrics + Others

Histogram with Imposed Density Curve  
(Miles per Gallon)



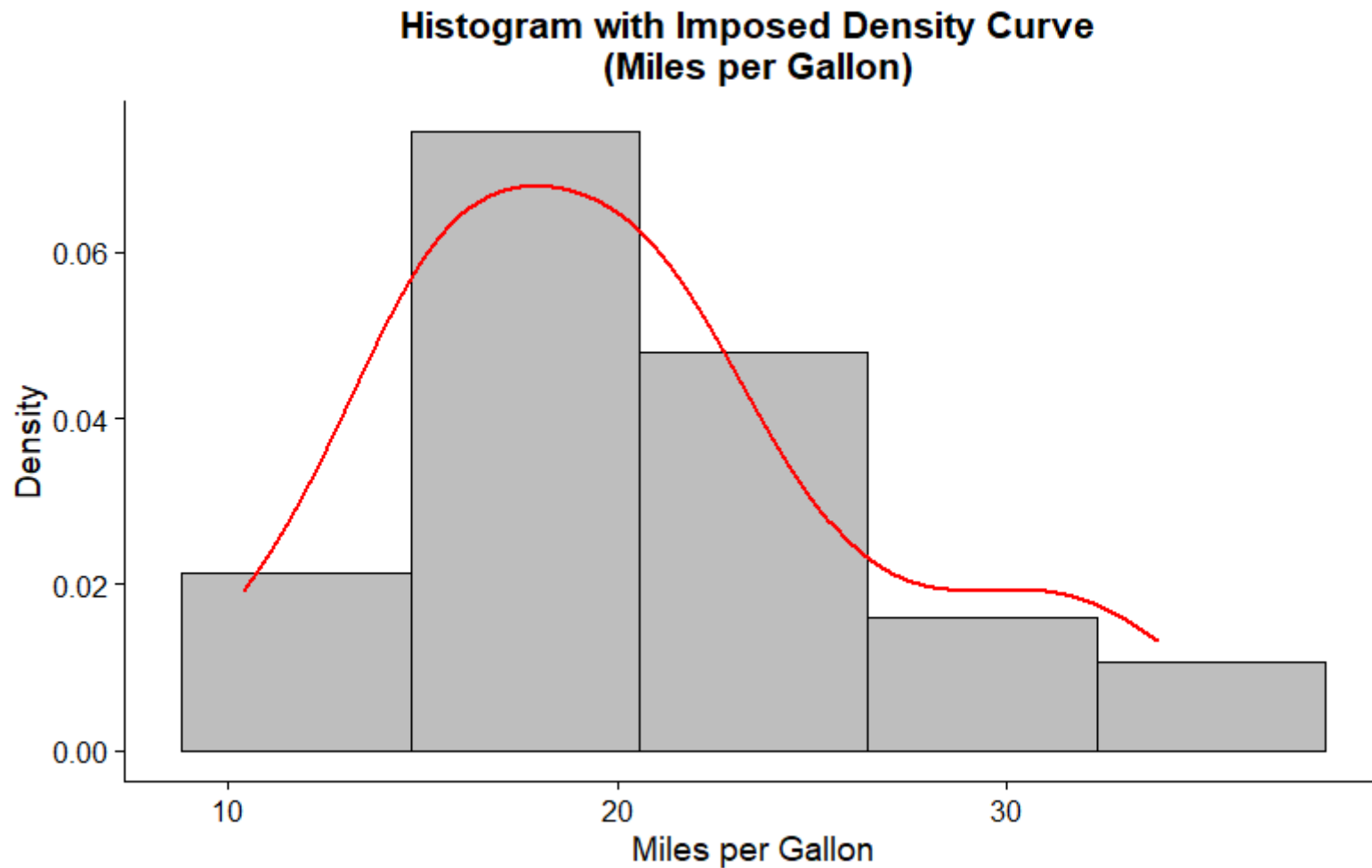
# ggplot绘图语法

**主题：感觉图中的灰格子有亿点点丑**

```
ggplot(data = mtcars, aes(x = mpg)) +  
  geom_histogram(aes(y = ..density..), fill = 'grey', color =  
'black', bins = 5) +  
  geom_density(color = 'red', lwd = 1) +  
  ggtitle('Histogram with Imposed Density Curve \n (Miles per  
Gallon)') +  
  ylab('Density') +  
  xlab('Miles per Gallon') +  
  theme_cowplot() +  
  theme(plot.title = element_text(hjust = 0.5))
```

# ggplot绘图语法

主题：感觉图中的灰格子主题有亿点点丑



# ggplot绘图语法

保存：那，怎么把我的图存下来？

- 存在变量中

```
fig1 <- ggplot(data = mtcars, aes(x = mpg)) +  
  geom_histogram(aes(y = ..density..), fill = 'grey', color =  
'black', bins = 5) +  
  geom_density(color = 'red', lwd = 1) +  
  ggtitle('Histogram with Imposed Density Curve \n (Miles per Gallon)') +  
  ylab('Density') +  
  xlab('Miles per Gallon') +  
  theme_cowplot() +  
  theme(plot.title = element_text(hjust = 0.5))
```



# ggplot绘图语法

保存：那，怎么把我的图存下来？

- 存在电脑中

```
ggsave("C:/Users/mi/SynologyDrive/Rclub/myfirstplot.pdf", fig1)
```

- 直接用鼠标点也可

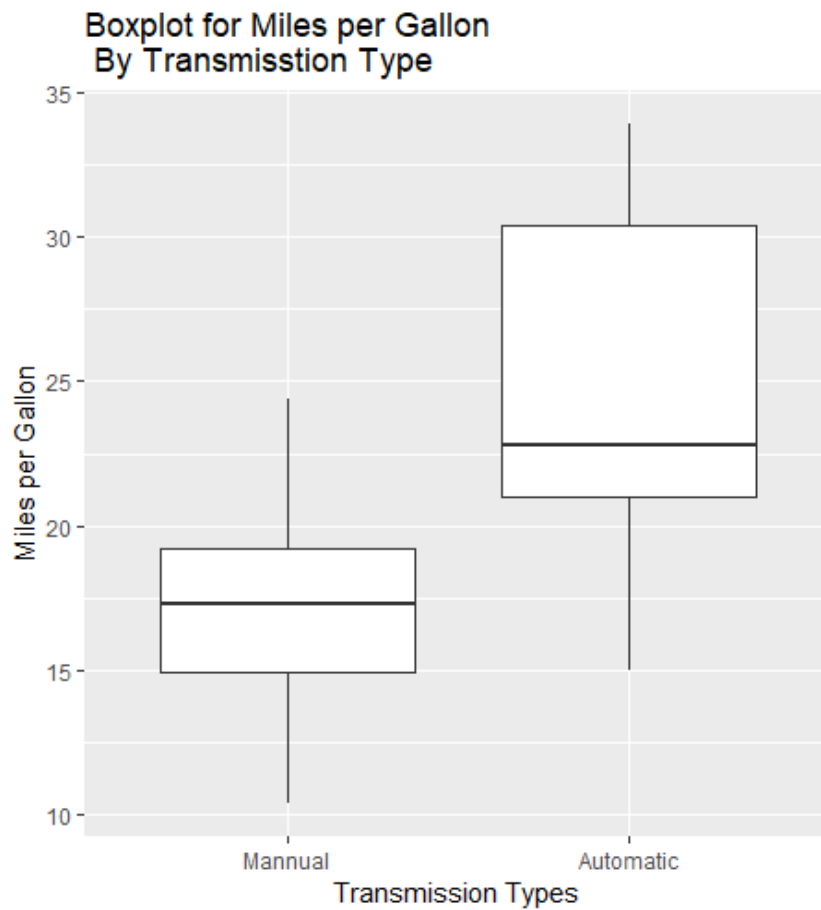
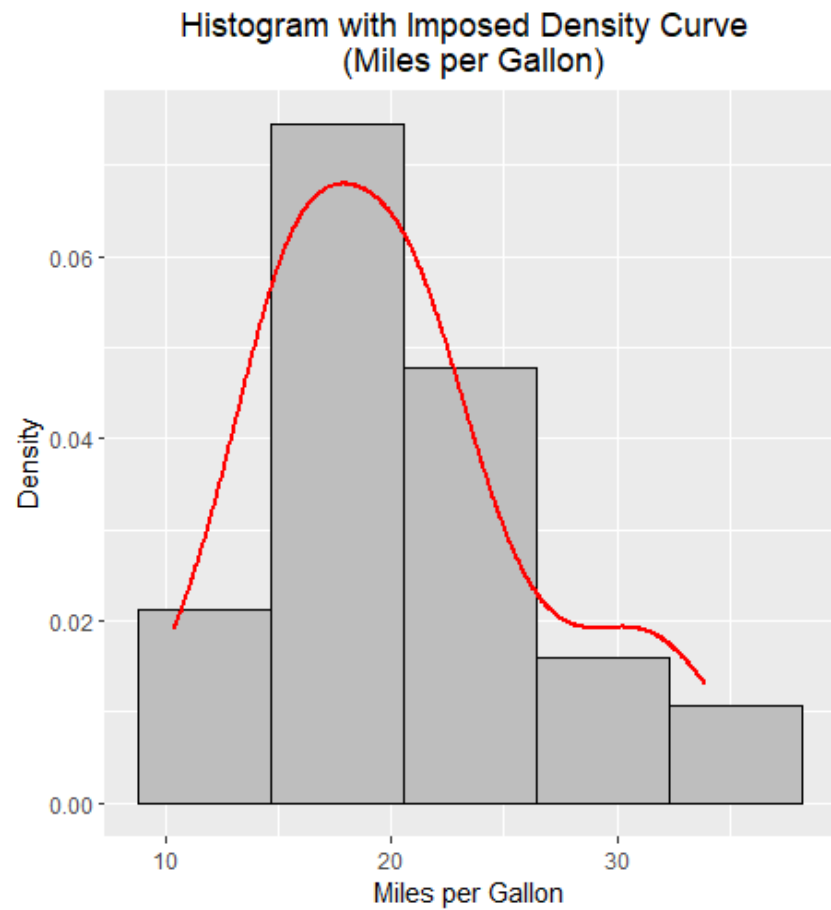
# ggplot绘图语法

**分格：放在同一个张图里之横着放**

```
mtcars$am <- factor(mtcars$am, labels = c('Manual', 'Automatic'))  
fig2 <- ggplot(data = mtcars, aes(y = mpg, x = am)) +  
  geom_boxplot() +  
  ggtitle('Boxplot for Miles per Gallon \n By Transmisstion Type') +  
  xlab('Transmission Types') +  
  ylab('Miles per Gallon')  
  
plot_grid(fig1, fig2, nrow = 1)
```

# ggplot绘图语法

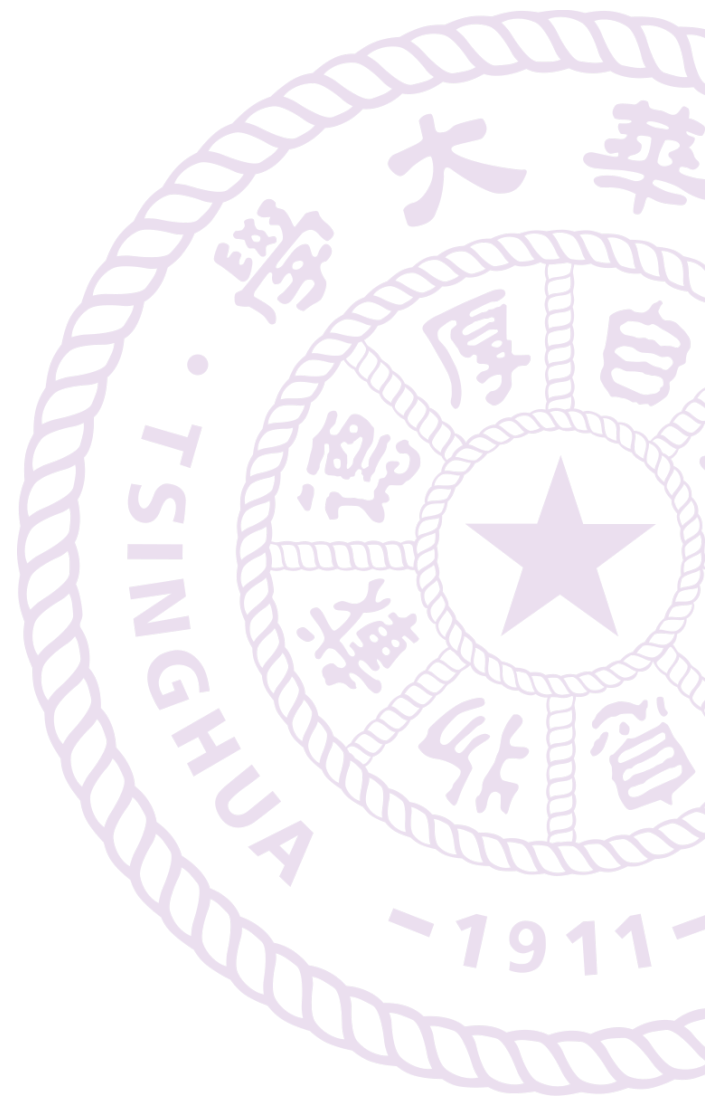
分格：放在同一个张图里之横着放



# ggplot绘图语法

**分格：放在同一个张图里之竖着放**

```
plot_grid(fig1, fig2, nrow = 2)
```

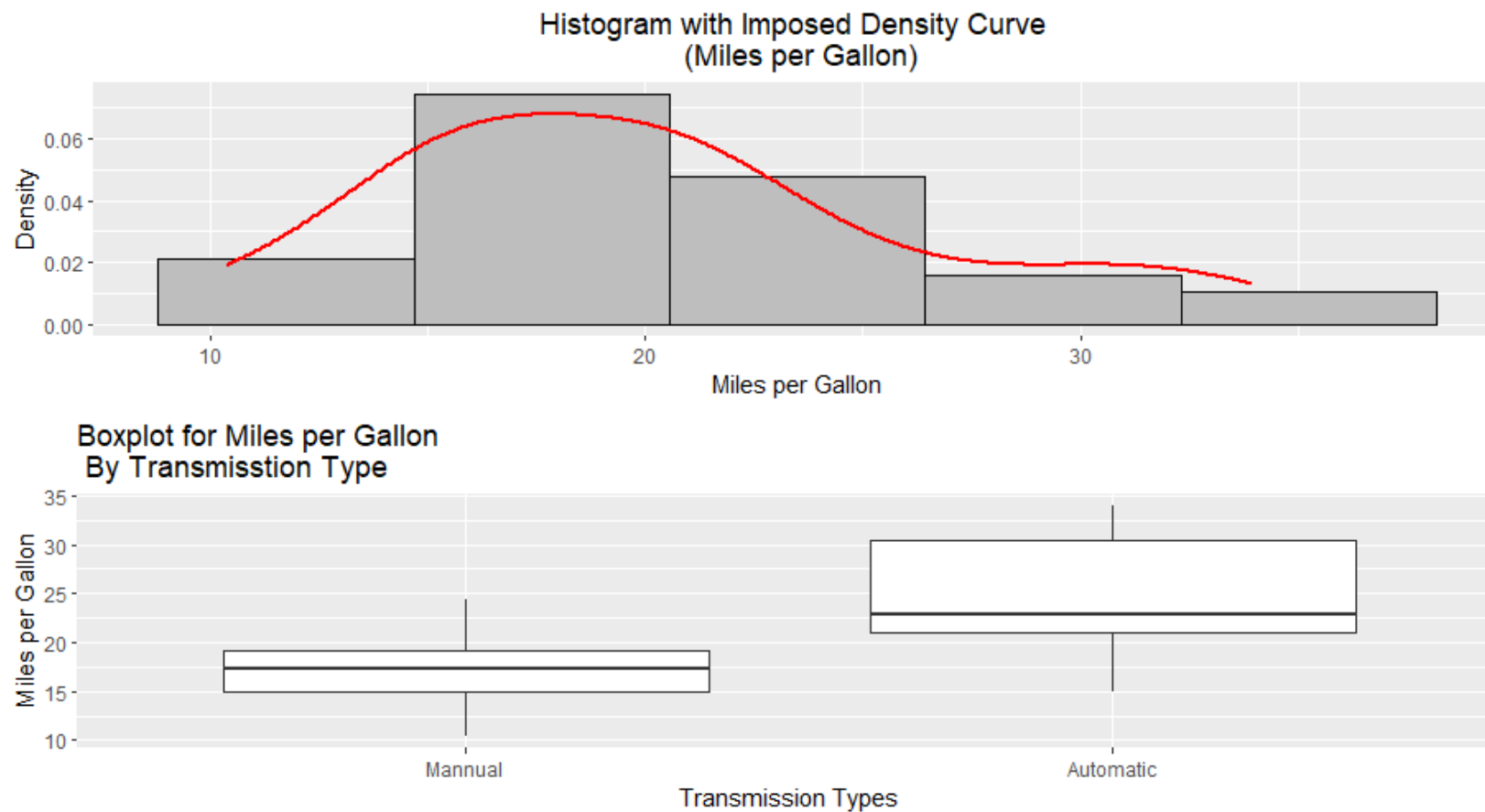


为什么要选择R  
ggplot绘图语法  
色彩原理  
基本图表类型  
单变量分析  
双变量分析

# ggplot绘图语法

**分格：放在同一个张图里之竖着放**

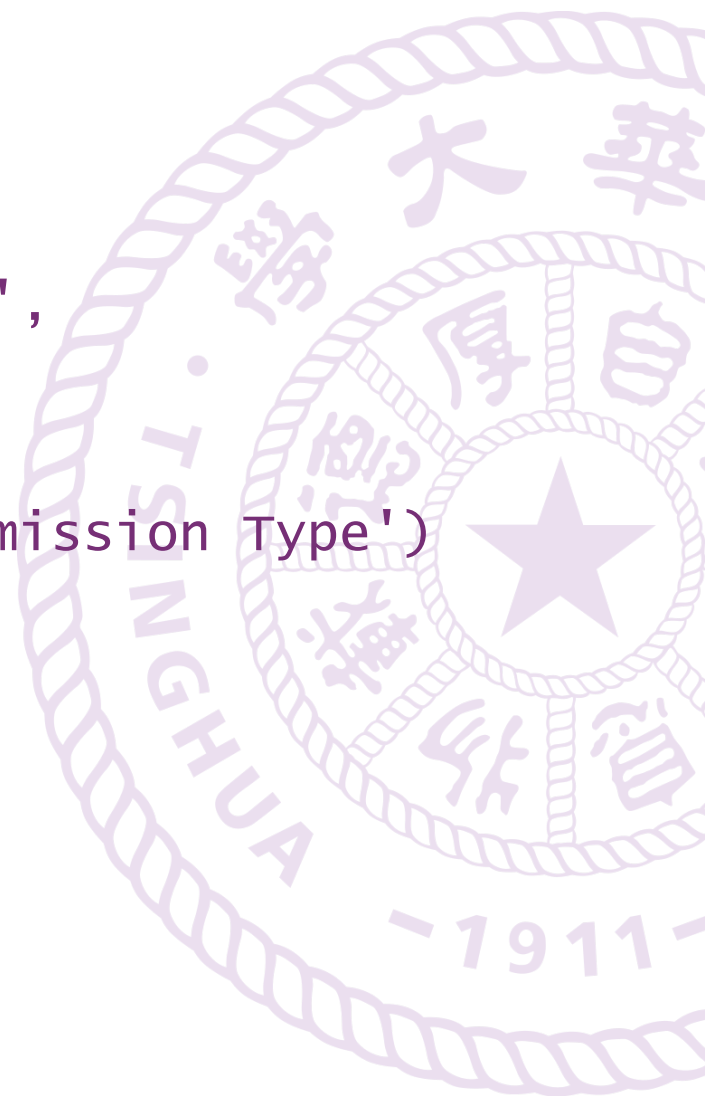
`plot_grid(fig1, fig2, nrow = 2)`



# ggplot绘图语法

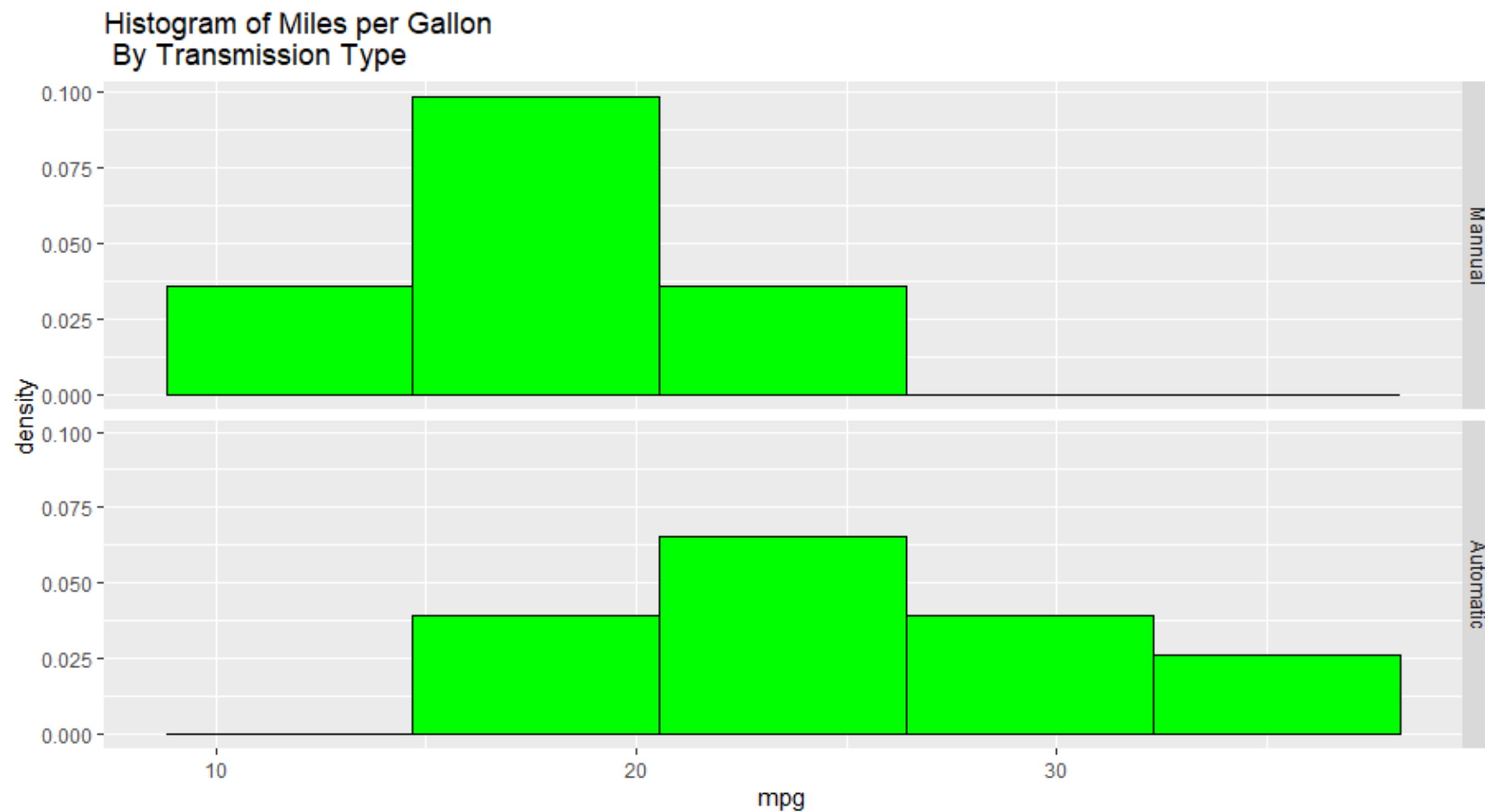
**分面：还想放在同一坐标轴里**

```
fig3 <- ggplot(data = mtcars, aes(x = mpg)) +  
  geom_histogram(aes(y = ..density..), fill = 'Green',  
                 color = 'black', bins = 5) +  
  facet_grid(am ~ .) +  
  ggtitle('Histogram of Miles per Gallon \n By Transmission Type')  
fig3
```



# ggplot绘图语法

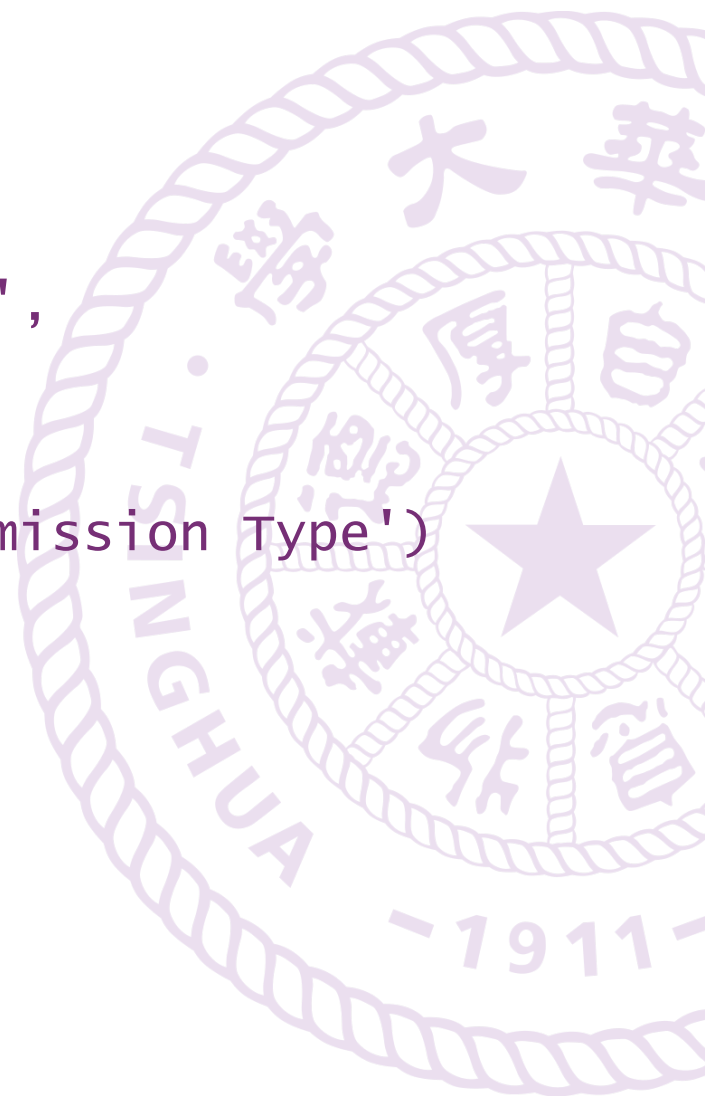
## 分面：还想放在同一坐标轴里



# ggplot绘图语法

**分面：还想放在同一坐标轴里**

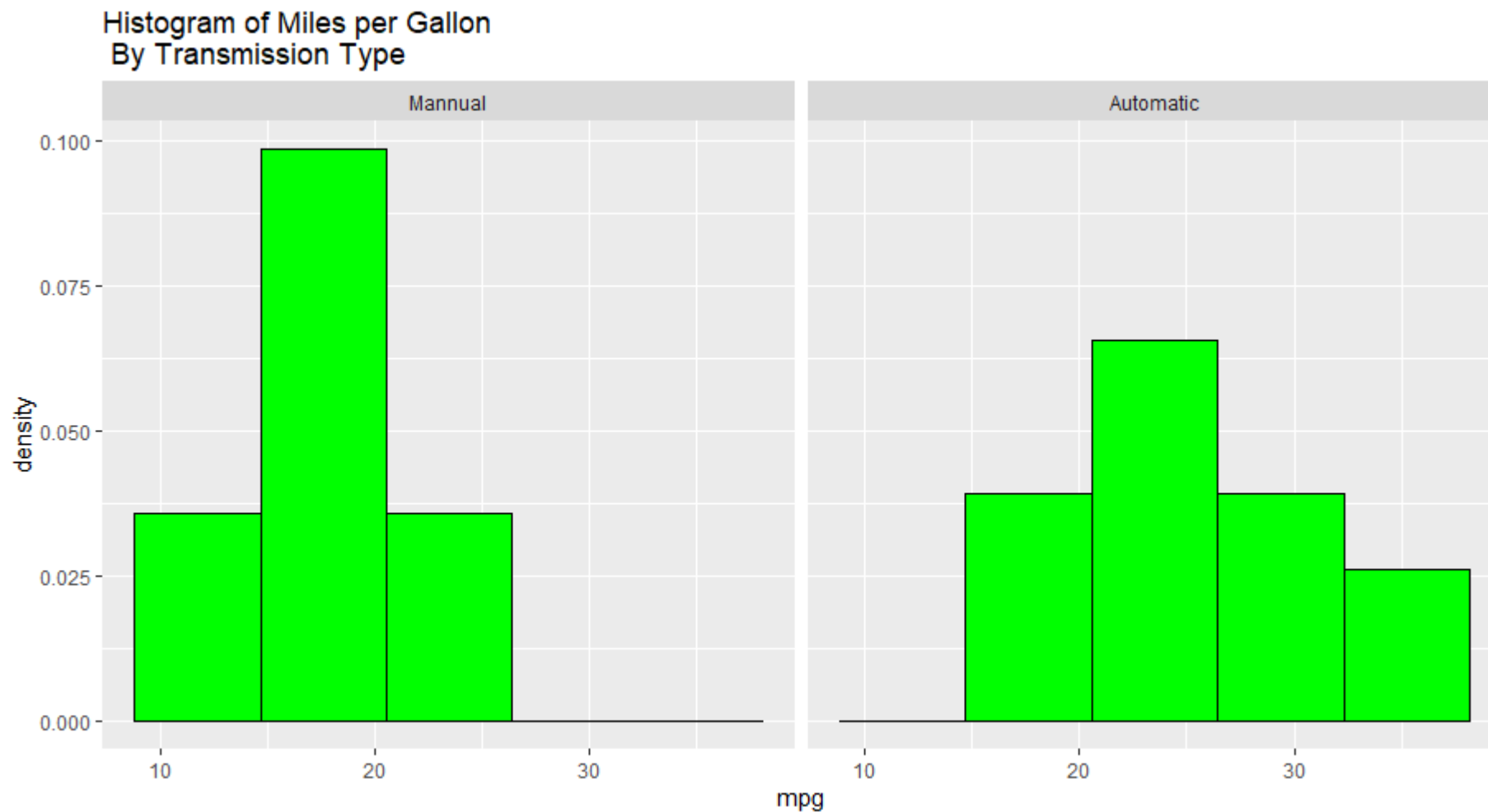
```
fig4 <- ggplot(data = mtcars, aes(x = mpg)) +  
  geom_histogram(aes(y = ..density..), fill = 'Green',  
                 color = 'black', bins = 5)+  
  facet_grid(. ~ am) +  
  ggtitle('Histogram of Miles per Gallon \n By Transmission Type')  
fig4
```





# ggplot绘图语法

分面：还想放在同一坐标轴里





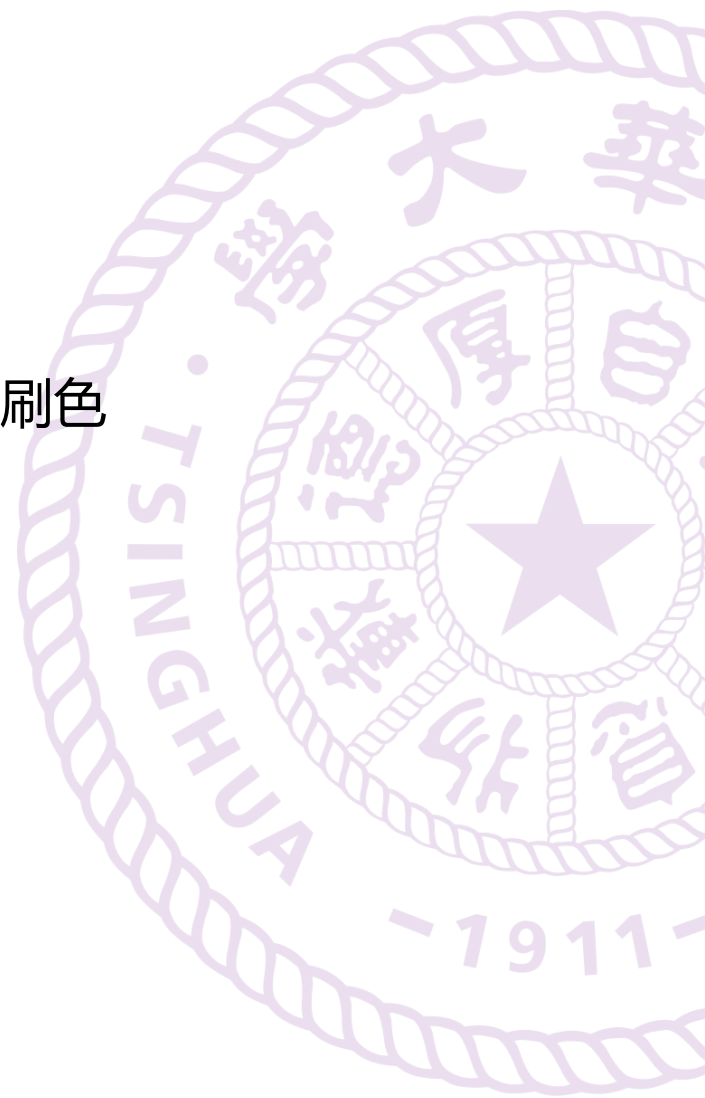
## 课程结构

- I. 为什么要选择R
- II. ggplot绘图语法
- III. 色彩原理**
- IV. 基本的图表类型
- V. 单变量分析
- VI. 双变量分析

# 色彩原理

## 色彩的表达

- RGB: 红、绿、蓝。显示色
- CMYK: 青 (Cyan)、洋红 (Magenta)、黄 (Yellow)、黑。印刷色
- HSL: 色相 (Hue)、饱和度 (Saturation)、亮度 (Lightness)



# 色彩原理

## 色彩的搭配

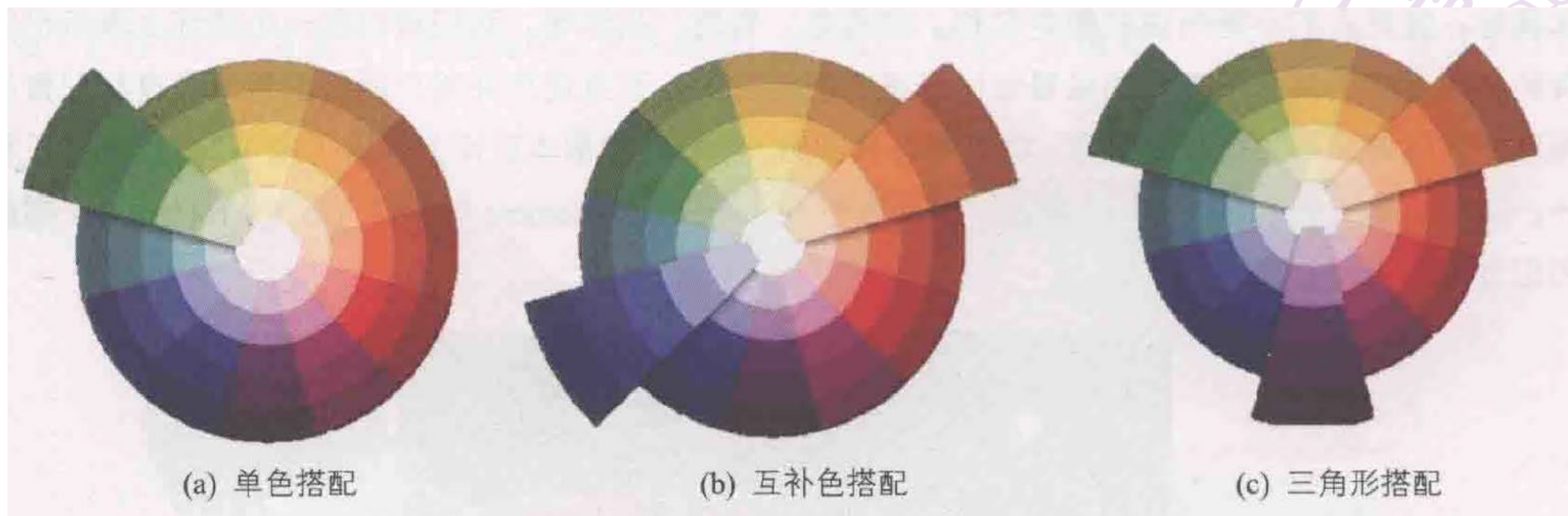
(1) 单色(monochromatic) 搭配色相由暗、中、明三种色调组成的单色。单色搭配并没有形成颜色的层次，但形成了明暗的层次。这种搭配在设计中应用时，效果永远不错。

(2) 互补色(complement) 搭配如果颜色方案只包括两种颜色，就会选择色环上对立的2个颜色（在色轮上直线相对的两种颜色称为补色，比如红色和绿色）。互补色搭配在正式的设计中比较少见，主要由于它色彩之间强烈对比所产生的特殊性和不稳定

(3) 三角形(triad) 搭配如果颜色方案只包括三种颜色，那么就会以 $120^\circ$  的间限选择3个颜色。三角形搭配是一种能使得画面生动的搭配方式，即使使用了低饱和度的色彩也是如此

# 色彩原理

## 色彩的搭配





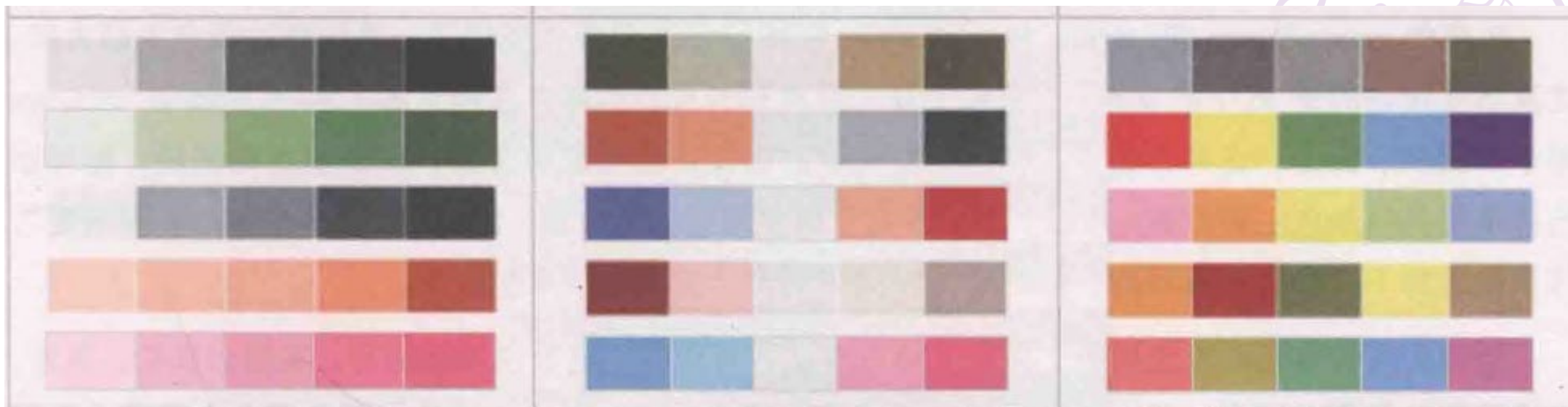
# 色彩原理

## 色彩如何运用于学术图表之中——Colorbrewer

单色系 ( sequential )	双色渐变系 ( dsiverging )	多色系 ( qualitative )
色相基本相同，饱和度呈单调递增的变化。有序数据一般从大到小排列，对应的颜色亮度也逐步增加。小数值通常使用较亮的颜色表示，而大数值通常使用较暗的颜色表示。单色系颜色搭配方案中可能存在颜色的色相不同，但它的主要特征还是颜色从亮到暗的亮度变化。比如地区的人口密度等通常使用单色系搭配方案	两个不同的色系使用于不同的两类情况，如正值与负值。双色渐变系搭配方案主要强调数据基于一个关键中间数值 (midpoint) 的级数分布情况。把关键的中间数值作为中间点，使用一个较亮的颜色表示，然后两端逐步变化到两个不同色相的颜色。比如基于某疾病平均死亡率的分布情况，就可以使用双色渐变系搭配方案	数据为非数值情况，不同色系的颜色用于表示不同类别，尤其是使用色相最轻或最暗的颜色强调关键的类别。多色系颜色搭配方案使用不同色相值的颜色，表示不同类别或数值的差异。这些颜色的亮度不一定要完全相等，但是要基本差不多。多色系还包括圆形分布的多色系
$[-A, 0]$ , $[0, A]$ , 或者 $[A, B]$	$[A, 0, B]$ , 或者 $[A, C, B]$ (C 为 mean, medium 等)	类别，特征， 时间类的周期性数据

# 色彩原理

## 色彩如何运用于学术图表之中——Colorbrewer



# 色彩原理

## 操作化：怎么在ggplot中调色

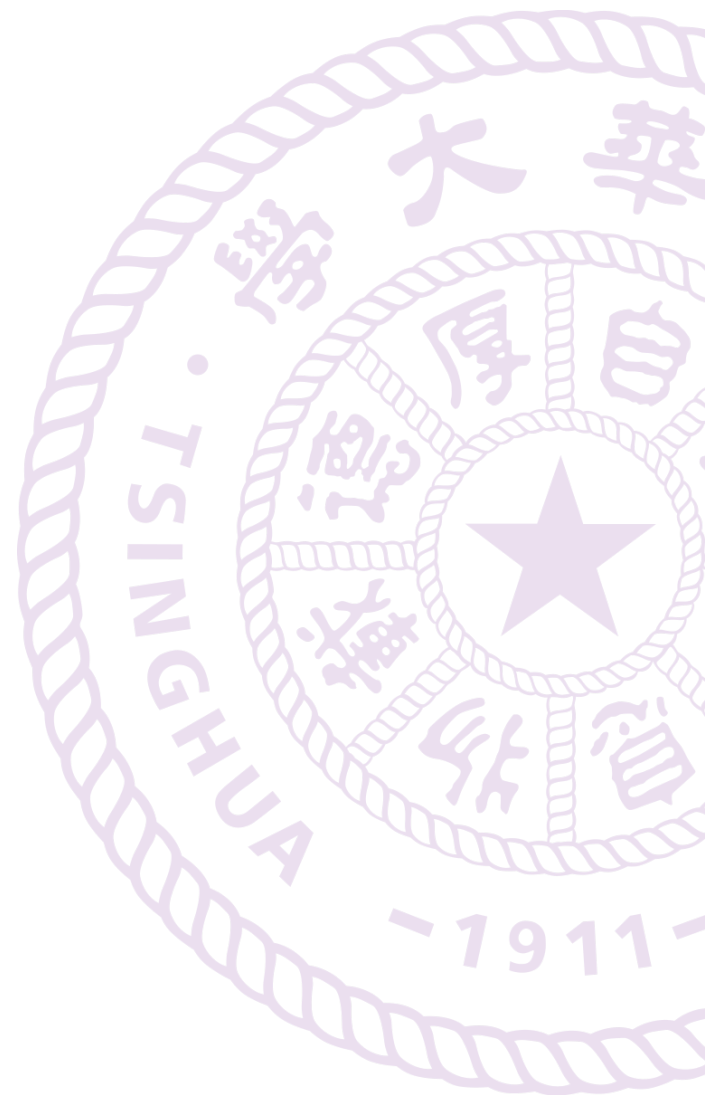
```
fig8 <- ggplot(data =diamonds) +  
  geom_boxplot(aes(x = color, y = price, fill = color)) +  
  xlab('Color of Diamonds') +  
  ylab('Price (USD)') +  
  ggtitle('Diamond Price by color')
```

fig8

```
fig8 <- fig8 + scale_fill_brewer(palette = 'Accent')  
fig8
```

更多调色方案及其他类型的配色： `?scale_fill_brewer()`

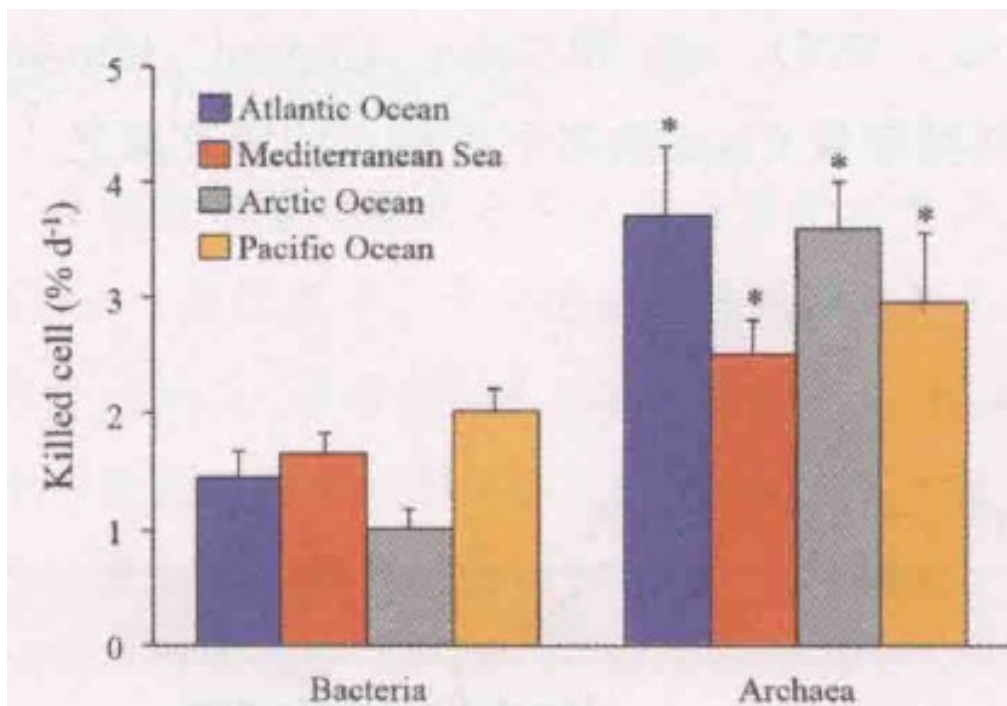
其他包：wesanderson、ggsci、viridis



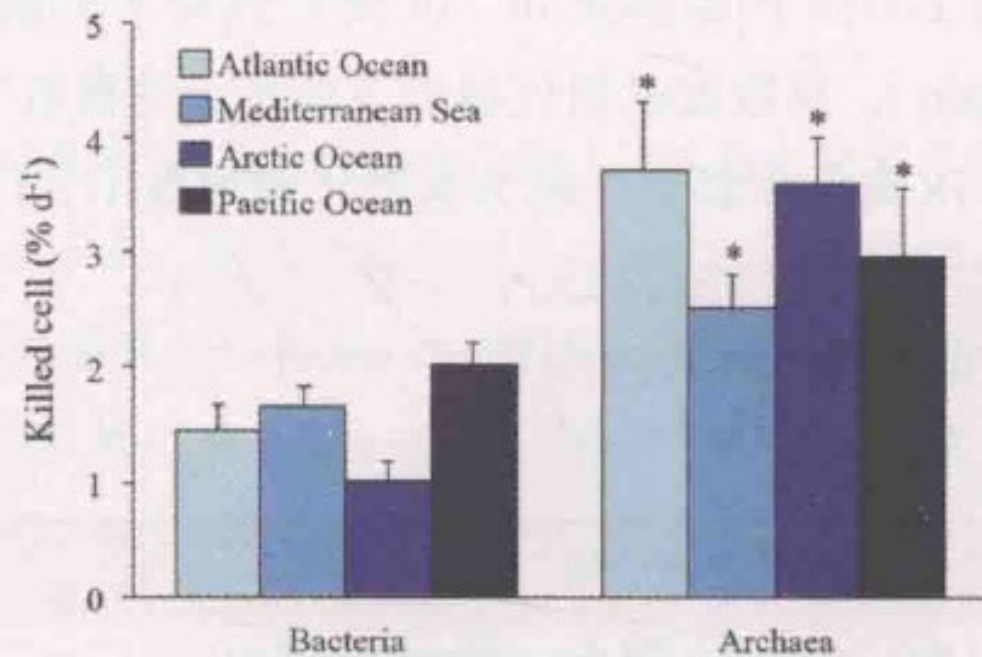


# 色彩原理

## 案例



(a) Excel 多色系颜色方案

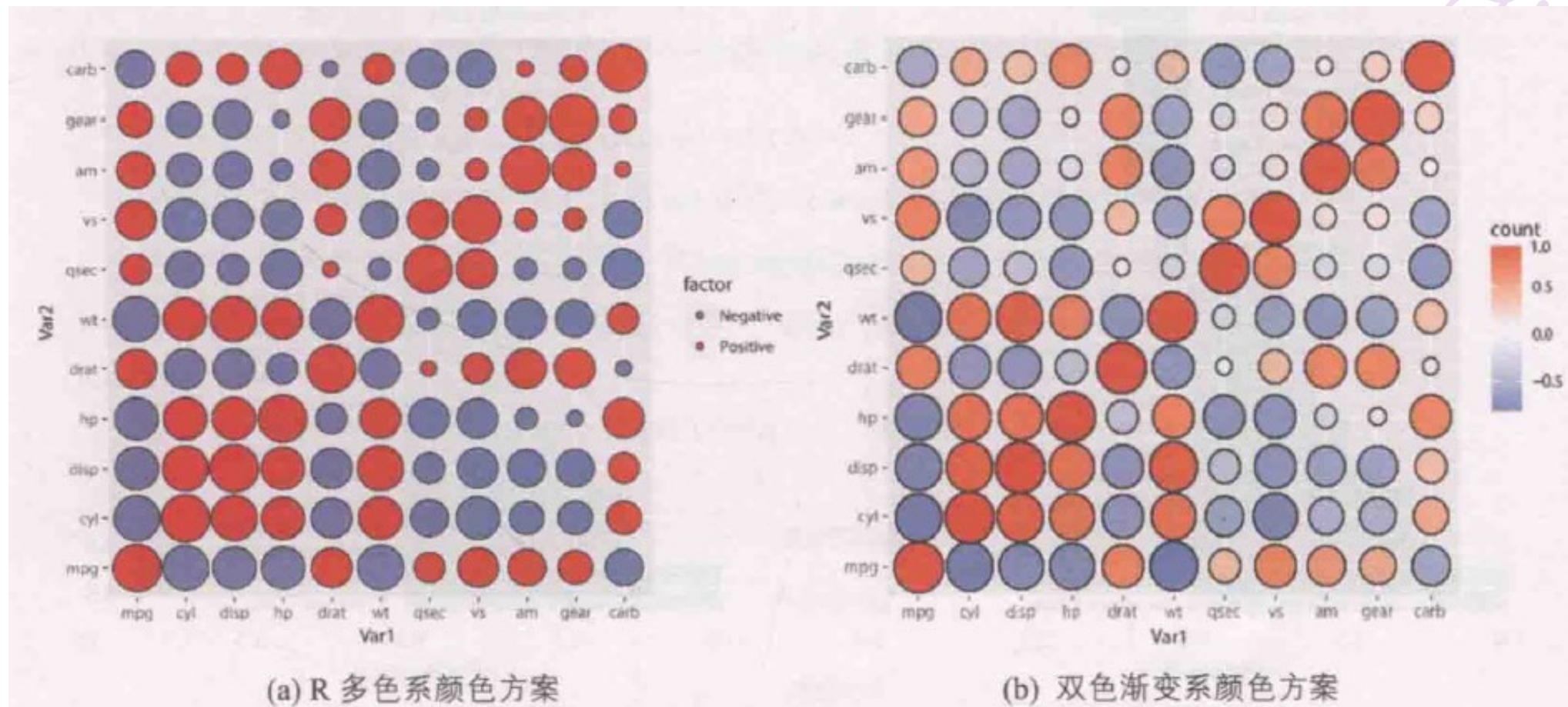


(b) <sup>[8]</sup> 单色系颜色方案

为什么要选择R  
ggplot绘图语法  
色彩原理  
基本图表类型  
单变量分析  
双变量分析

# 色彩原理

## 案例





## 课程结构

- I. 为什么要选择R
- II. ggplot绘图语法
- III. 色彩原理
- IV. 基本的图表类型**
- V. 单变量分析
- VI. 双变量分析

# 基本图表类型

变量数	类型	函数	常用图表类型
1	连续型	geom_histogram() 、 geom_density() 、 geom_dotplot() 、 geom_freqpoly()、 geom_qq()、 geom_area()	统计直方图、核密度估计曲线图
	离散型	geom_bar()	柱形图系列
2	x-连续型 y-连续型	geom_point()、 geom_area()、 geom_line()、 geom_jitter()、 geom_smooth()、 geom_label()、 geom_text()、 geom_bin2d()、 geom_hex()、 geom_density2d()、 geom_map()、 geom_step()、 geom_quantile()、 geom_rug()	散点图系列、面积图系列、折线图系列，包括抖动散点图、平滑曲线图、文本、标签、二维统计直方图、二维核密度估计图、地理空间图表
	x-离散型 y-连续型	geom_boxplot()、 geom_violin()、 geom_dotplot()、 geom_col()	箱形图、小提琴图、点阵图、统计直方图
	x-离散型 y-离散型	geom_count()	二维统计直方图
3	x, y, z- 连续型	geom_contour(), geom_raster(), geom_tile()	等高线图、热力图





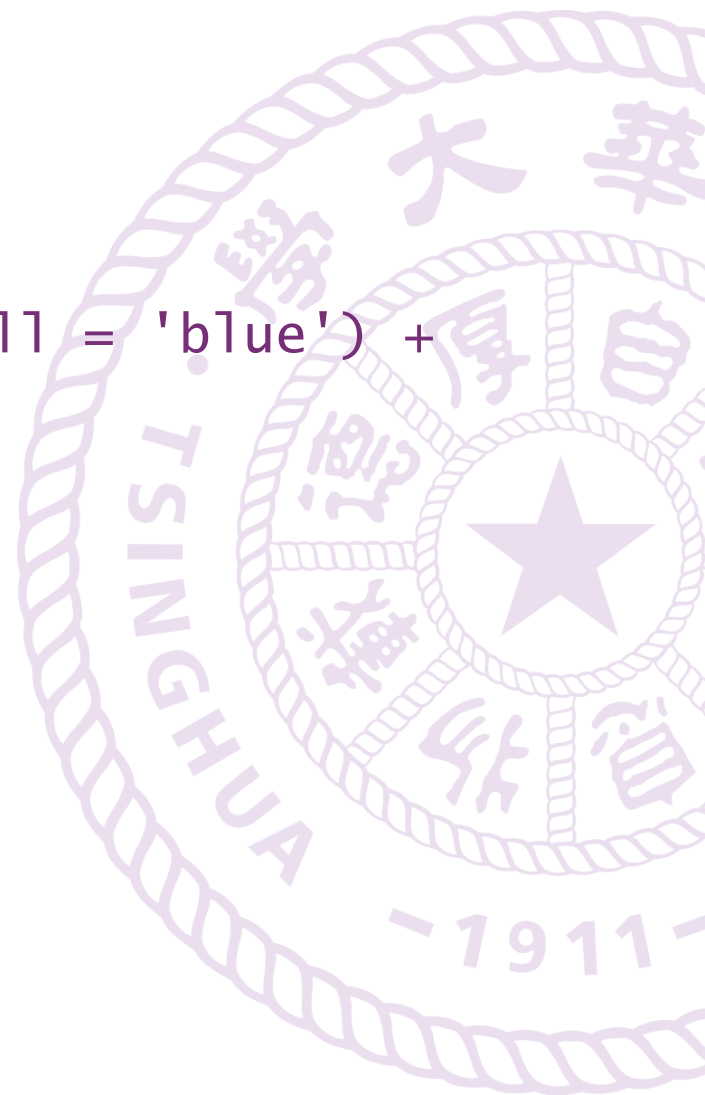
## 课程结构

- I. 为什么要选择R
- II. ggplot绘图语法
- III. 色彩原理
- IV. 基本图表类型
- V. 单变量分析**
- VI. 双变量分析

# 单变量分析

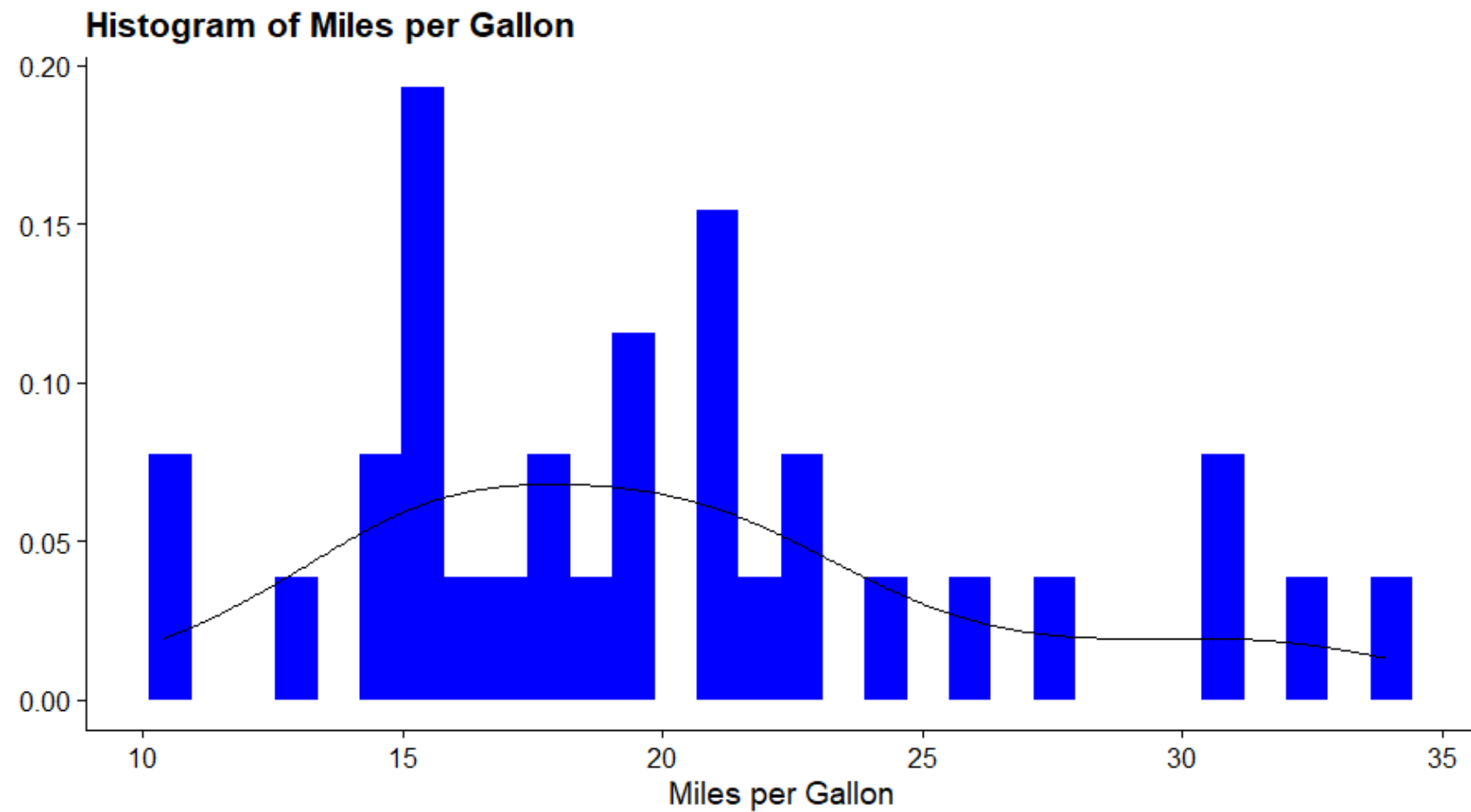
## 直方图 histogram

```
fig1 <- ggplot(data = mtcars, aes(x = mpg)) +  
  geom_histogram(aes(y = ..density..), bins = 30, fill = 'blue') +  
  geom_density(color = 'black') +  
  xlab('Miles per Gallon') +  
  ylab('') +  
  ggtitle('Histogram of Miles per Gallon')  
fig1
```



# 单变量分析

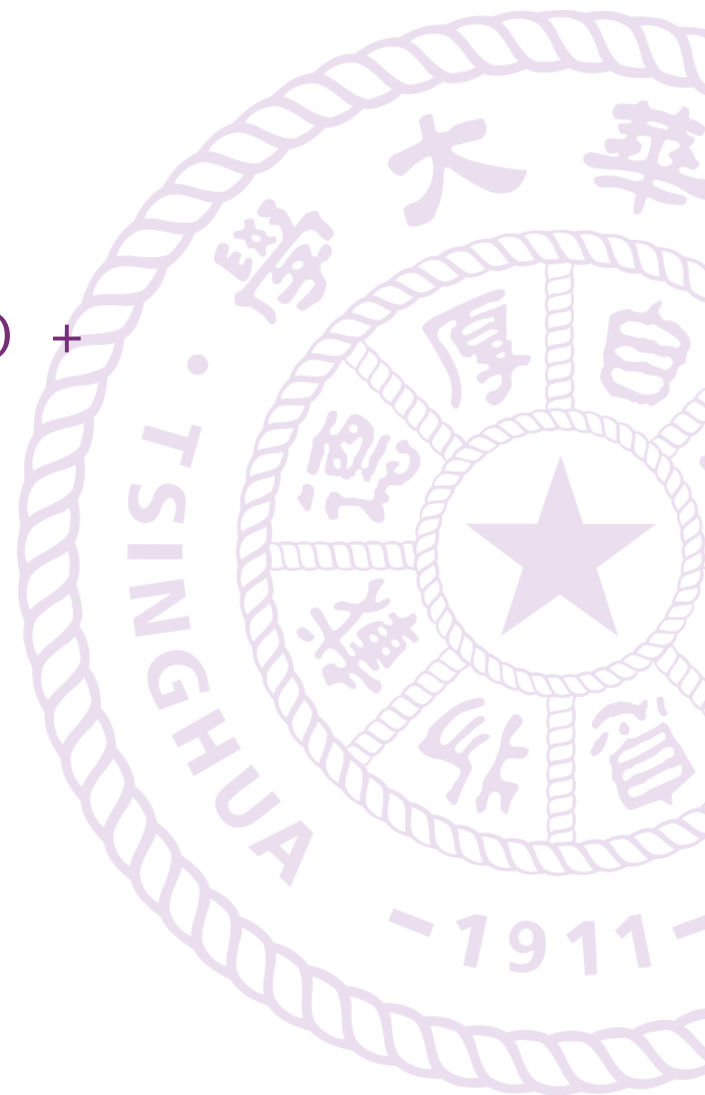
## 直方图 boxplot



# 单变量分析

## 箱型图 boxplot

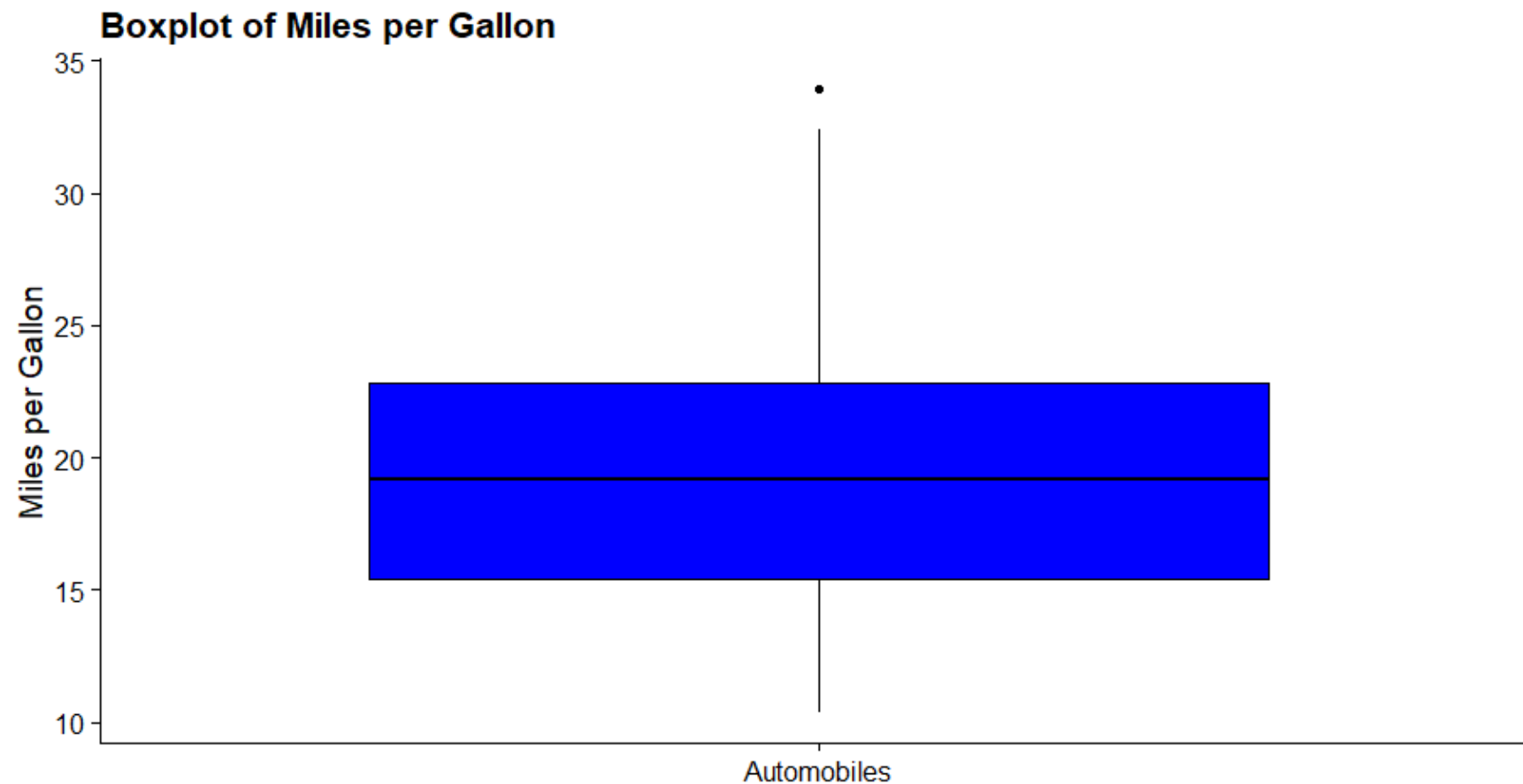
```
mtcars$type <- 'Automobiles'
fig2 <- ggplot(data = mtcars, aes(y = mpg, x = type)) +
  geom_boxplot(fill = 'blue', color = 'black') +
  xlab('') +
  ylab('Miles per Gallon') +
  ggtitle('Boxplot of Miles per Gallon')
fig2
```





# 单变量分析

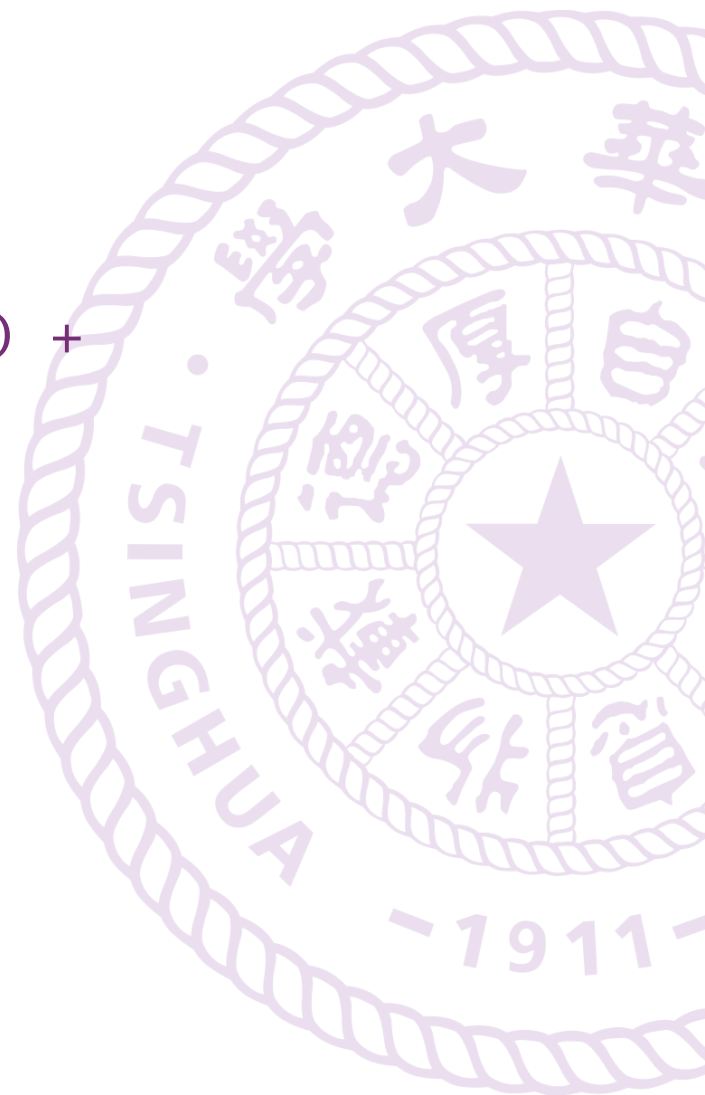
## 箱型图 boxplot



# 单变量分析

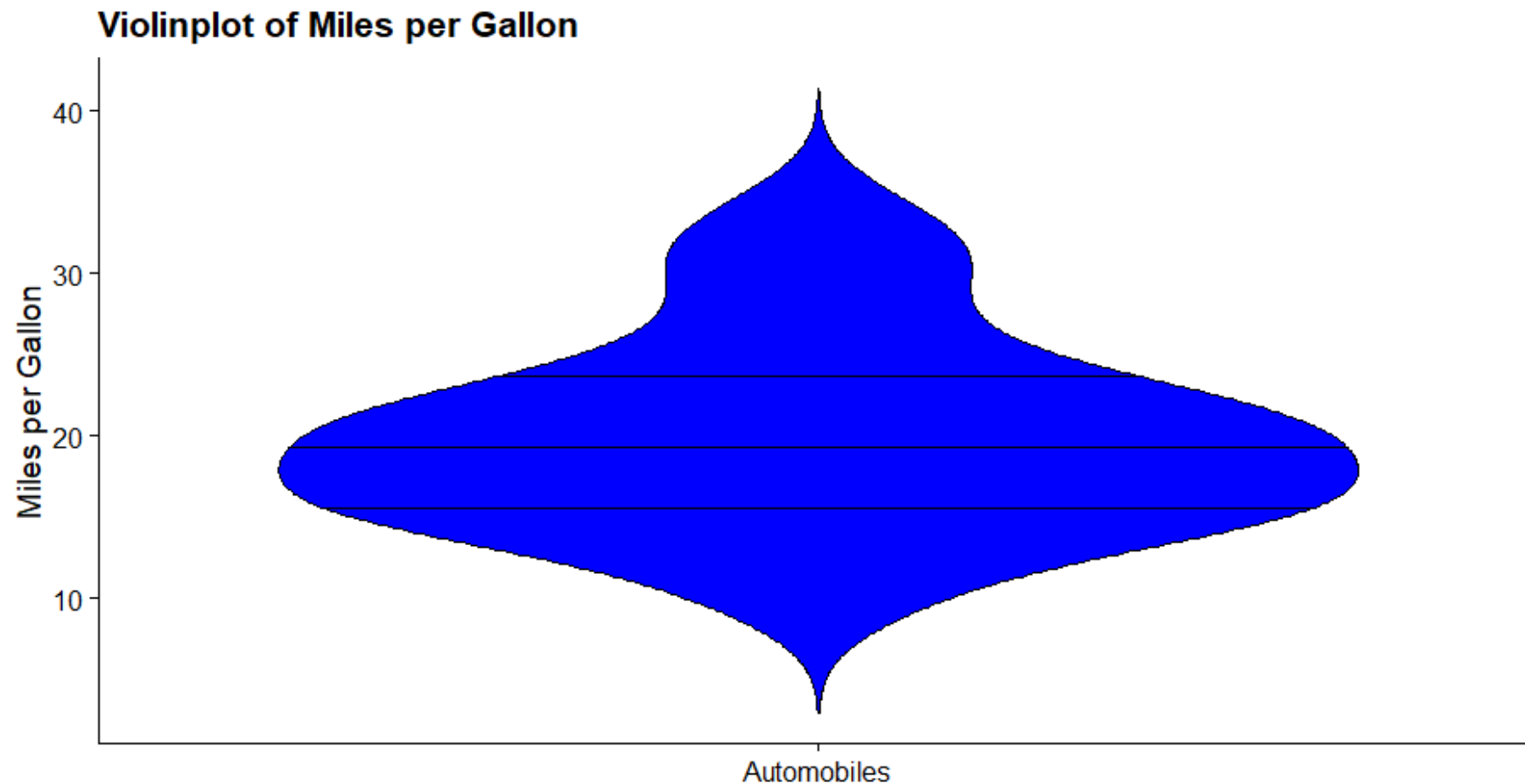
## 小提琴图 boxplot

```
mtcars$type <- 'Automobiles'
fig3 <- ggplot(data = mtcars, aes(y = mpg, x = type)) +
  geom_violin(fill = 'blue', color = 'black',
    draw_quantiles = c(0.25, 0.5, 0.75),
    trim = F) +
  xlab('') +
  ylab('Miles per Gallon') +
  ggtitle('Violinplot of Miles per Gallon')
fig3
```



# 单变量分析

## 小提琴图 boxplot

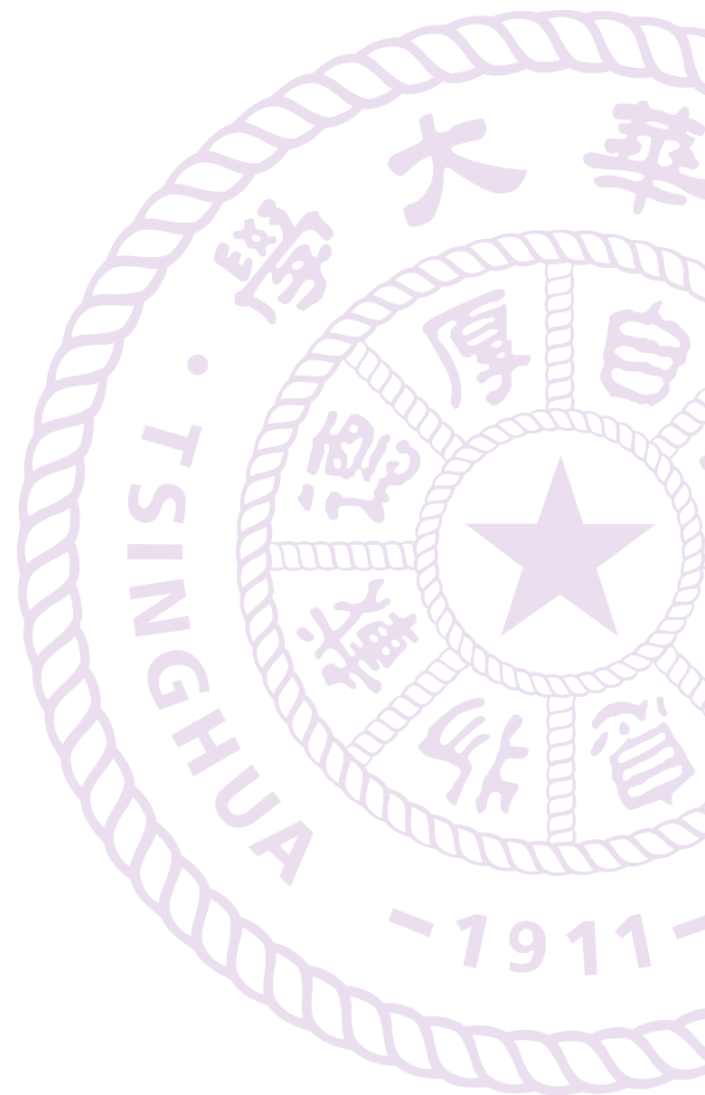


# 单变量分析

## 条形图 bar chart

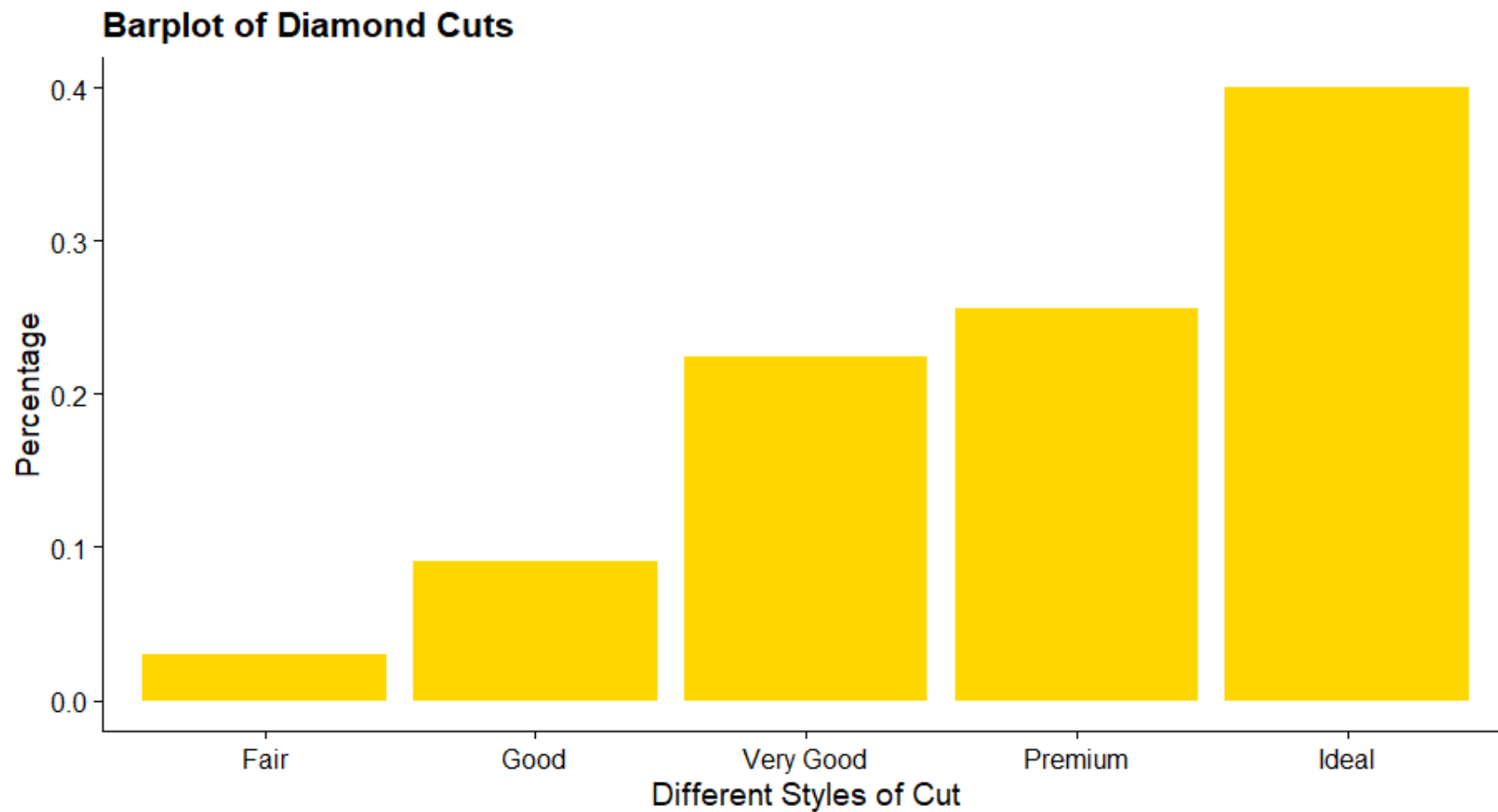
```
fig4 <- ggplot(data = diamonds) +  
  geom_bar(aes(x = cut, y = ..prop..,  
               group = 1), fill = 'gold') +  
  xlab('Different styles of Cut') +  
  ylab('Percentage') +  
  ggtitle('Barplot of Diamond Cuts')
```

fig4



# 单变量分析

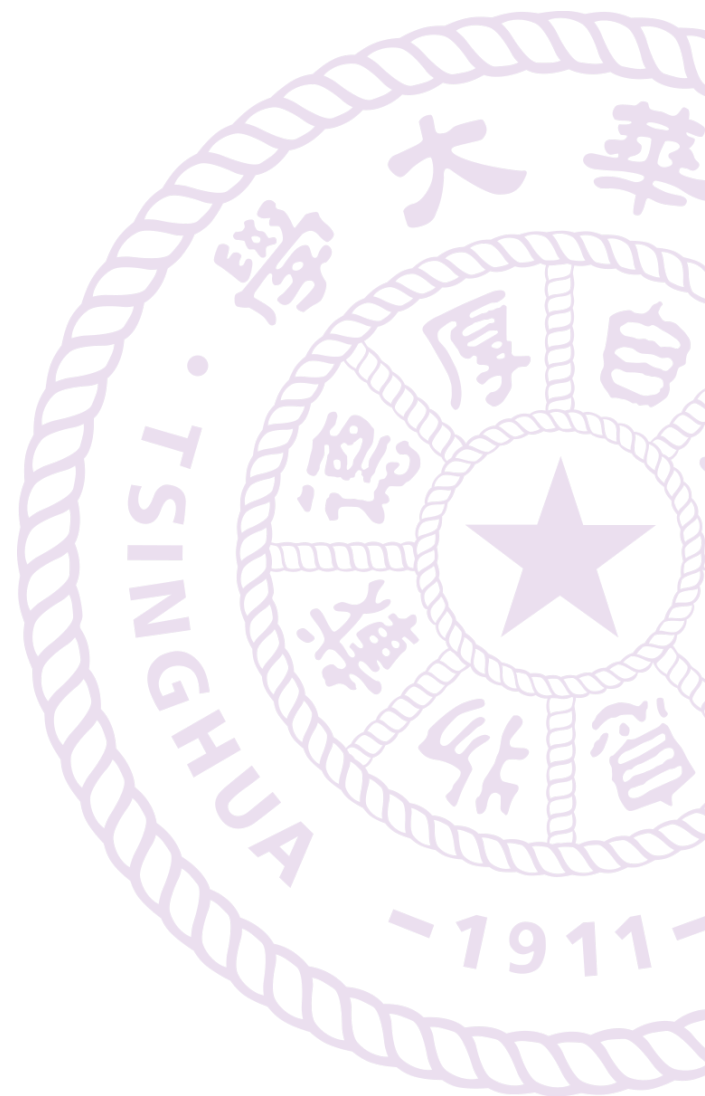
## 条形图 bar chart



# 单变量分析

## 饼图 bar chart

```
pie.data <- diamonds %>%  
  group_by(cut) %>%  
  summarise(perct = n()/nrow(diamonds))
```



# 单变量分析

## 饼图 bar chart

```
fig5 <- ggplot(data = pie.data, aes(x = '')) +  
  geom_bar(aes(y = perct, fill = cut), stat = 'identity') +  
  coord_polar('y', start = 0) +  
  xlab('') +  
  ylab('') +  
  ggtitle('Pie Chart of Diamond Cuts') +  
  theme(axis.ticks = element_blank(),  
        axis.text = element_blank(),  
        axis.line = element_blank())
```

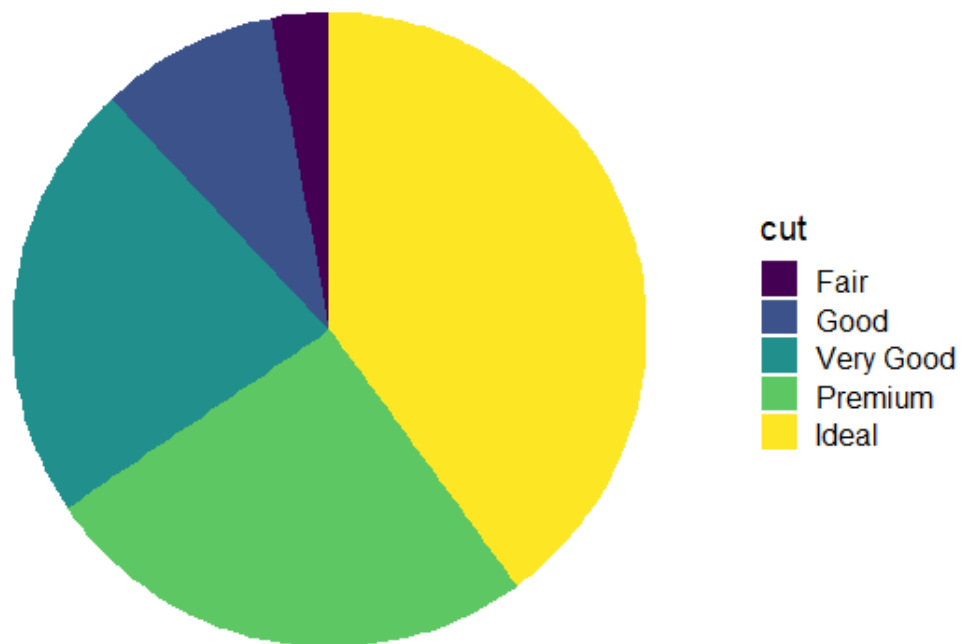
fig5



# 单变量分析

## 饼图 bar chart

Pie Chart of Diamond Cuts







## 课程结构

- I. 为什么要选择R
- II. ggplot绘图语法
- III. 色彩原理
- IV. 基本图表类型
- V. 单变量分析
- VI. 双变量分析**

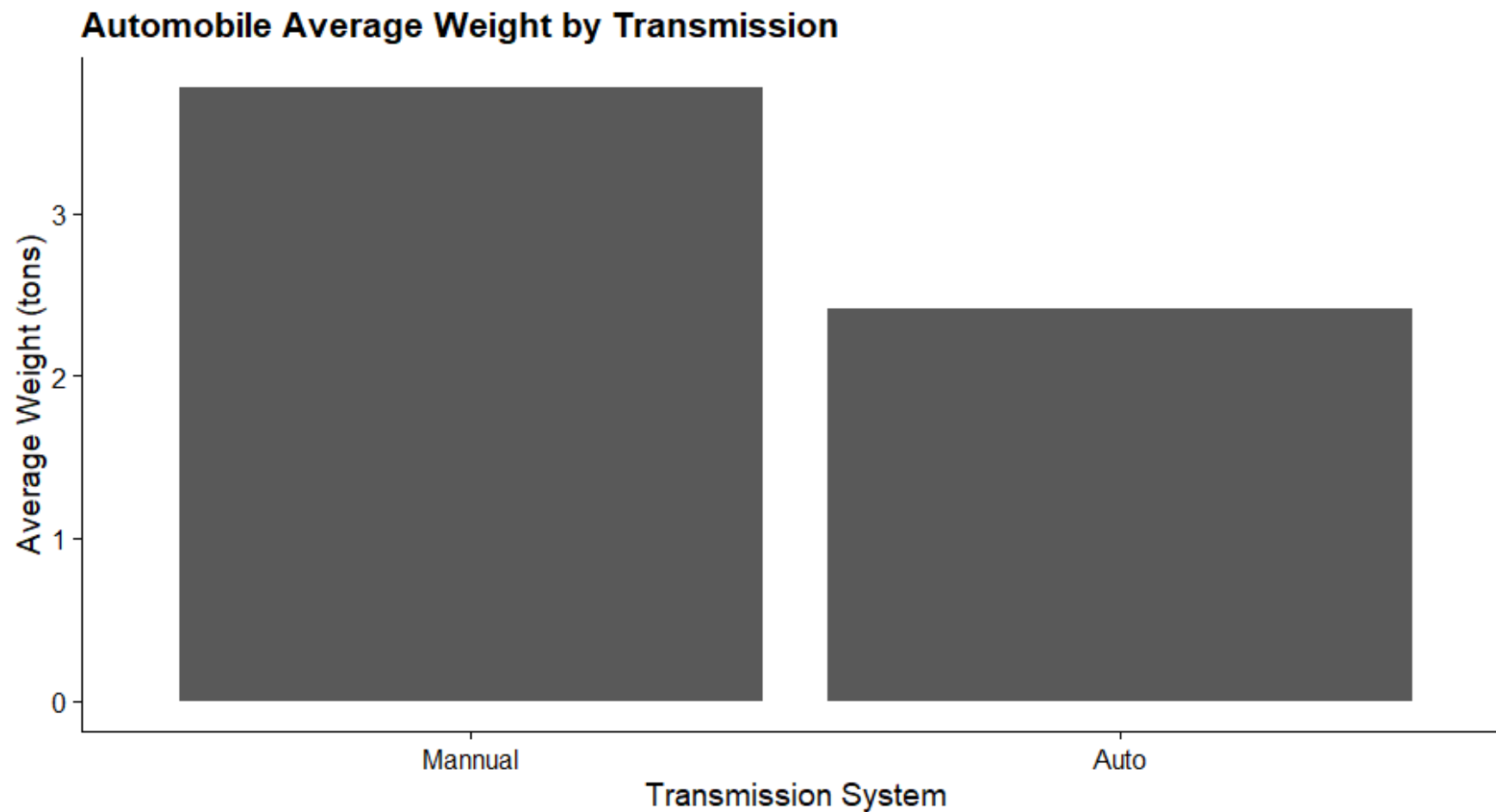
# 双变量分析

## 集中趋势

```
cars.new <- mtcars %>%  
  group_by(am) %>%  
  summarise(wt.avg = mean(wt))  
cars.new$am <- factor(cars.new$am , labels = c('Manual', 'Auto'))  
  
fig6 <- ggplot(data = cars.new) +  
  geom_bar(aes(x = am, y = wt.avg), stat = 'identity') +  
  xlab('Transmission System') +  
  ylab('Average weight (tons)') +  
  ggtitle('Automobile Average weight by Transmission')  
fig6
```

# 双变量分析

## 集中趋势



# 双变量分析

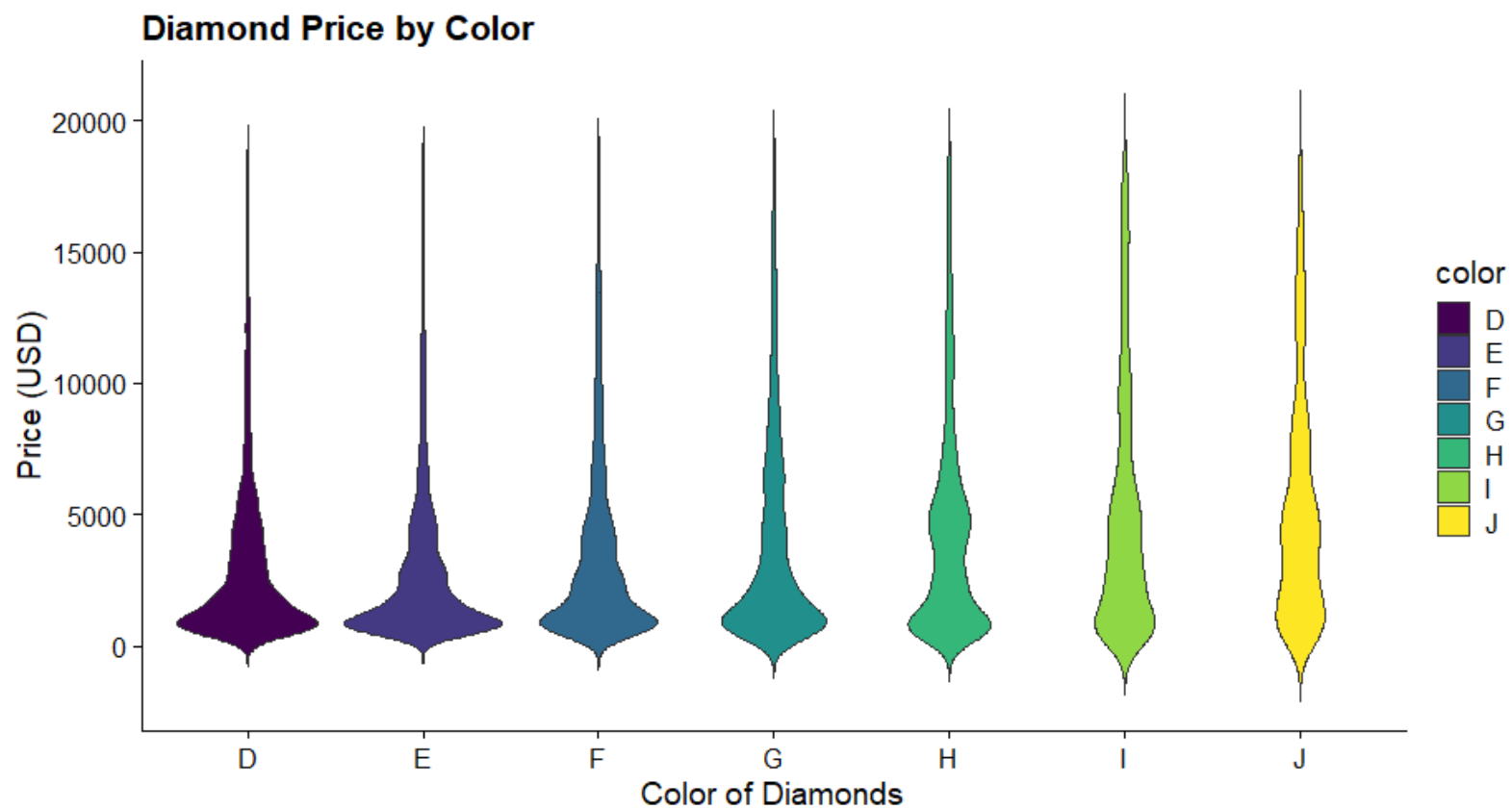
## 离散趋势

```
fig7 <- ggplot(data = diamonds) +  
  geom_violin(aes(x = color, y = price, fill = color),  
             trim = F) +  
  xlab('Color of Diamonds') +  
  ylab('Price (USD)') +  
  ggtitle('Diamond Price by Color')  
fig7
```



# 双变量分析

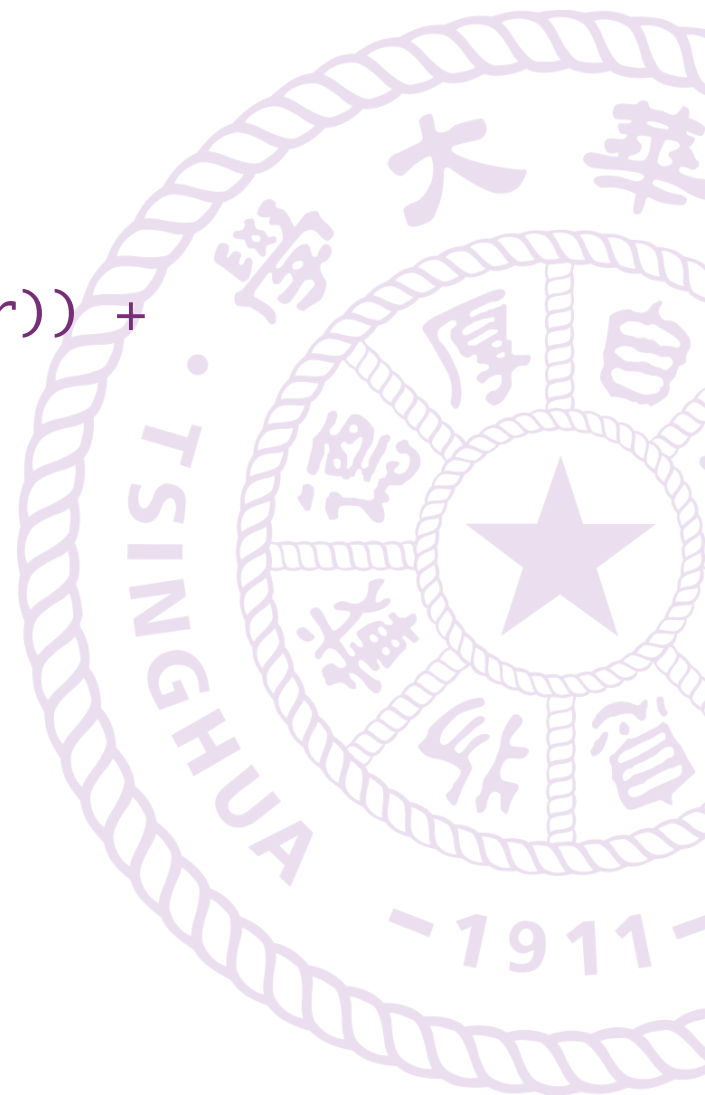
## 离散趋势



# 双变量分析

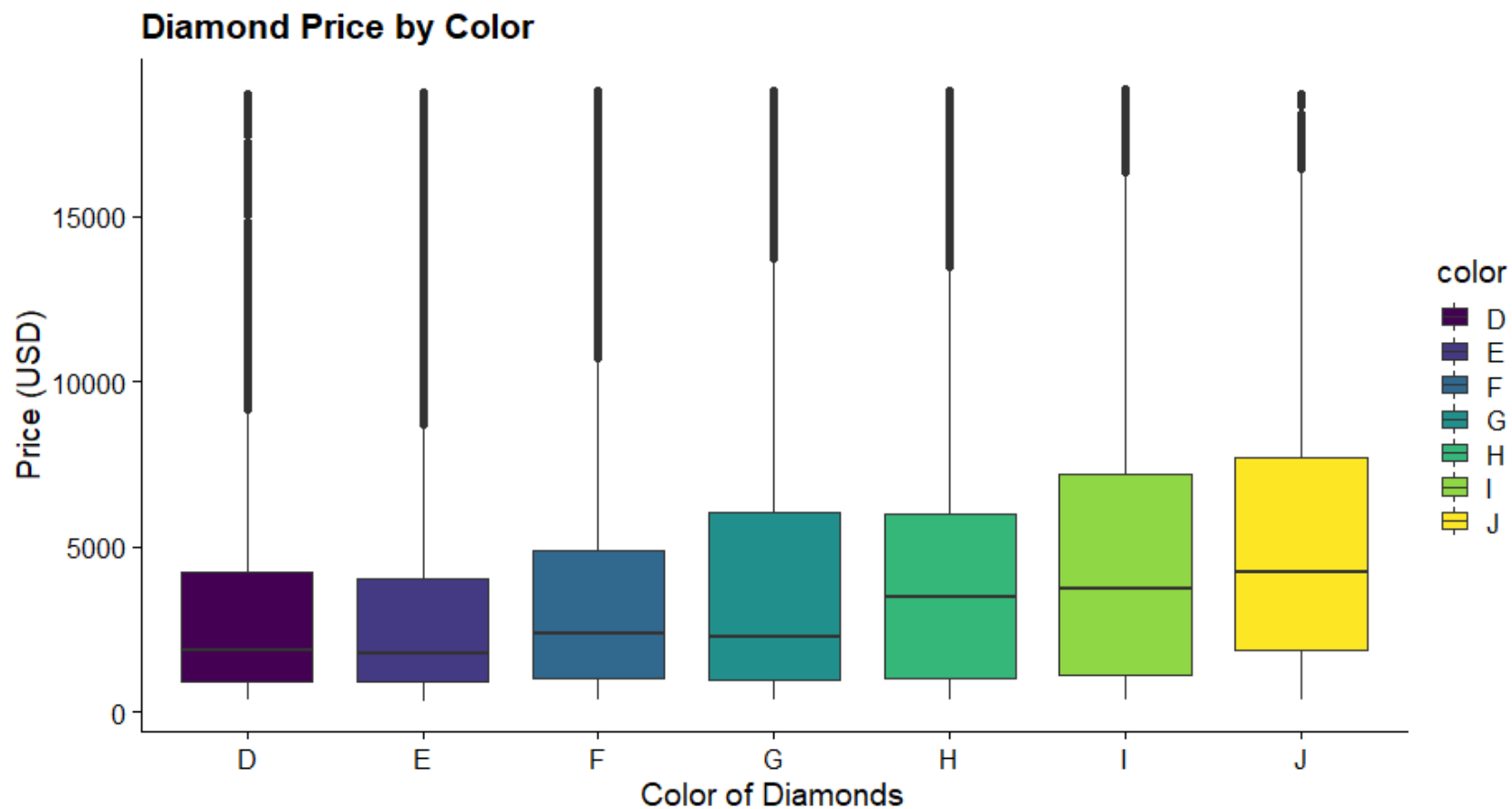
## 均值+离散

```
fig8 <- ggplot(data =diamonds) +  
  geom_boxplot(aes(x = color, y = price, fill = color)) +  
  xlab('Color of Diamonds') +  
  ylab('Price (USD)') +  
  ggtitle('Diamond Price by Color')  
fig8
```



# 双变量分析

## 均值+离散



# 双变量分析

## 相关关系

```
fig9 <- ggplot(data = diamonds, aes(x =carat, y = price )) +  
  geom_point() +  
  geom_smooth(method = 'auto', color = 'red', se = T)+  
  xlab('Diamond weight (carat)') +  
  ylab('Price (USD)') +  
  ggtitle('diamond Price by weight')
```

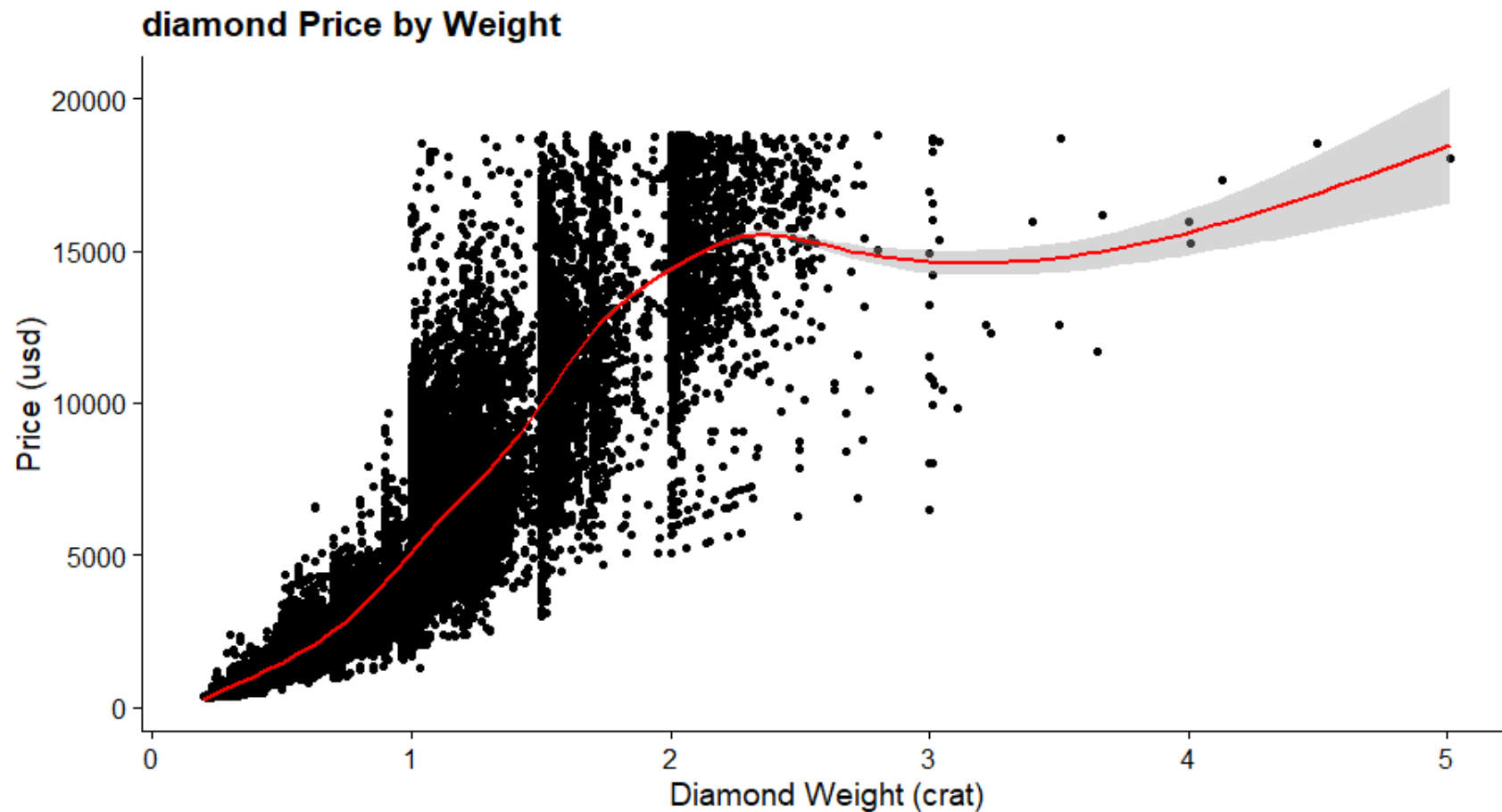
fig9





# 双变量分析

## 相关关系

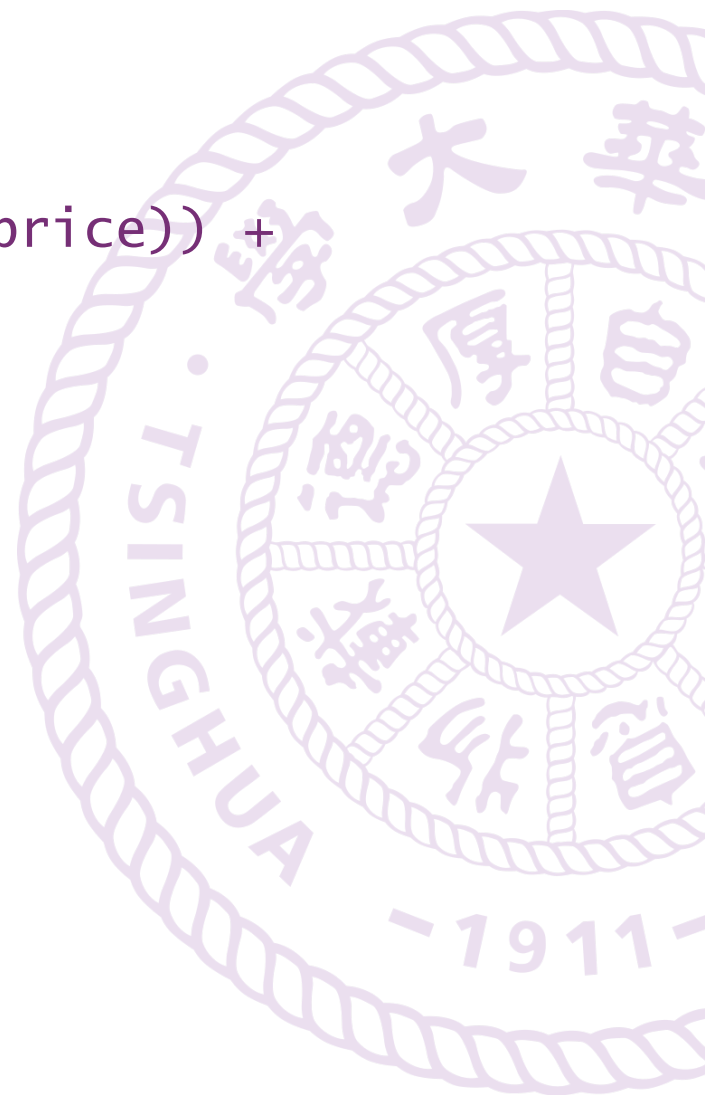


# 双变量分析

## 相关关系

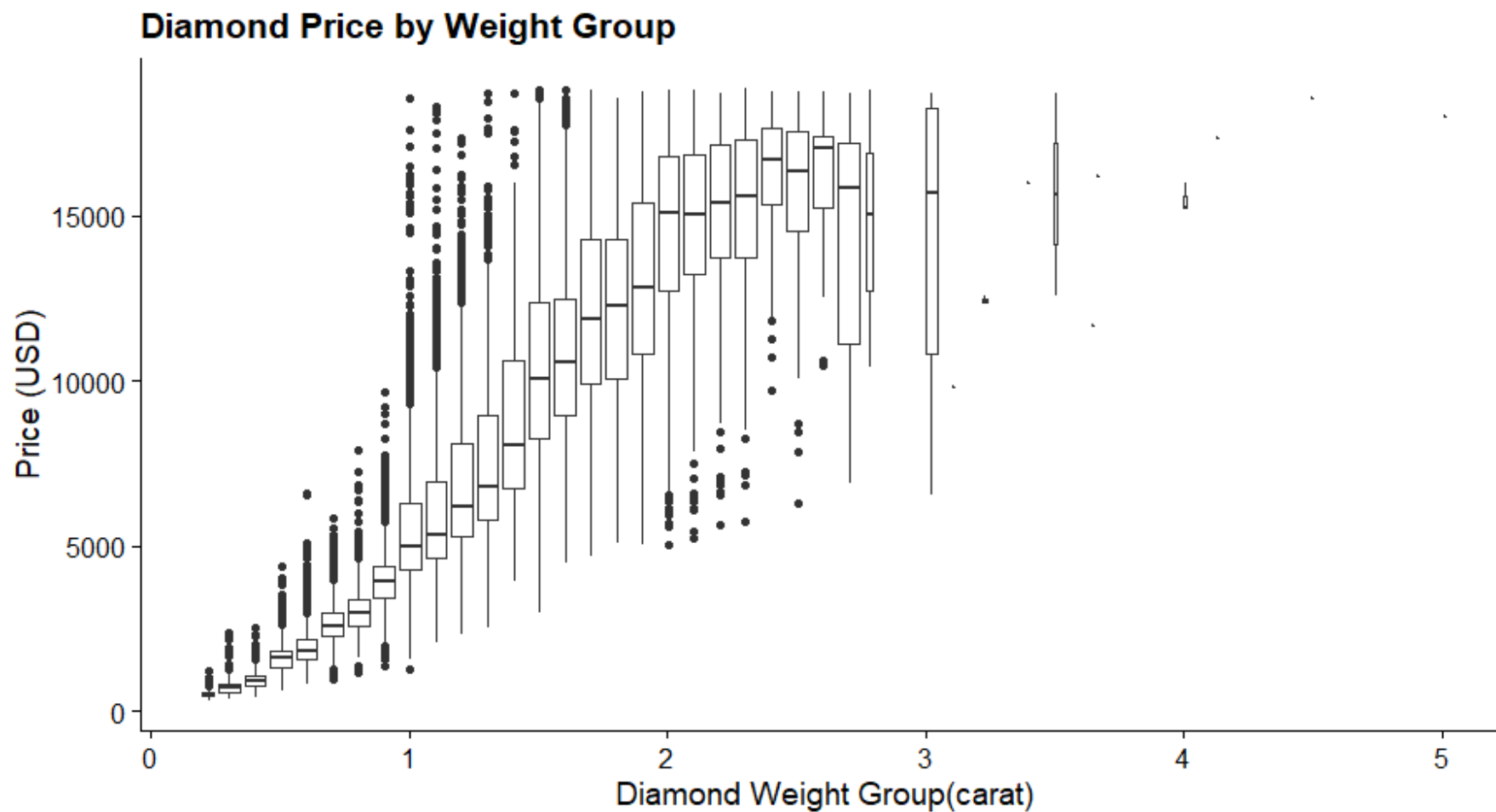
```
fig10 <- ggplot(data = diamonds, aes(x = carat, y = price)) +  
  geom_boxplot(aes(group = cut_width(carat, 0.1))) +  
  xlab('Diamond weight Group(carat)') +  
  ylab('Price (USD)') +  
  ggtitle('Diamond Price by weight Group')
```

fig10



# 双变量分析

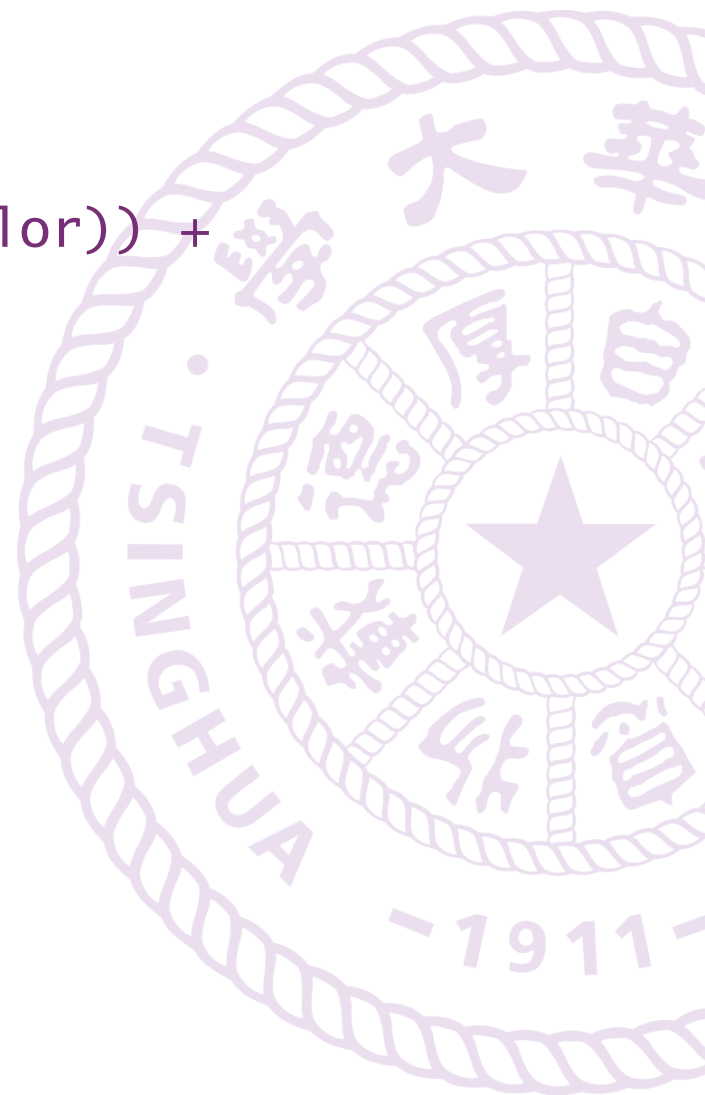
## 相关关系



# 双变量分析

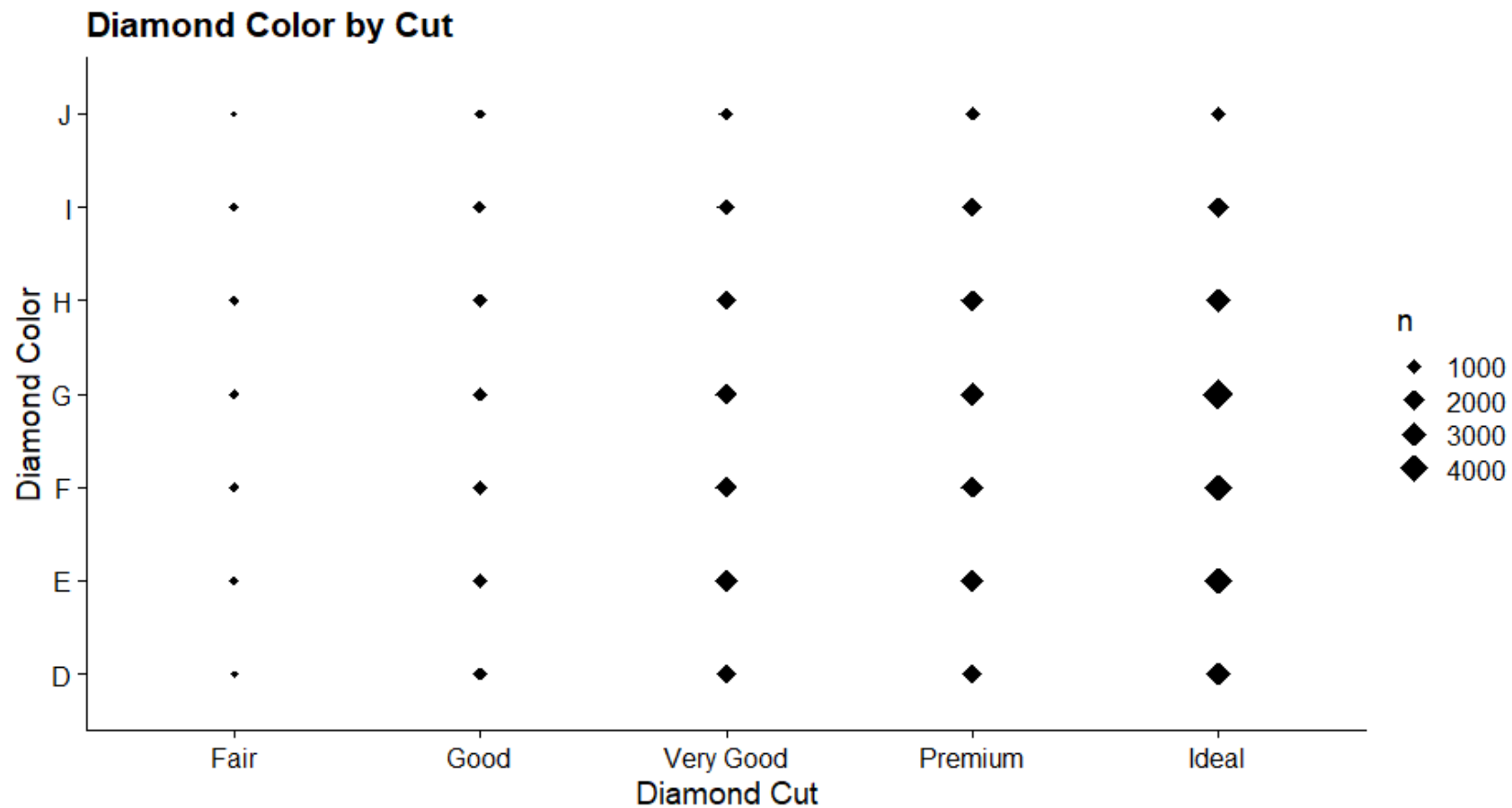
## 相关关系

```
fig11 <- ggplot(data = diamonds, aes(x = cut, y = color)) +  
  geom_count(shape = 'diamond') +  
  xlab('Diamond Cut') +  
  ylab('Diamond Color') +  
  ggtitle('Diamond Color by Cut')  
fig11
```



# 双变量分析

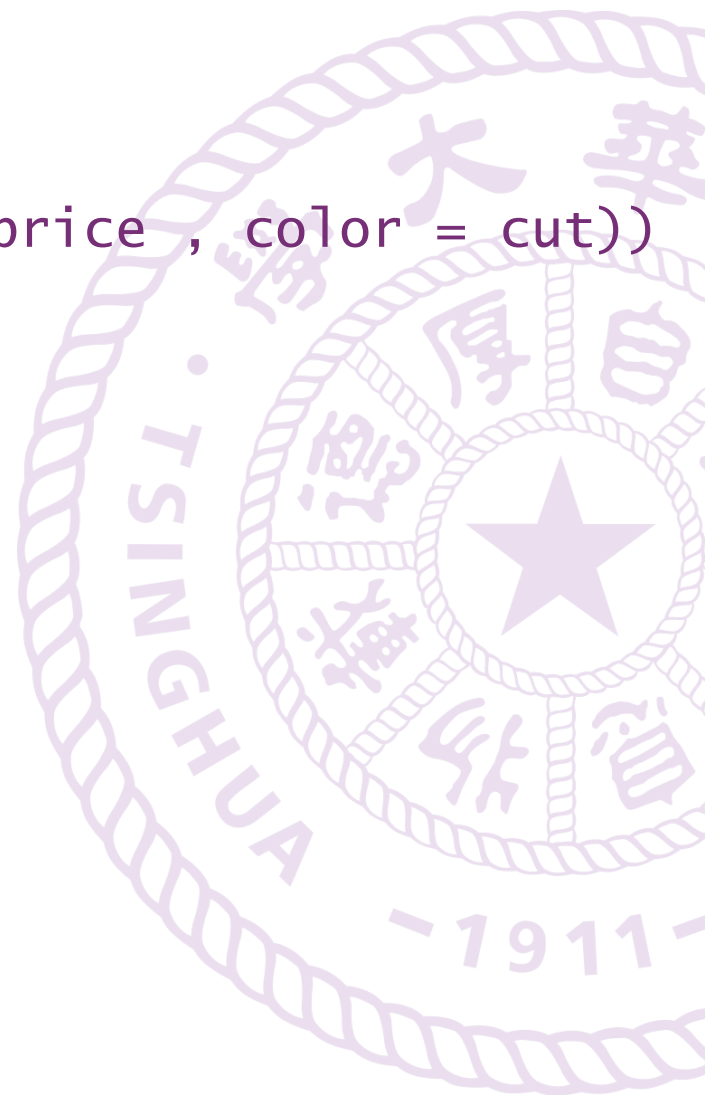
## 相关关系



# 双变量分析

等等，真的是那条曲线吗

```
fig14 <- ggplot(data = diamonds, aes(x = carat, y = price , color = cut))  
+  
  geom_point() +  
  geom_smooth(method = 'lm') +  
  xlab('Diamond weight (carat)') +  
  ylab('Diamond Price (USD)') +  
  ggtitle('Diamond Price by weight by Cut')  
fig14
```



## 双变量分析

等等，真的是那条曲线吗

