《生物计算机程序设计语言》

“Biology Computer Programming Language”

杨建华、郑凌伶

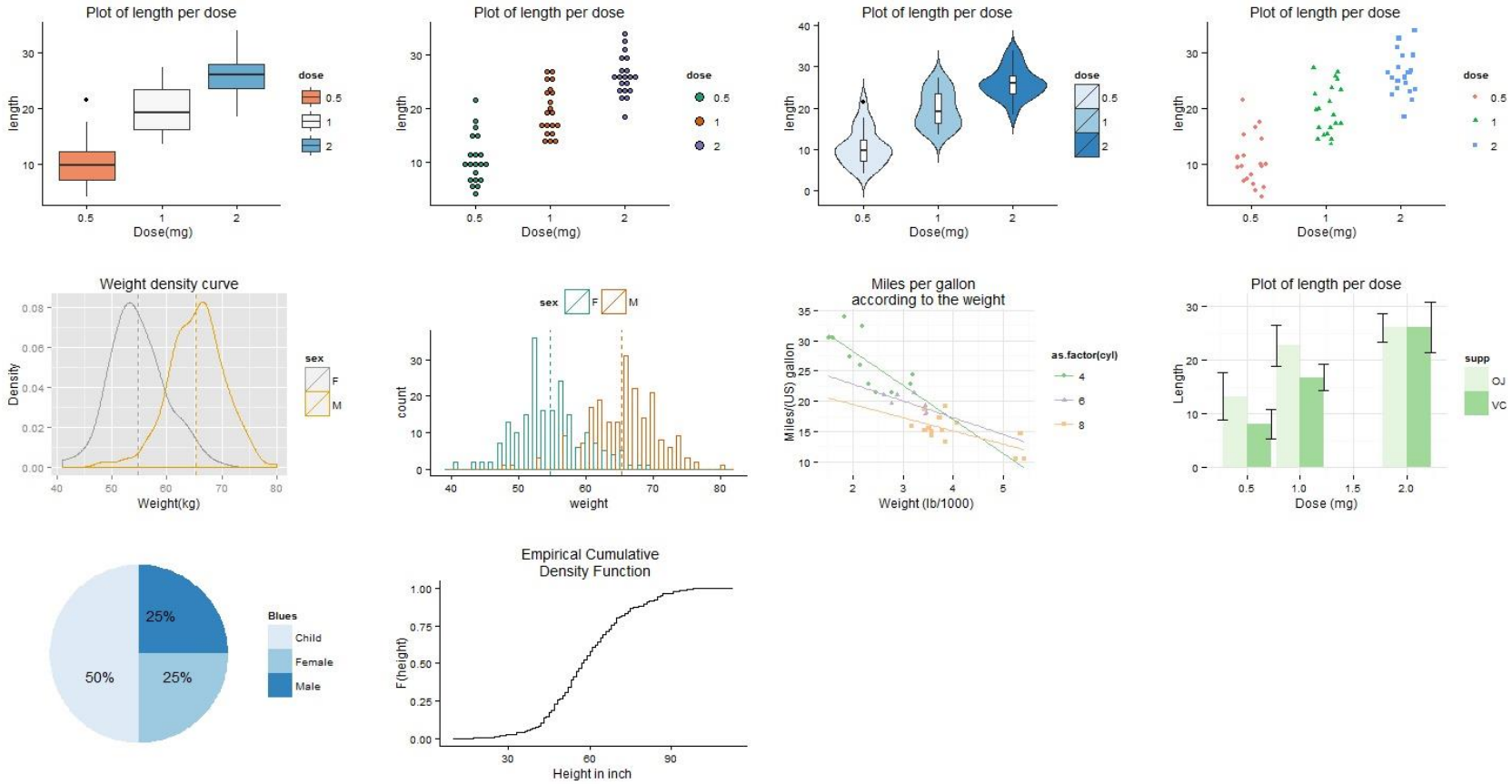
生命科学学院

2019年10月15日

Chapter 7: R-ggplot2

ggplot2

- **ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics**



1. Install ggplot2 package

install.packages('ggplot2')

2. Loading

library(ggplot2)

ggplot2

- ggplot2是R语言最流行的第三方扩展包，是RStudio首席科学家Hadley Wickham读博期间的作品。
- 包名中“gg”是grammar of graphics的简称，是一套优雅的绘图语法。
- **设计的基本概念：Plot = data + Aesthetics + Geometry.**

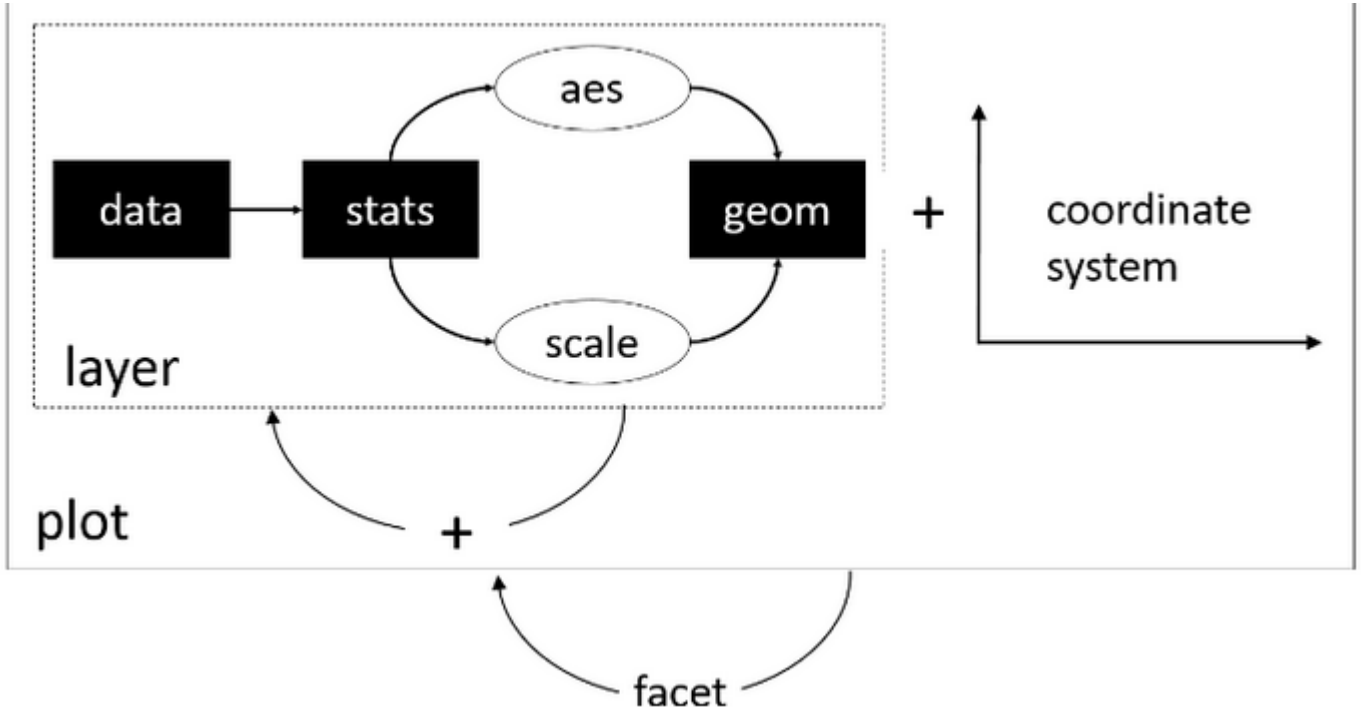
数据（data）+图形属性（aesthetic attribute）+几何对象（geometric object）

The grammar of graphics

- Wilkinson 在2005年所写的关于统计图形的总结性抽象
- 主要概念：一张统计图形就是从**数据**（data）到**几何对象**（geometric object，缩写geom: **points, lines, bars**）的**图形属性**（aesthetic attribute，缩写aes: **colour, shape, size** ）的一个**映射**（**mapping**）。
- 此外，图形中还可能包含数据的**统计变换**（statistical transformation，缩写**stats**），最后绘制在某个特定的**坐标系**（coordinate system，缩写**coord**）中，而**分面**（**facet**）则可以用来生成数据不同子集的图形。

ggplot重要特性

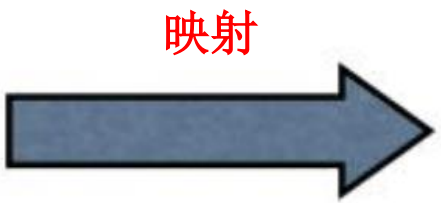
- 数据：**data**
- 统计变换：**stat**istics
- 几何对象：**geom**etric
- 标尺：**scale**
- 图层：layer
- 坐标系：**Coord**inate
- 分面：facet
- 图形属性： **aes**thetic attribute



数据（data）和映射（mapping）

- 将数据中的**变量映射到图形属性**。映射控制了二者之间的关系

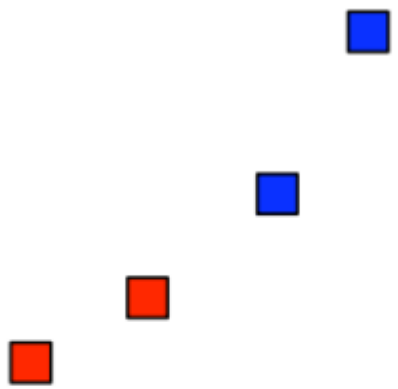
length	width	depth	trt
2	3	4	a
1	2	1	a
4	5	15	b
9	10	80	b



x	y	colour
2	3	a
1	2	a
4	5	b
9	10	b

几何对象（geometric）及图形属性（aesthetic attribute）

- 代表你在图中实际看到的图形元素，如点、线、多边形等



Geoms

几何对象

- 几何对象，说的直观一些，就是你要**画什么图**来表示这组数据。ggplot2提供了众多几何对象**geom_xyz()**供大家选择。
- 举两个常见的例子，**geom_point()**用于表示两个连续变量之间的关系，几何形状是点；
- **geom_bar()**用于表示x轴为离散变量，y轴为连续连续变量之间的关系，几何形状是条块。

图形属性

- 每个几何对象都有自己的属性，这些**属性的取值需要通过数据提供**。
- 数据与图形属性之间的**映射**关系称为**mapping**，在ggplot2中用**aes()进行定义**。
- 常见的图形属性有：x，y，size，color，group。图形属性的**任意一项**都可以用**数据的某一个变量**来表示。

统计变换（statistics）

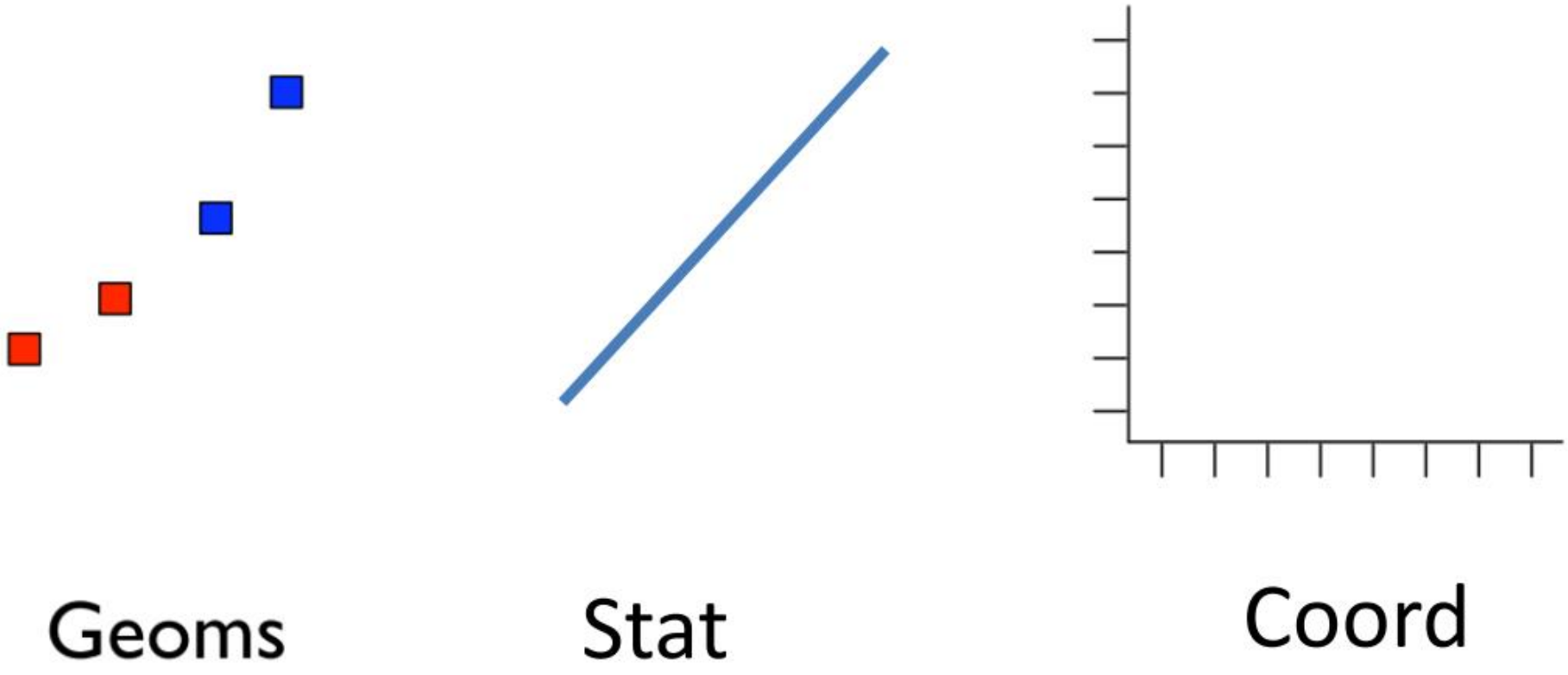
- 对原始数据进行某种计算，例如对**二元散点图加上一条回归线**。

Geoms

Stat

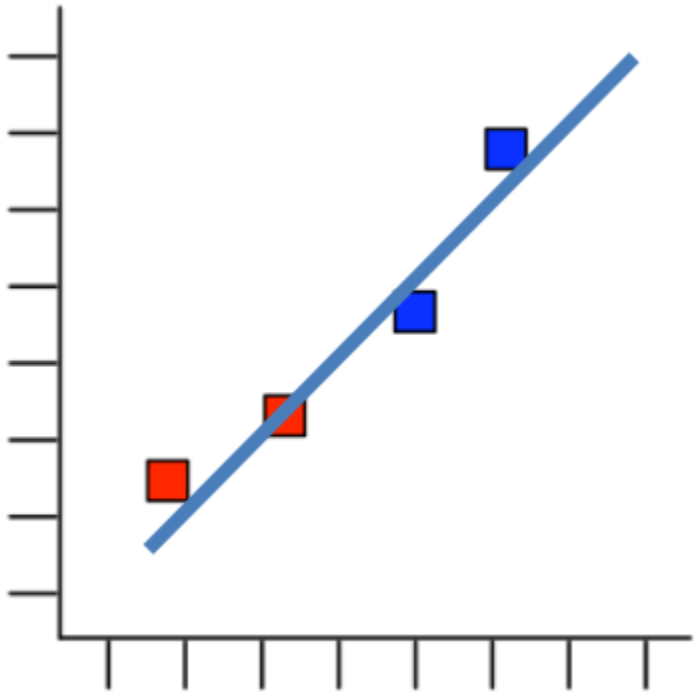
坐标系统（Coordinate）

- 坐标系统控制坐标轴并影响所有图形元素，坐标轴可以进行变换以满足不同的需要



图层（Layer）

- 图层可以允许用户一步步的构建图形，方便单独对图层进行修改、增加统计量、甚至改动数据

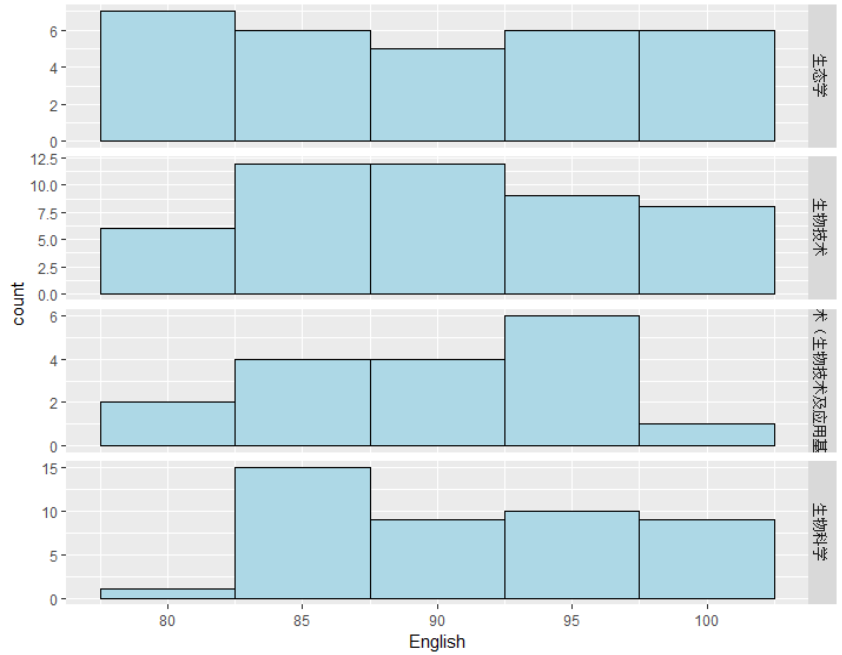


图层

- ggplot2的绘图过程有点像Photoshop，有一个图层的理念，每个图层可以有自己的**图形对象和图形属性**，**通过+**将不同图层叠加起来生成最后的统计图形。
- 如果将**数据定义在ggplot()**中，那么所有图层都可以共用这个数据；如果将**数据定义在几何对象geom_xyz()**中，那么这个数据就只供这个几何对象使用。

分面（Facet）

- 条件绘图，将**数据按某种方式分组，然后分别绘图**。分面就是控制分组绘图的方法和排列形式



数据

- ggplot2接受的输入数据一般是data.frame，每一行是一个观测（observation），每一列是一个变量（variable）

```
sysu_student<-read.table("sysu_student_score.csv", header=TRUE)
```

	Id	Name	Major	English	Math	Computer	total
1	16336001	阿比达·阿布来提	生物技术	80	92	95	267
2	16336007	蔡静	生物技术	96	91	93	280
3	16336008	蔡奇	生物技术	85	86	99	270
4	16336010	蔡响	生物科学	82	91	85	258
5	16336014	曾思琳	生物科学	95	88	83	266
6	16336019	陈嘉杰	生物科学	97	87	93	277
7	16336025	陈琳琪	生物技术	87	95	88	270
8	16336029	程海涛	生物技术	90	83	100	273
9	16336030	程凯平	生物技术	83	82	94	259

使用str()查看数据集的结构

```
str(sysu_student)
```

```
## 'data.frame': 138 obs. of 7 variables:
## $ Id : int 16336001 16336007 16336008 16336010 16336014 16336019 16336025 16336029 16336030 16336031 ...
## $ Name : Factor w/ 138 levels "阿比达·阿布来提",..: 1 7 8 9 13 16 18 21 22 23 ...
## $ Major : Factor w/ 4 levels "生态学","生物技术",..: 2 2 2 4 4 4 2 2 2 2 ...
## $ English : int 80 96 85 82 95 97 87 90 83 94 ...
## $ Computer: num 95 93 99 85 83 93 88 100 94 92 ...
## $ total : num 267 280 270 258 266 277 270 273 259 276 ...
```

用summary()对每一个变量进行统计

```
summary(sysu_student)
```

```
       Id                                Name                                Major
Min.   :15335058   阿比达·阿布来提:  1   生态学                                :30
1st Qu.:16336042   阿合博塔·哈孜太:  1   生物技术                              :47
Median :16336096   安建婷          :  1   生物技术（生物技术及应用基地班）:17
Mean   :16328744   包玲慧          :  1   生物科学                              :44
3rd Qu.:16336175   毕结仪          :  1
Max.   :16352036   蔡浩然          :  1
                   (Other)         :132
    English            Math          Computer          total
Min.   : 80.00   Min.   : 80.00   Min.   : 74.00   Min.   :243.0
1st Qu.: 84.00   1st Qu.: 86.00   1st Qu.: 84.00   1st Qu.:262.0
Median : 88.50   Median : 90.00   Median : 88.00   Median :267.0
Mean   : 89.32   Mean   : 89.93   Mean   : 88.55   Mean   :267.8
3rd Qu.: 95.00   3rd Qu.: 94.00   3rd Qu.: 93.00   3rd Qu.:275.0
Max.   :100.00   Max.   :100.00   Max.   :100.00   Max.   :290.0
```

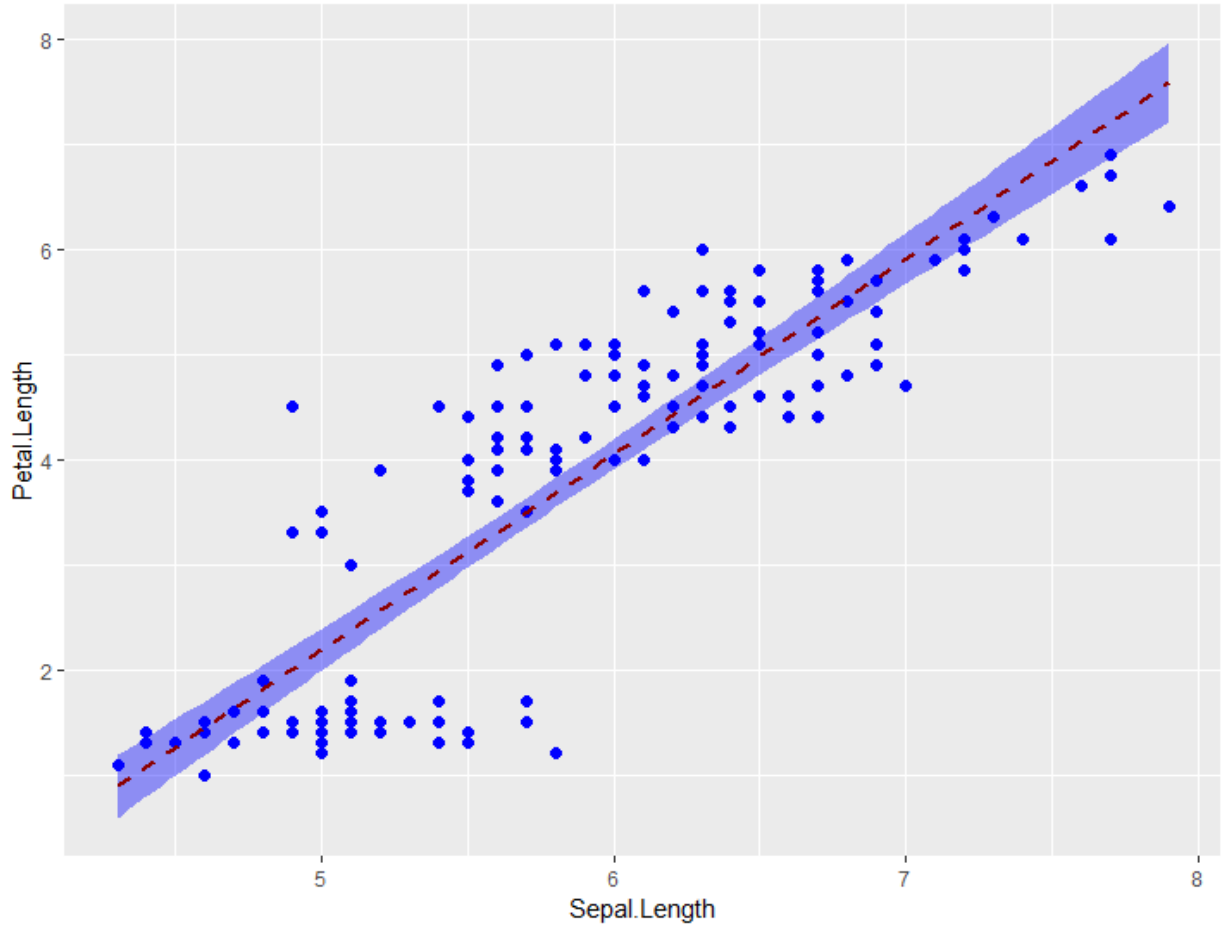
散点图

geom_point(size, color, shape)

```
library(ggplot2)
1.  ggplot(data=iris, mapping=aes(x=Sepal.Length,
    y=Petal.Length)) + geom_point()
2.  ggplot(data=iris, mapping=aes(x=Sepal.Length,
y=Petal.Length)) + geom_point(color="blue" ,size=2,
shape=19) + geom_smooth(method=lm,
linetype="dashed", color="darkred", fill="blue")
```

1.对鸢尾花的三个亚种（setosa、versicolor、virginica）分别统计它们的花萼长度（Sepal.Length）、花萼宽度（Sepal.Width）、花瓣长度（Petal.Length）、花瓣宽度（Petal.Width）的最大值、最小值、均值、中位数等特征，并绘制箱线图进行展示

2. 请问鸢尾花的四种属性花萼长度（Sepal.Length）、花萼宽度（Sepal.Width）、花瓣长度（Petal.Length）、花瓣宽度（Petal.Width）之间是否存在相关性？如何展示？ ##



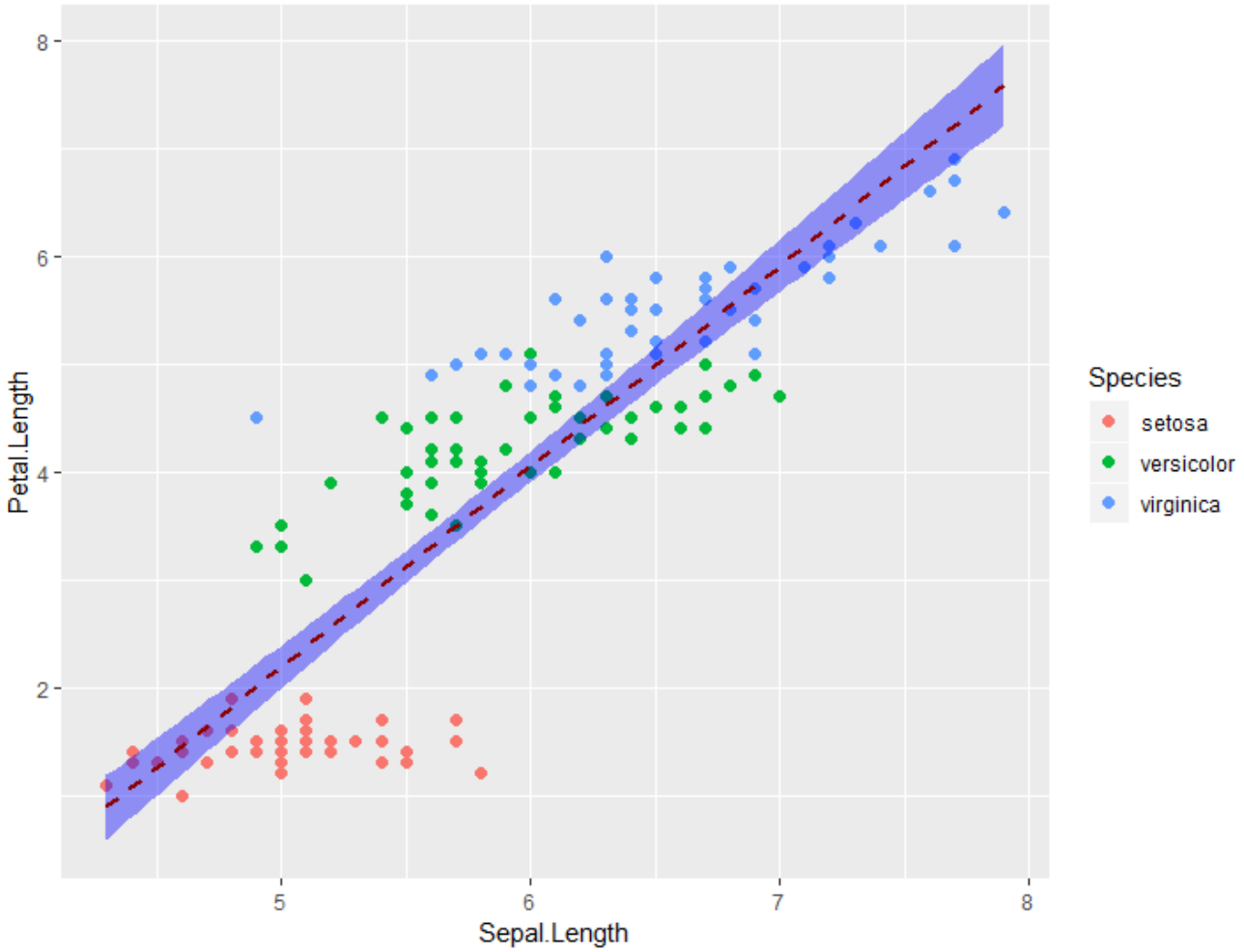
散点图-分Species上颜色

geom_point(size, color, shape)

```
library(ggplot2)
3. ggplot(data=iris, mapping=aes(x=Sepal.Length, y=Petal.Length, color=Species)) + geom_point(size=2, shape=19) + geom_smooth(method=lm, linetype="dashed", color="darkred", fill="blue")
```

1.对鸢尾花的三个亚种（setosa、versicolor、virginica）分别统计它们的花萼长度（Sepal.Length）、花萼宽度（Sepal.Width）、花瓣长度（Petal.Length）、花瓣宽度（Petal.Width）的最大值、最小值、均值、中位数等特征，并绘制箱线图进行展示

2. 请问鸢尾花的四种属性花萼长度（Sepal.Length）、花萼宽度（Sepal.Width）、花瓣长度（Petal.Length）、花瓣宽度（Petal.Width）之间是否存在相关性？如何展示？ ##

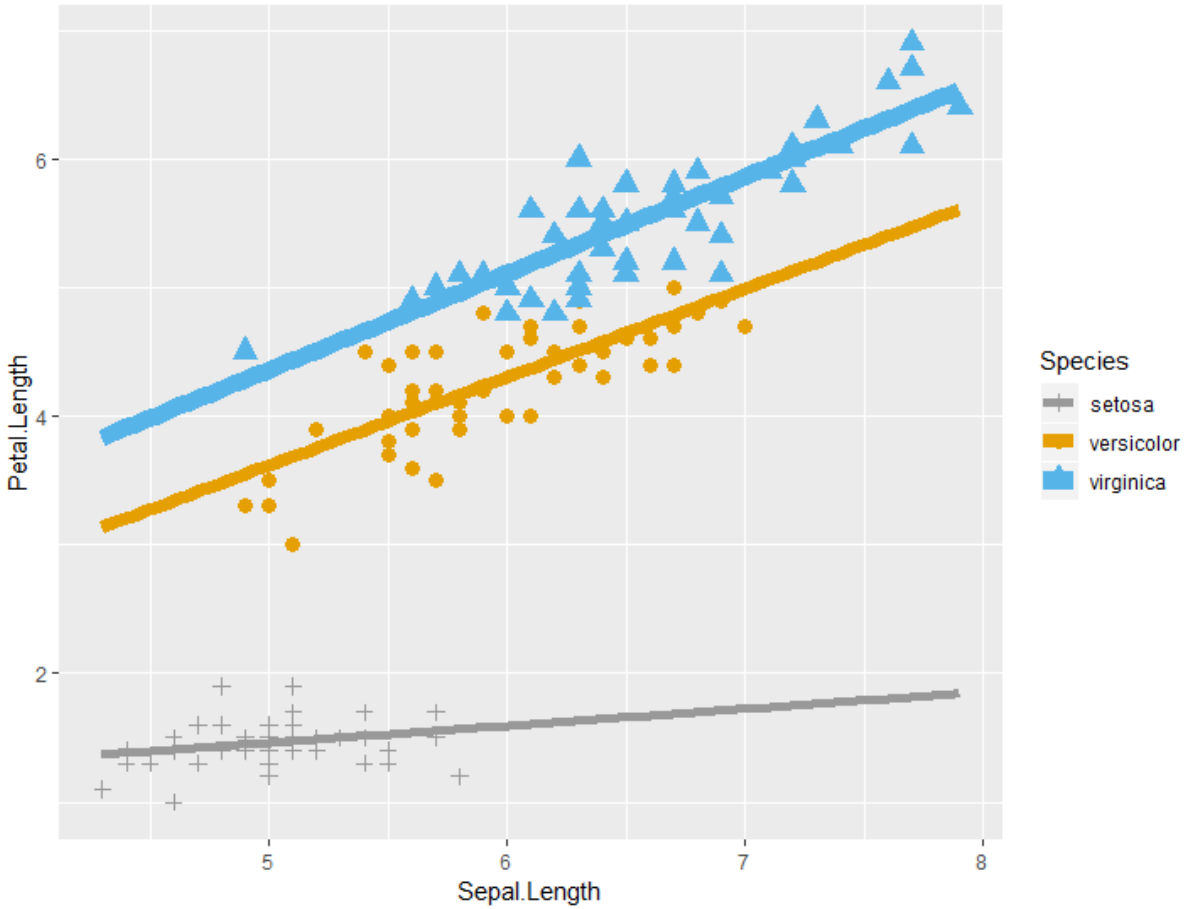


散点图-手动修改点的颜色、形状、大小

geom_point(size, color, shape)

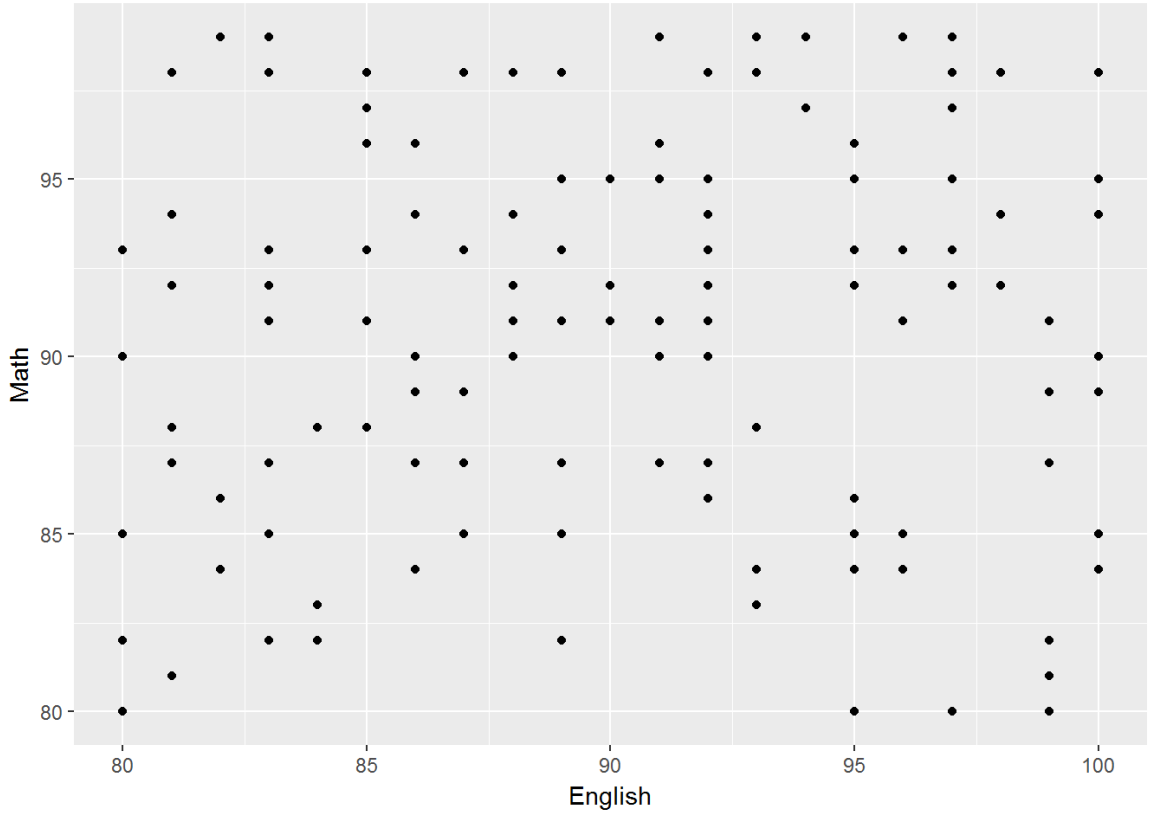
使用scale标尺来修改图形

```
library(ggplot2)
3. ggplot(data=iris, mapping=aes(x=Sepal.Length,
y=Petal.Length, color=Species, shape=Species,
size=Species)) + geom_point() +
geom_smooth(method=lm, se=FALSE, fullrange=TRUE)
+scale_shape_manual(values=c(3, 16, 17))
+scale_color_manual(values=c('#999999','#E69F00',
'#56B4E9'))
+scale_size_manual(values=c(2,3,4))
```



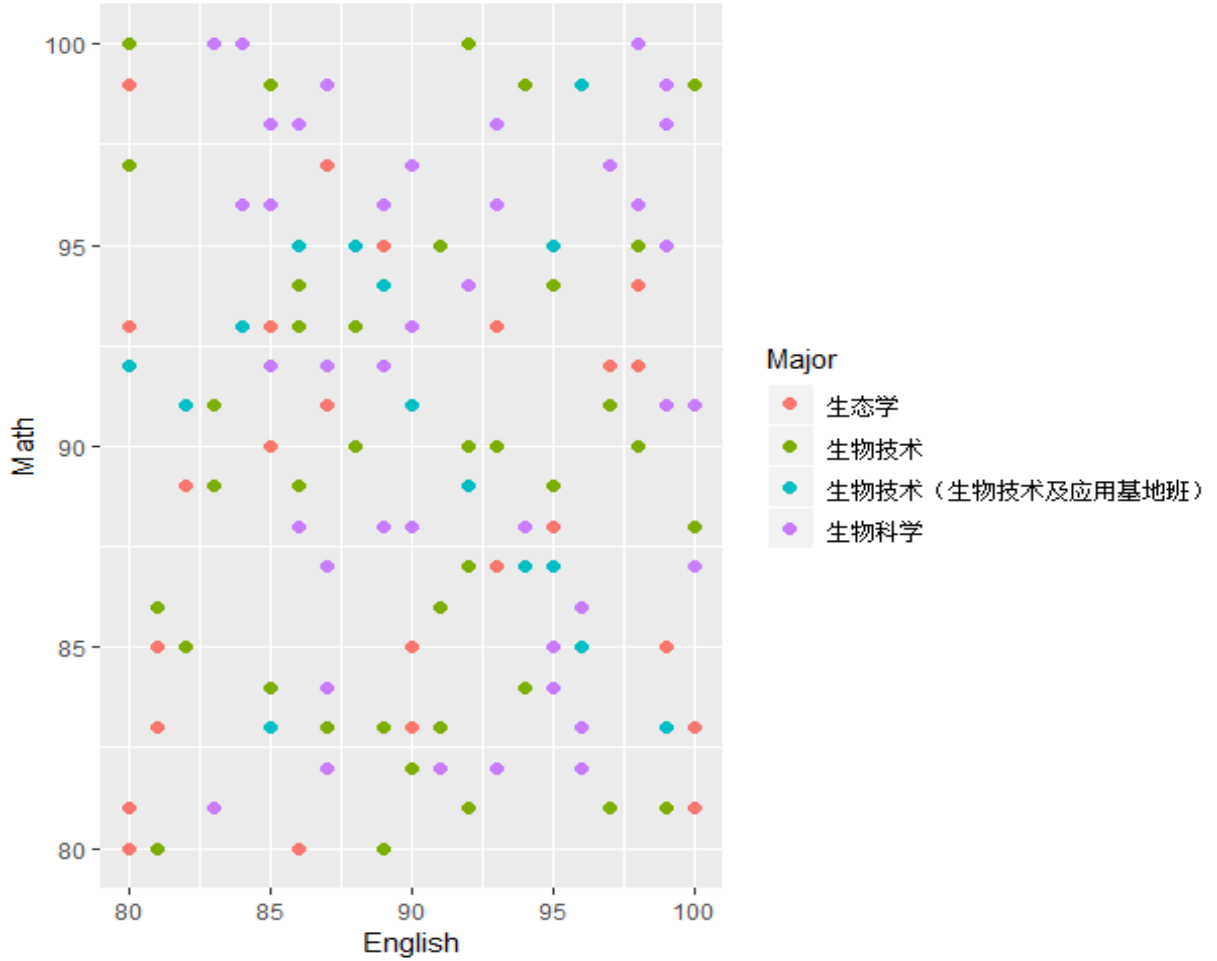
散点图

```
library(ggplot2)

# Make scatter plot of English and Math

p.scatter <- ggplot(data=sysu_student) +

geom_point(mapping=aes(x=English, y=Math))

p.scatter
```



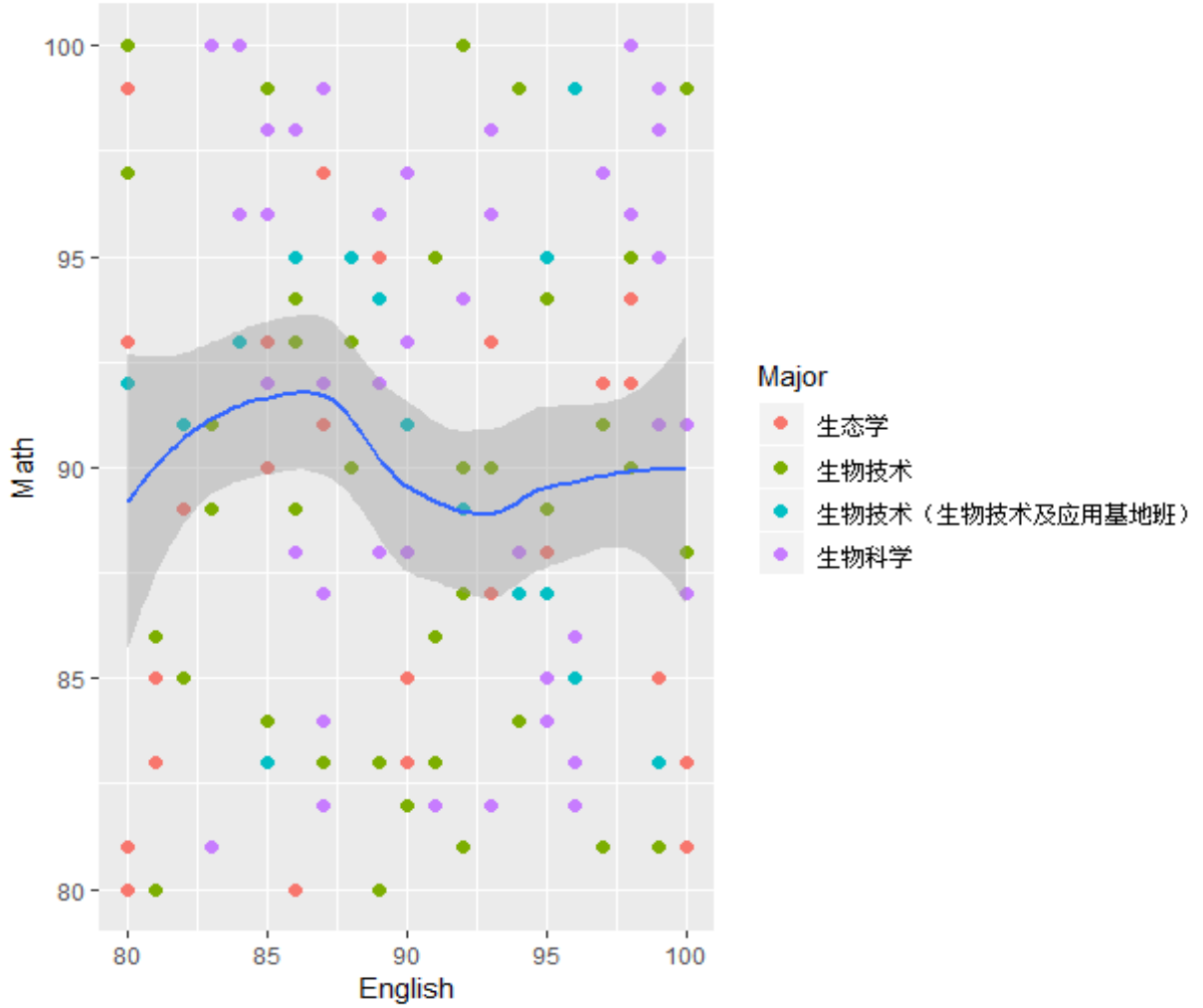
分专业上色

```
p.scatter <- ggplot(data=sysu_student) +

geom_point(aes(x=English, y=Math, color=Major), size=2)

p.scatter
```



添加回归线

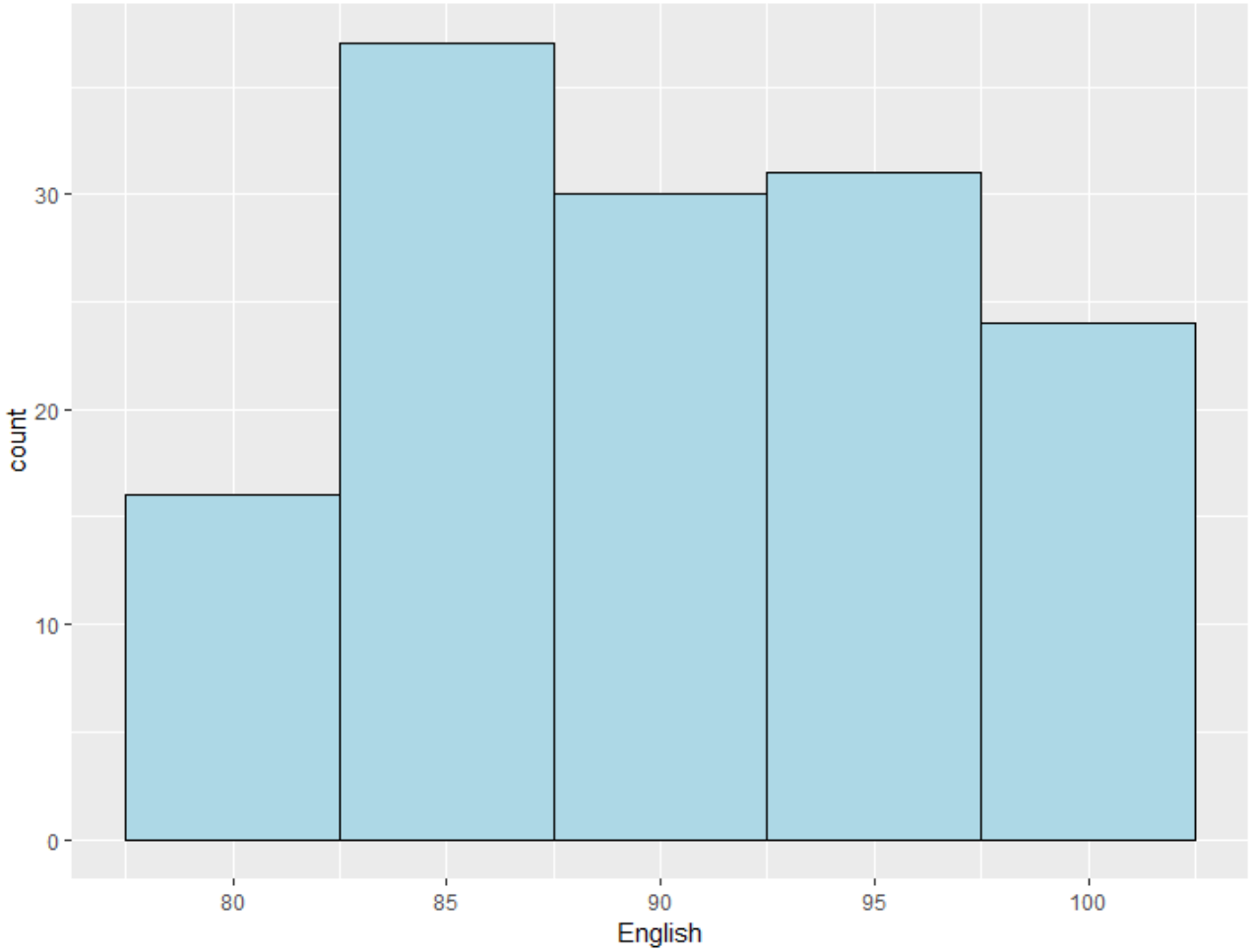
```
p.scatter <- ggplot(data=sysu_student) +
geom_point(aes(x=English, y=Math,
color=Major),size=2) +
geom_smooth(mapping=aes(x=English, y=Math))
p.scatter
```



可以修改smooth的方法等

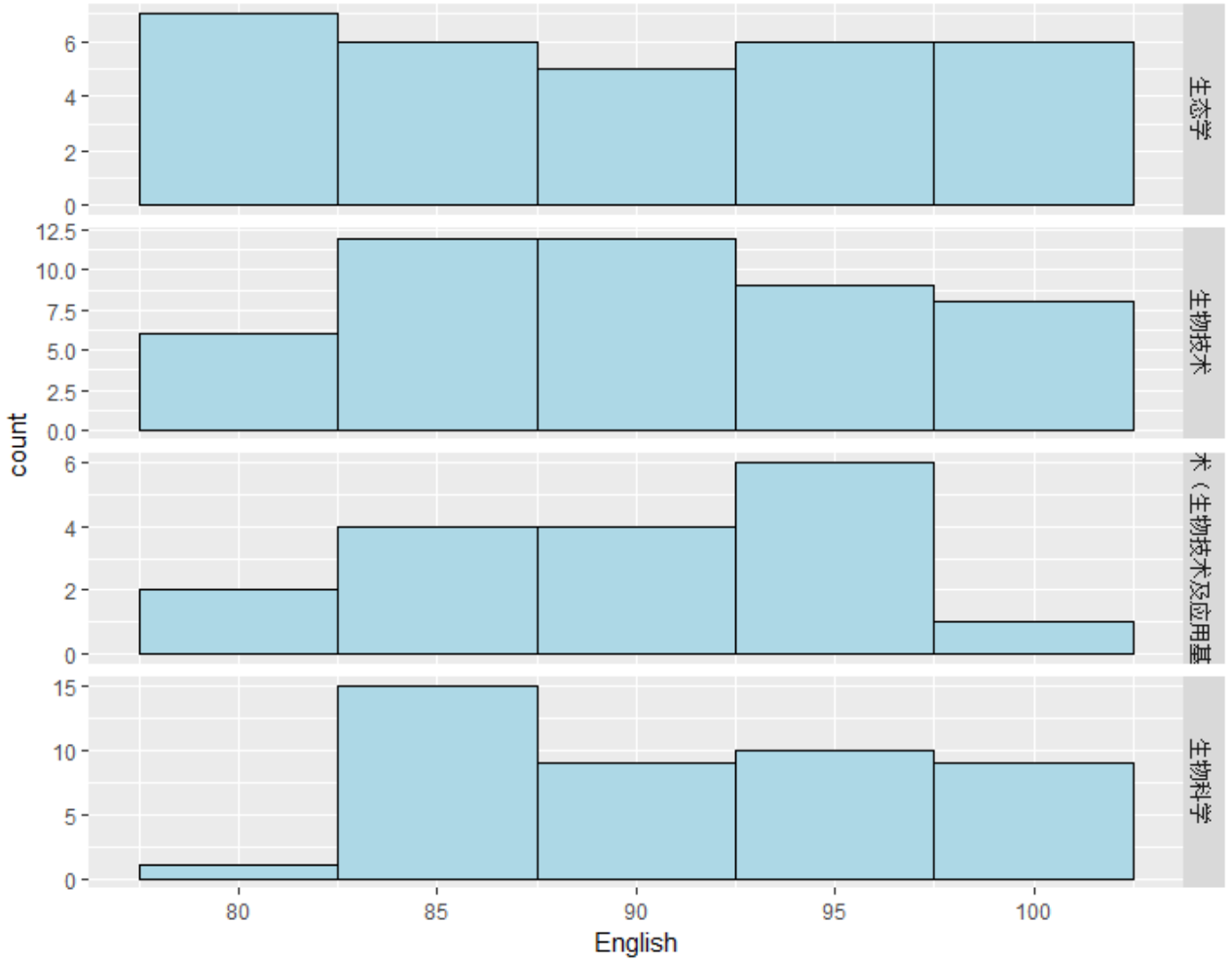
直方图

```
#histogram
ggplot(data=sysu_student,
mapping=aes(x=English)) +
geom_histogram(binwidth = 5,
fill="lightblue", color="black")
```



分专业绘制直方图

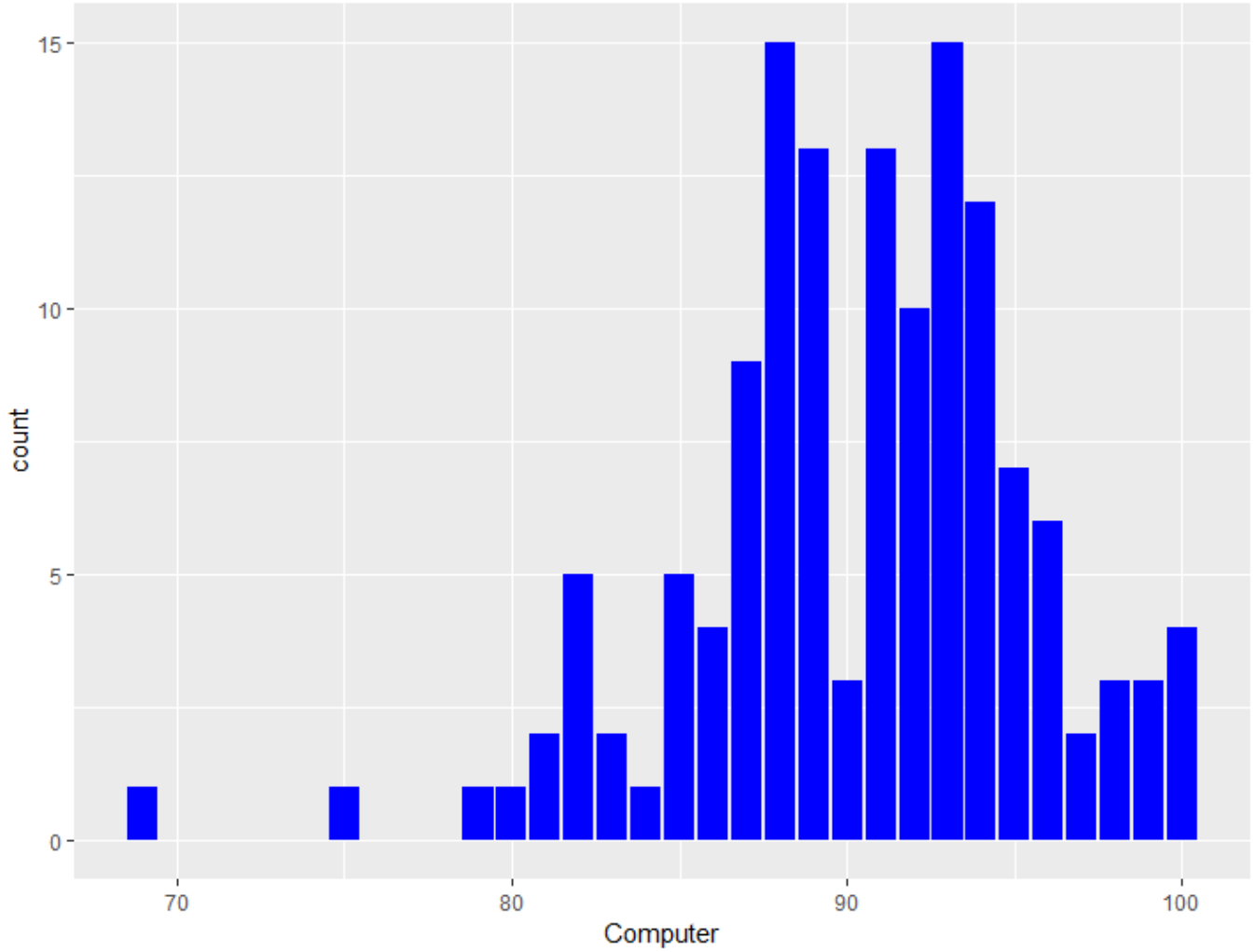
```
ggplot(data=sysu_student, mapping =
aes(x=English)) +
geom_histogram(binwidth = 5,
fill="lightblue", color="black") +
facet_grid(Major ~ ., scales = "free")
```



Major的顺序改变会是如何？

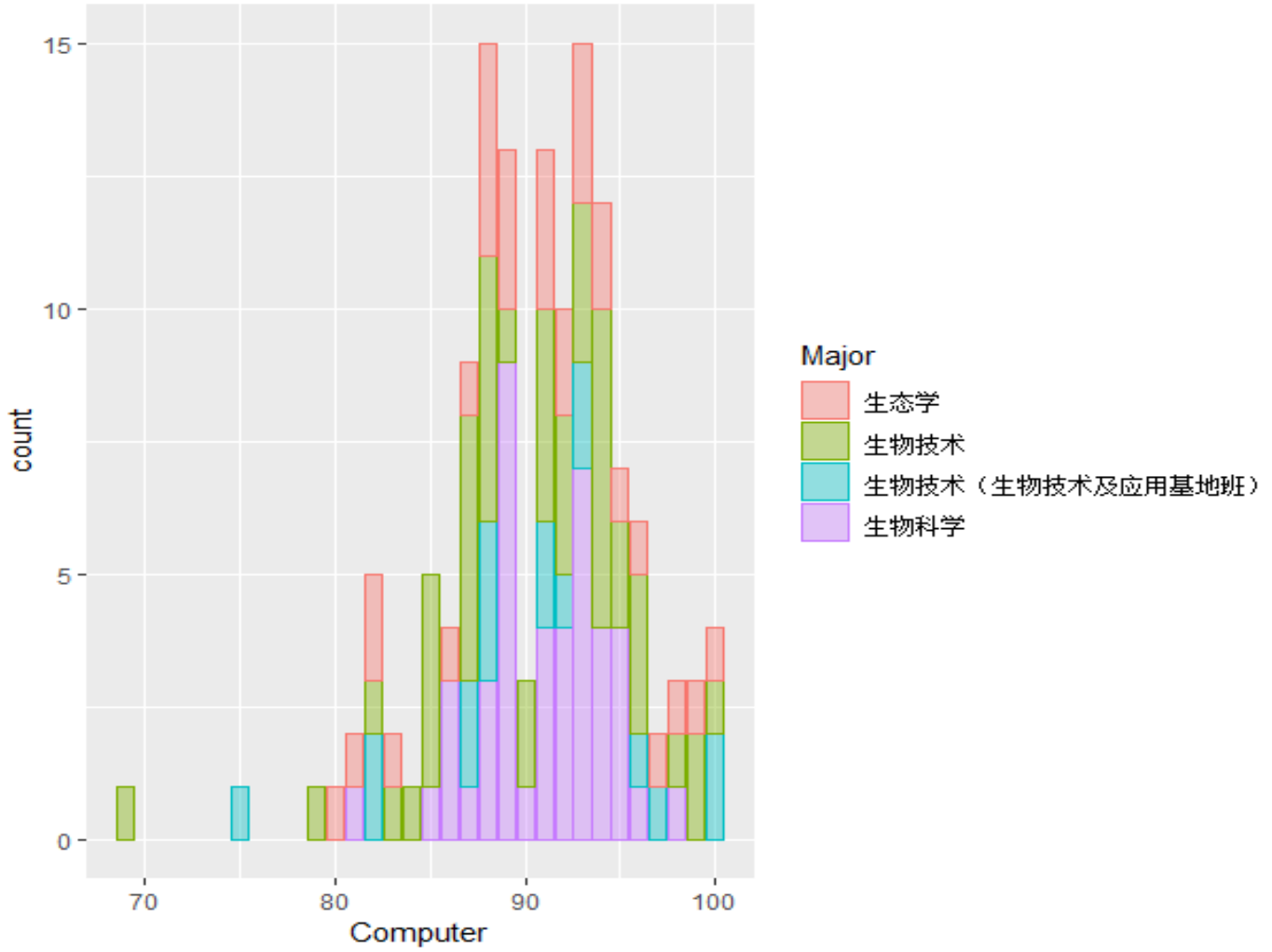
条形图

```
ggplot(data = sysu_student) +

geom_bar(aes(x=Computer),fill="blue")
```



分专业上色

```
ggplot(data = sysu_student) +
geom_bar(aes(x=Computer, color=Major,
fill=Major) , alpha = 0.4,
position="stack")
```



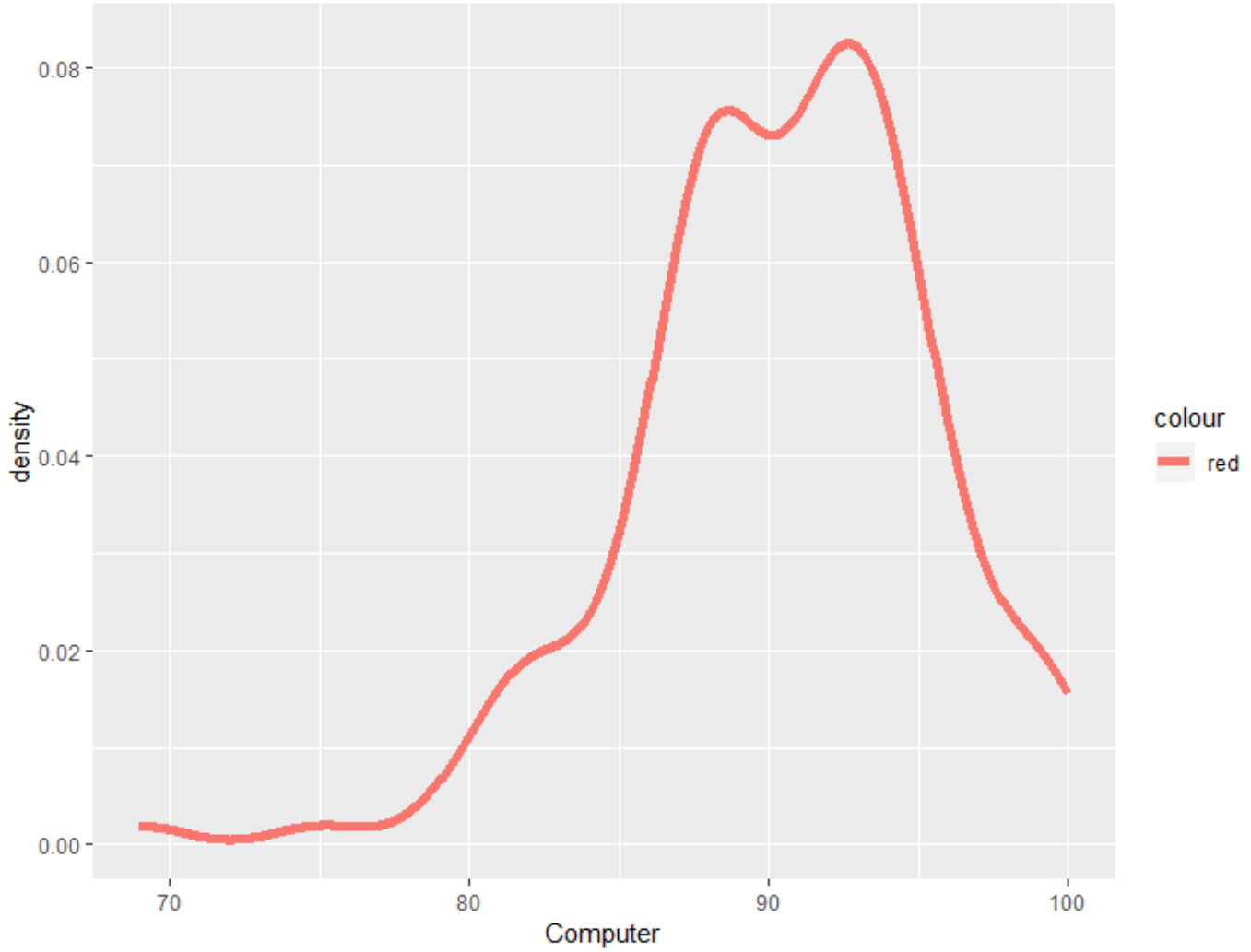
改变position参数会是如何？

图层的位置调整参数

- **dodge：** “避让”方式，即往旁边闪，如柱形图的并排方式就是这种。
- **fill：** 填充方式， 先把数据归一化，再填充到绘图区的顶部。
- **identity：** 原地不动，不调整位置
- **jitter：** 随机抖一抖，让本来重叠的露出点头来
- **stack：** 叠罗汉

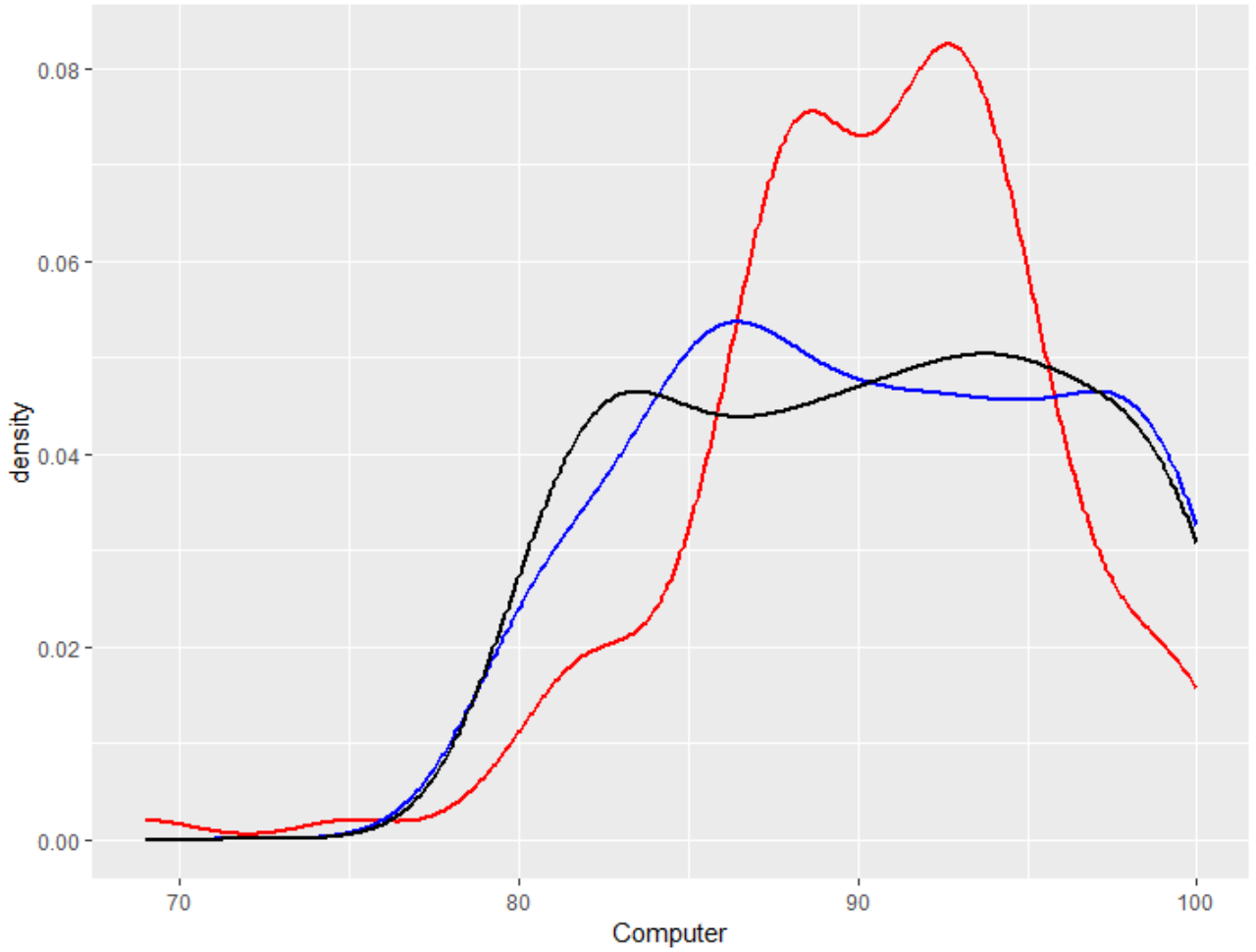
密度图

```
ggplot(data = sysu_student, mapping =

aes(x=Computer,color="red")) +

geom_line(stat="density",size=2)
```



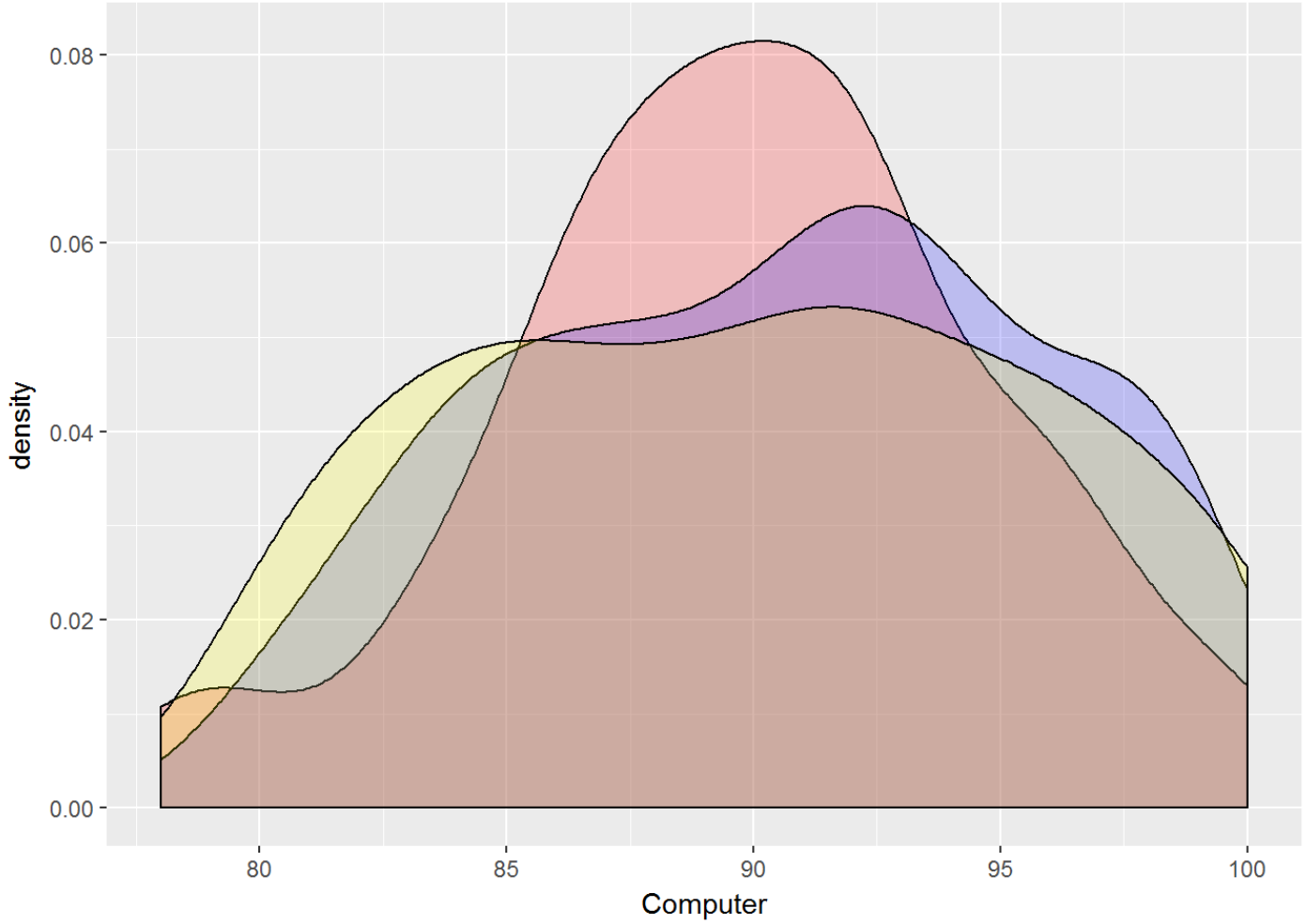
密度图叠加

```
ggplot(data = sysu_student) +
geom_line(aes(x=Computer),stat =
"density", color = "red", size=1)
+geom_line(aes(x=English), stat =
"density", color="blue", size=1) +
geom_line(aes(x=Math), stat =
"density", size=1)
```



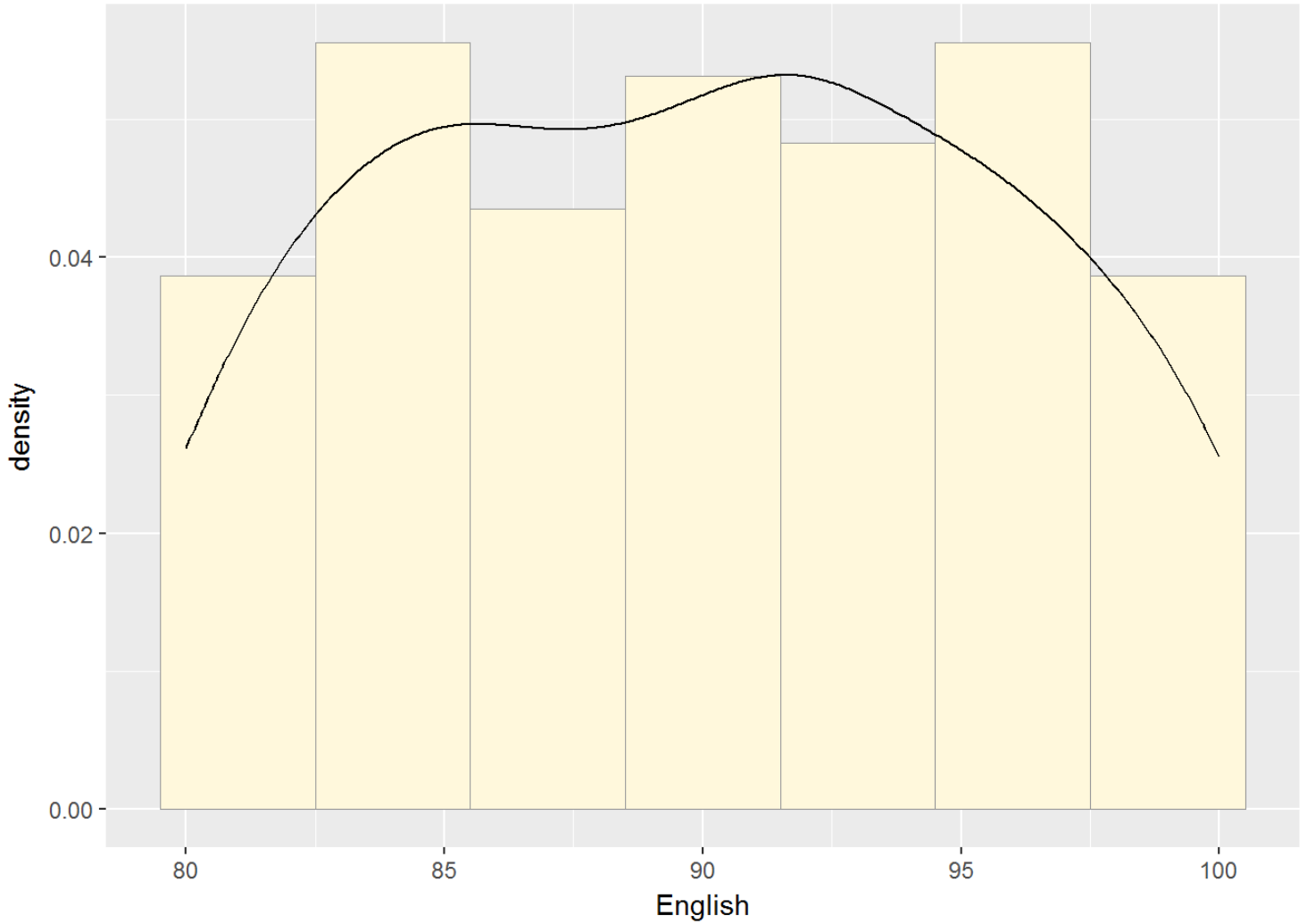
带阴影密度图

```
ggplot(data = sysu_student) +
geom_density(aes(x=Computer),f
ill="red", alpha=.2) +
geom_density(aes(x=Math),fill=
"blue", alpha=.2) +
geom_density(aes(x=English),fi
ll="yellow", alpha=.2)
```

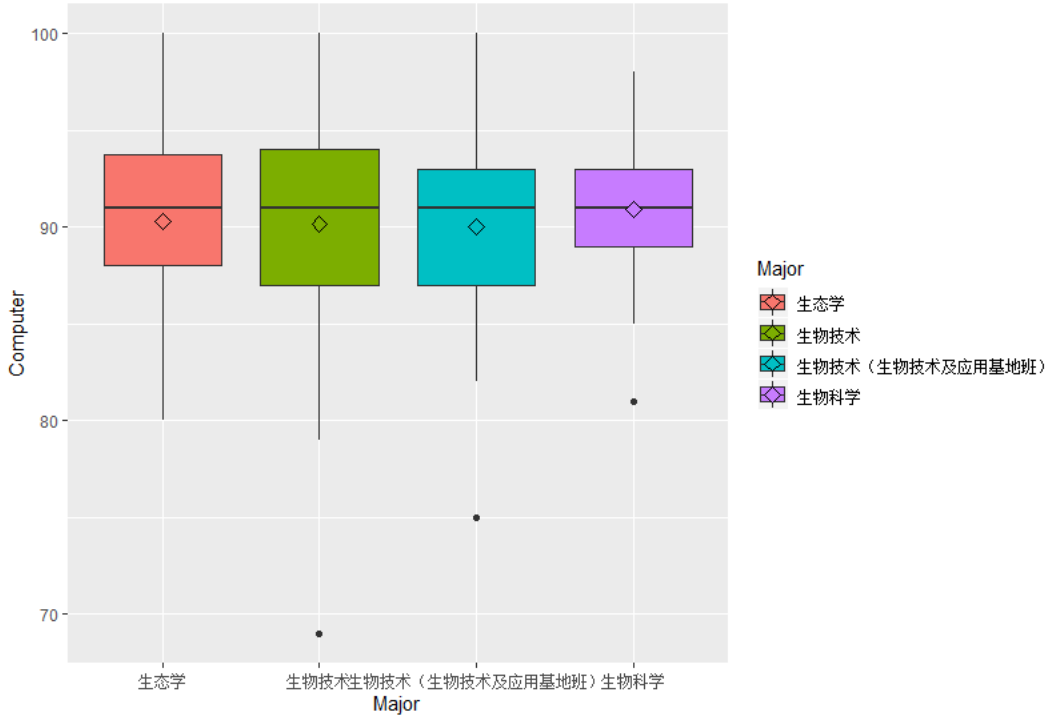


叠加密度图

```
#Density curve overlaid on a histogram
ggplot(data=sysu_student) +
geom_histogram( mapping = aes(x=English,
y = ..density..),
binwidth = 3, fill="cornsilk",
colour="grey60", size=.2) +
geom_line(aes(x=English), stat =
"density")
```



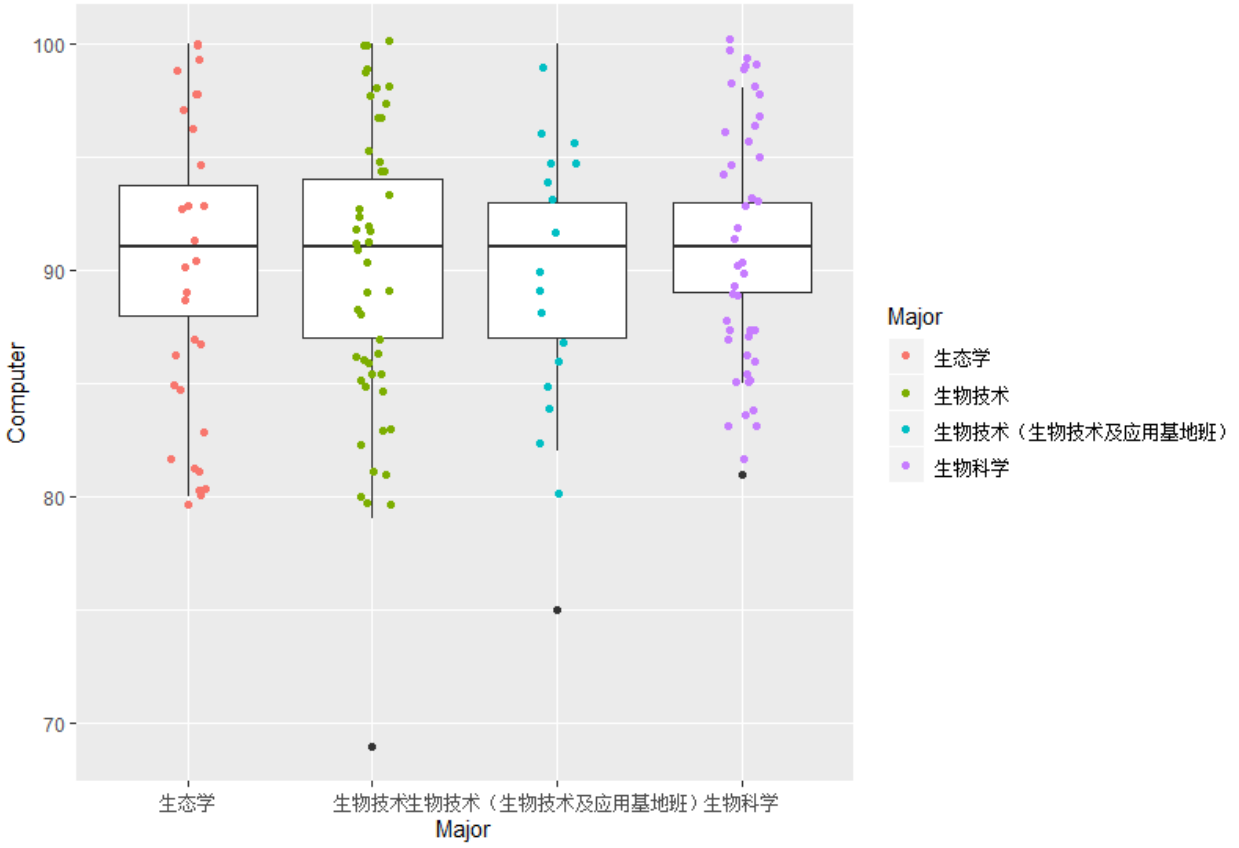
箱线图




```
#Use geom_boxplot(), mapping a continuous variable to y and a discrete variable to x
bp<-ggplot(data = sysu_student, mapping = aes(x=Major, y =Computer, fill=Major)) +
geom_boxplot()
bp + stat_summary(fun.y="mean", geom="point", shape=23, size=3)
```

增加点

```
# One way we can extend this plot is
adding a layer of individual points on
top of it
p.box.jitter <- ggplot(data =
sysu_student, mapping = aes(x=Major, y
=Computer)) + geom_boxplot() +
geom_jitter(aes(x=Major, y=English,
color=Major), width = 0.1)
p.box.jitter
```



分学科箱线图

如何按学科来分组显示？

Converting Data from Wide to Long

长数据有一列数据是变量的类型，有一列是变量的值

```
1. install.packages("reshape2")
2. library(reshape2)
```

	Id	Name	Major	English	Math	Computer	total
1	16336001	阿比达·阿布来提	生物技术	80	92	95	267
2	16336007	蔡静	生物技术	96	91	93	280
3	16336008	蔡奇	生物技术	85	86	99	270
4	16336010	蔡响	生物科学	82	91	85	258
5	16336014	曾思琳	生物科学	95	88	83	266
6	16336019	陈嘉杰	生物科学	97	87	93	277
7	16336025	陈琳琪	生物技术	87	95	88	270
8	16336029	程海涛	生物技术	90	83	100	273
9	16336030	程凯平	生物技术	83	82	94	259

melt函数对宽数据进行处理，得到长数据

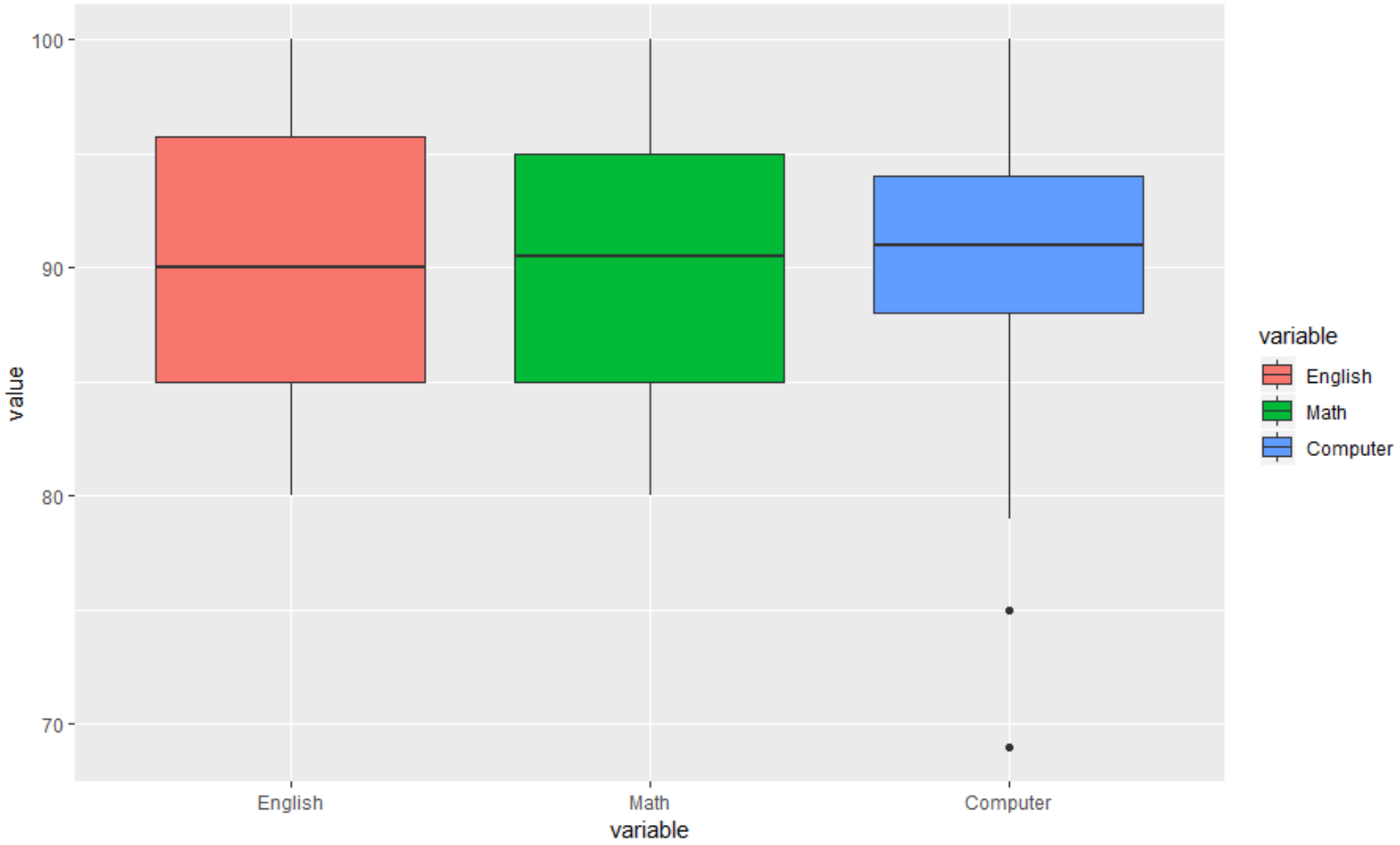
Id	Name	Major	variable	value
16336001	阿比达·阿布来提	生物技术	English	80
16336007	蔡静	生物技术	English	96
16336008	蔡奇	生物技术	English	85
16336010	蔡响	生物科学	English	82
16336014	曾思琳	生物科学	English	95
16336019	陈嘉杰	生物科学	English	97
16336025	陈琳琪	生物技术	English	87
16336029	程海涛	生物技术	English	90
16336030	程凯平	生物技术	English	83
16336031	迟可欣	生物技术	English	94
16336033	代智允	生物技术	English	87
16336035	邓柯	生物技术	English	80

```
sysu_student_long<-melt(sysu_student[,-7], id.vars = c("Id","Name","Major"))
```

思考：如何对长数据进行处理，得到宽数据

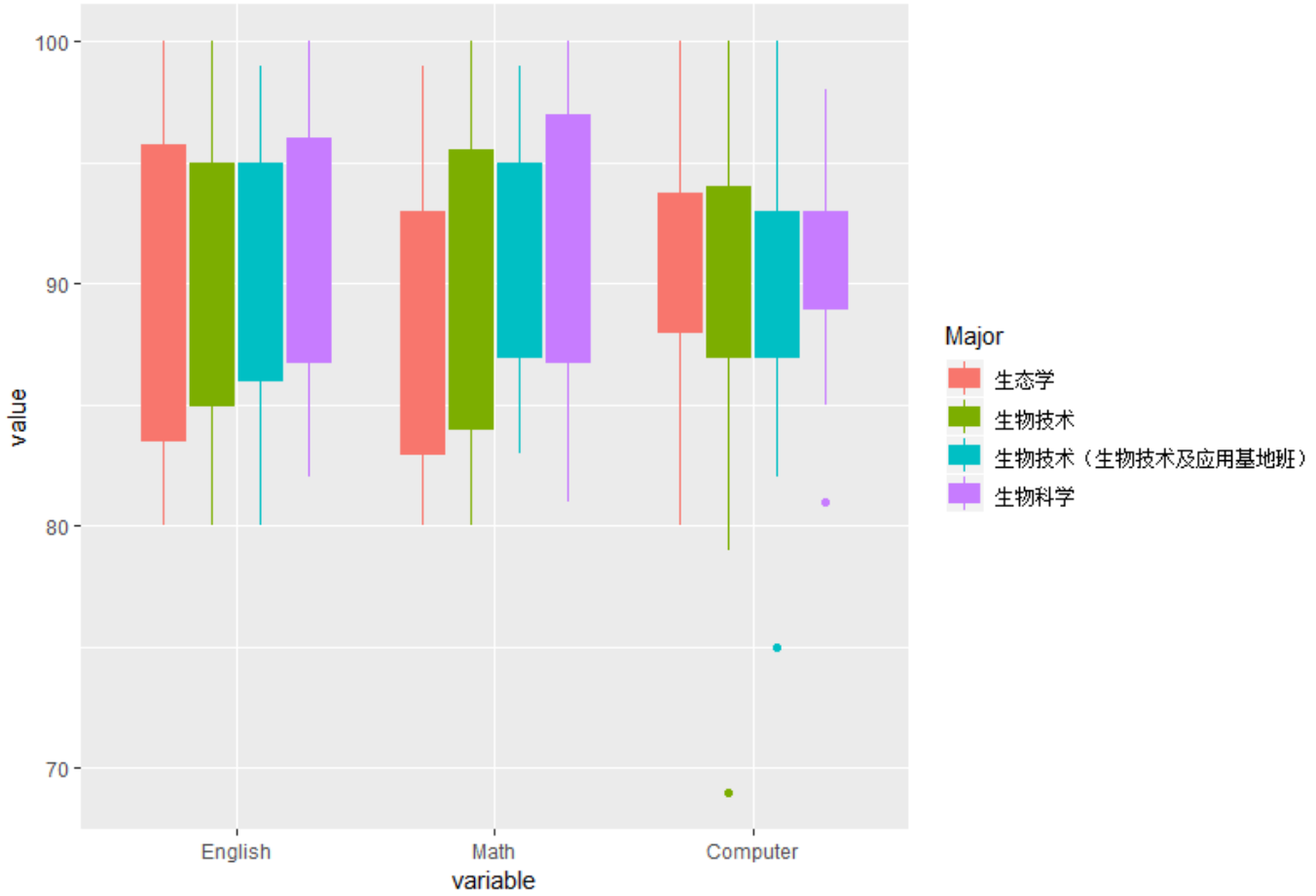
分学科箱线图

```
ggplot(data =
sysu_student_long) +
geom_boxplot(aes(x=variable,y=
value, fill=variable))
```



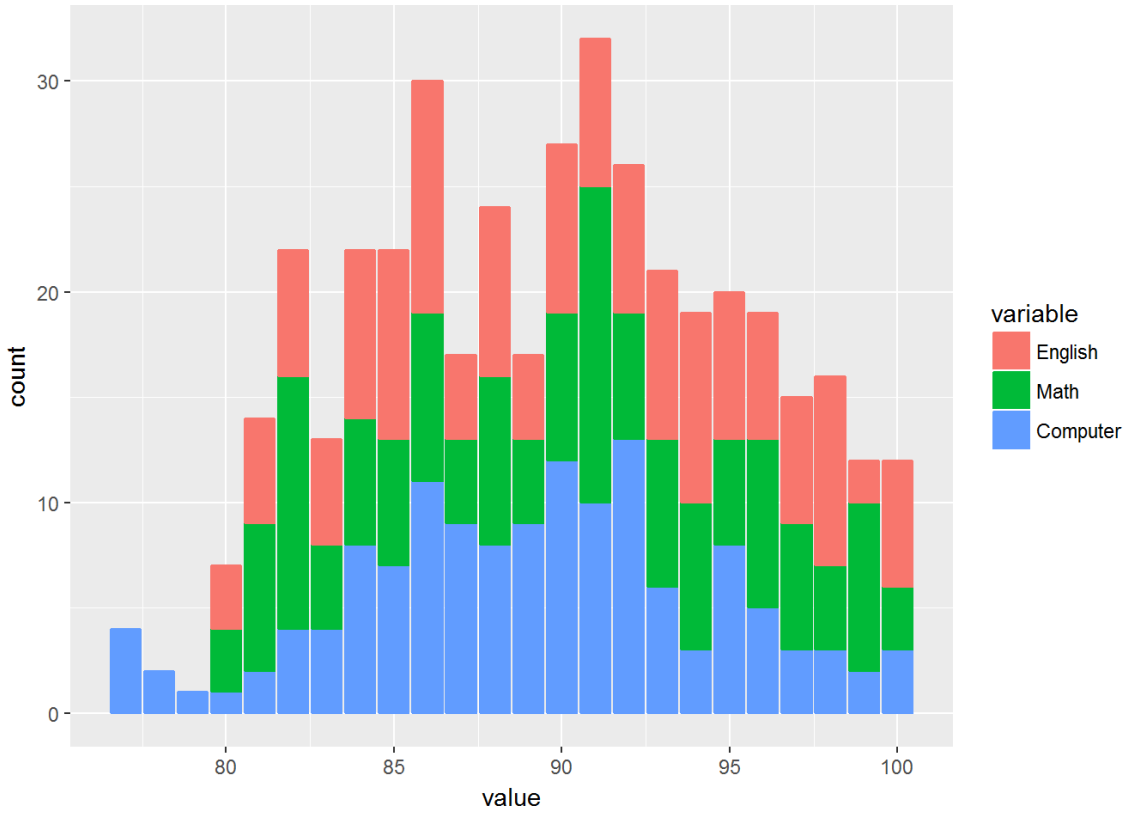
分学科及专业箱线图

```
ggplot(data = sysu_student_long) +
geom_boxplot(aes(x=variable,y=value,
color = Major, fill=Major))
```



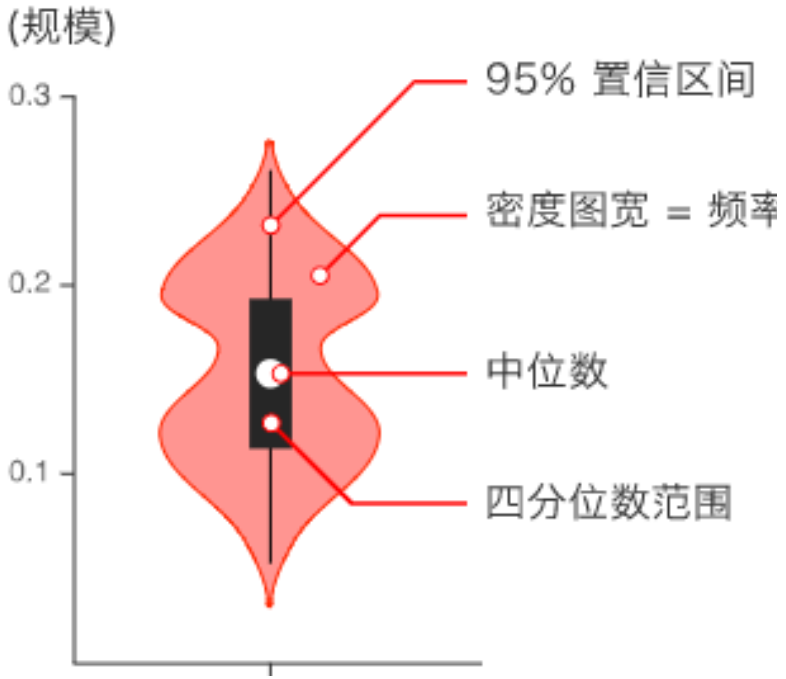
分学科直方图

```
ggplot(data =
sysu_student_long) +
geom_bar(aes(x=value,color =
variable, fill=variable))
```



小提琴图

geom_violin()　　小提琴图 (Violin Plot) 用于显示数据分布及其概率密度。



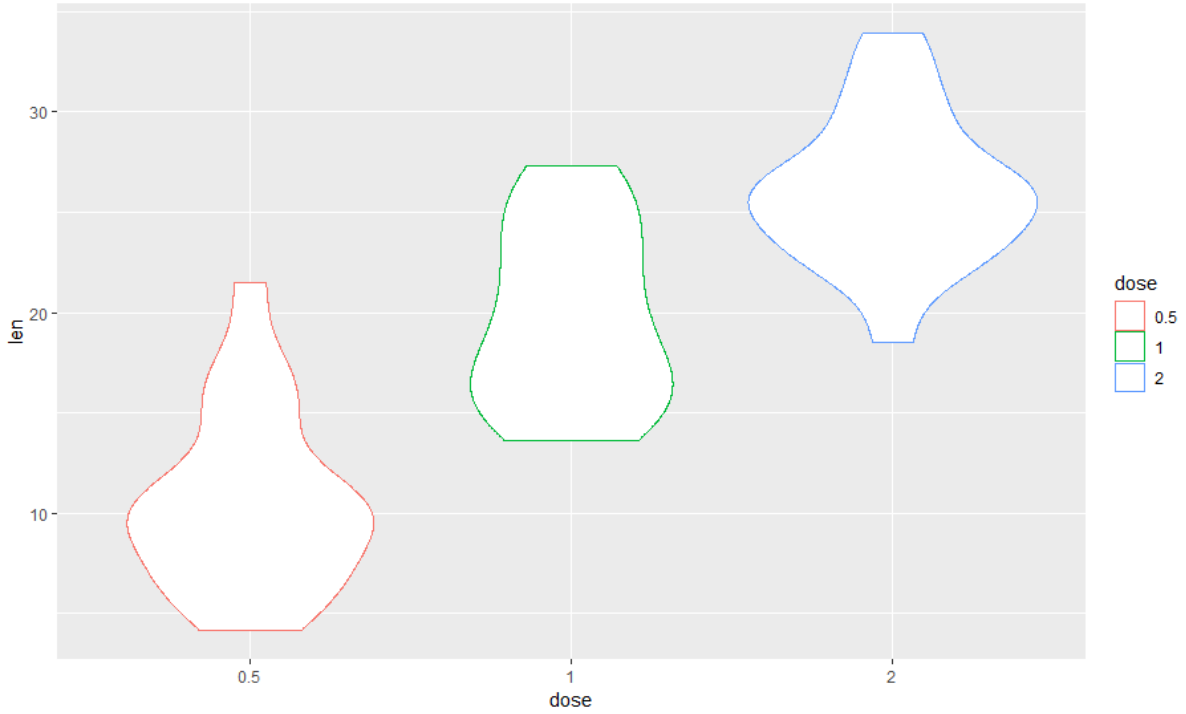
ToothGrowth: The Effect of Vitamin C on Tooth Growth in Guinea Pigs
(Vitamin C在豚鼠牙齿生长的效果)

豚鼠;天竺鼠

小提琴图

geom_violin()　　**小提琴图 (Violin Plot) 用于显示数据分布及其概率密度。**

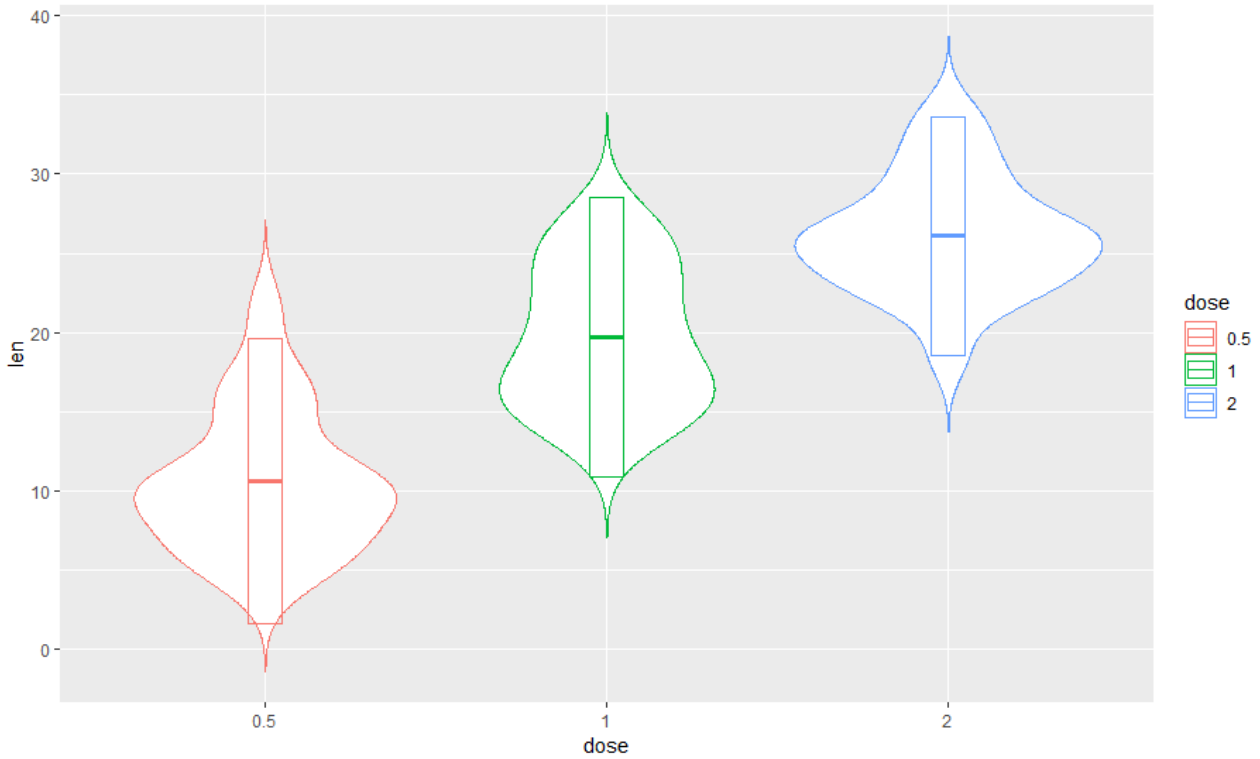
```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)  # 转换dose成factor
ggplot(ToothGrowth, aes(x=dose, y=len, color=dose)) + geom_violin()
```



小提琴图+统计量

geom_violin()　　小提琴图 (Violin Plot) 用于显示数据分布及其概率密度。

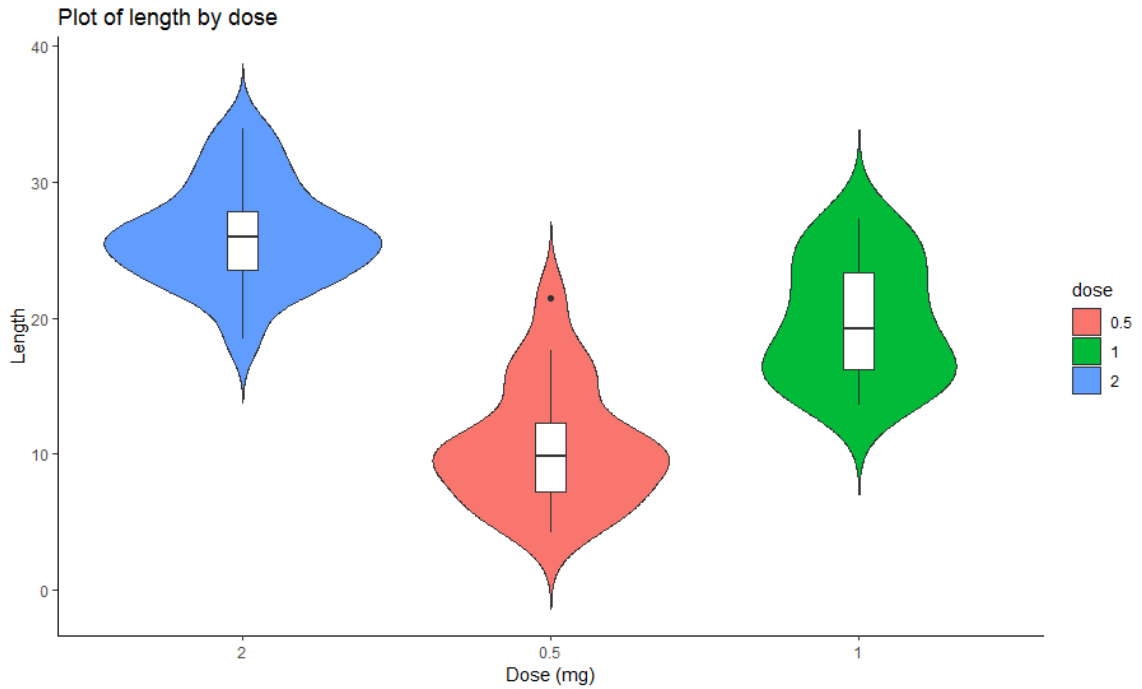
```
p <- ggplot(ToothGrowth, aes(x=dose, y=len, color=dose)) + geom_violin(trim=FALSE)
p + stat_summary(fun.data="mean_sdl",  geom="crossbar", width=0.1 )
```



小提琴图+个性化修改

geom_violin()　　小提琴图 (Violin Plot) 用于显示数据分布及其概率密度。

```
dp <- ggplot(ToothGrowth, aes(x=dose, y=len, fill=dose)) + geom_violin(trim=FALSE)+
geom_boxplot(width=0.1, fill="white")+ labs(title="Plot of length by dose",x="Dose
(mg)", y = "Length") + scale_x_discrete(limits=c("2", "0.5", "1"))
dp + theme_classic()
```




Thank You