使用ggplot2进行数据可视化

Alex / 2020-06-07 / free_learner@163.com / learning-archive.org

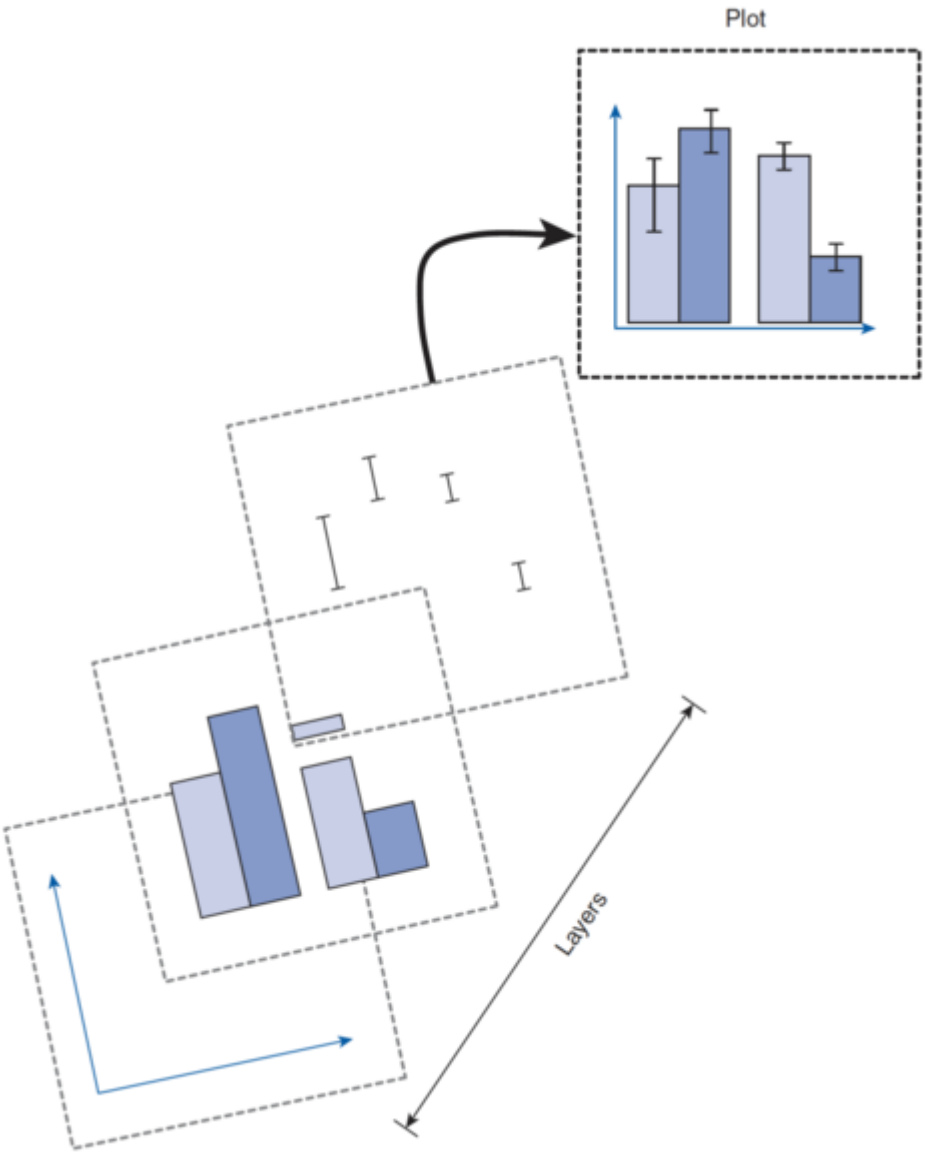
更新于2023-08-29，主要是文字排版上的更新，内容基本保持不变。

使用R中的ggplot2包进行常见的数据可视化的方法，包括箱线图、散点图、柱状图等。

一、ggplot2画图的基本结构

1. 一幅图由多个层（layer）构成

在ggplot2中，一幅图就是多个层叠加在一起显示的效果。如下图所示，坐标轴、柱（bar）和误差线（errorbar）分别用三个层来表示：



图片来源：Field, Andy, Jeremy Miles, and Zoë Field. *Discovering statistics using R*. Sage publications, 2012.

2. 一个层由几何对象（geoms）和美学元素（asethetics）构成

一个层包括几何对象（geometric objects, 比如points, bar, line等）和这些对象的美学元素（asethetics, 比如color, size, style等）构成。

常用的几何对象有: `geom_bar()`, `geom_point()`, `geom_line()`, `geom_smooth()`, `geom_histogram()`, `geom_boxplot()`, `geom_text()`, `geom_density()`, `geom_errorbar()`, `geom_hline()`, `geom_vline()`。

对不同的几何对象，可以设置的美学元素不同。有两种方法设置美学元素，比如color="red"或者aes(color=group)，后一种情况就是根据变量来设置美学元素。

3. 位置（position）参数和主题（theme）函数

通过设置位置参数，可以避免数据重叠在一起。position有5种选项，dodge/stack/fill/identity/jitter；通过主题函数可以设置跟数据无关的一些元素，比如background, gridlines等。

4. 统计量

我们往往希望同时呈现原始数据和一些统计量（比如均值和方差），除了可以提前计算好统计量，也可以使用`stat_summary`函数实现（使用方法见下面的例子）。

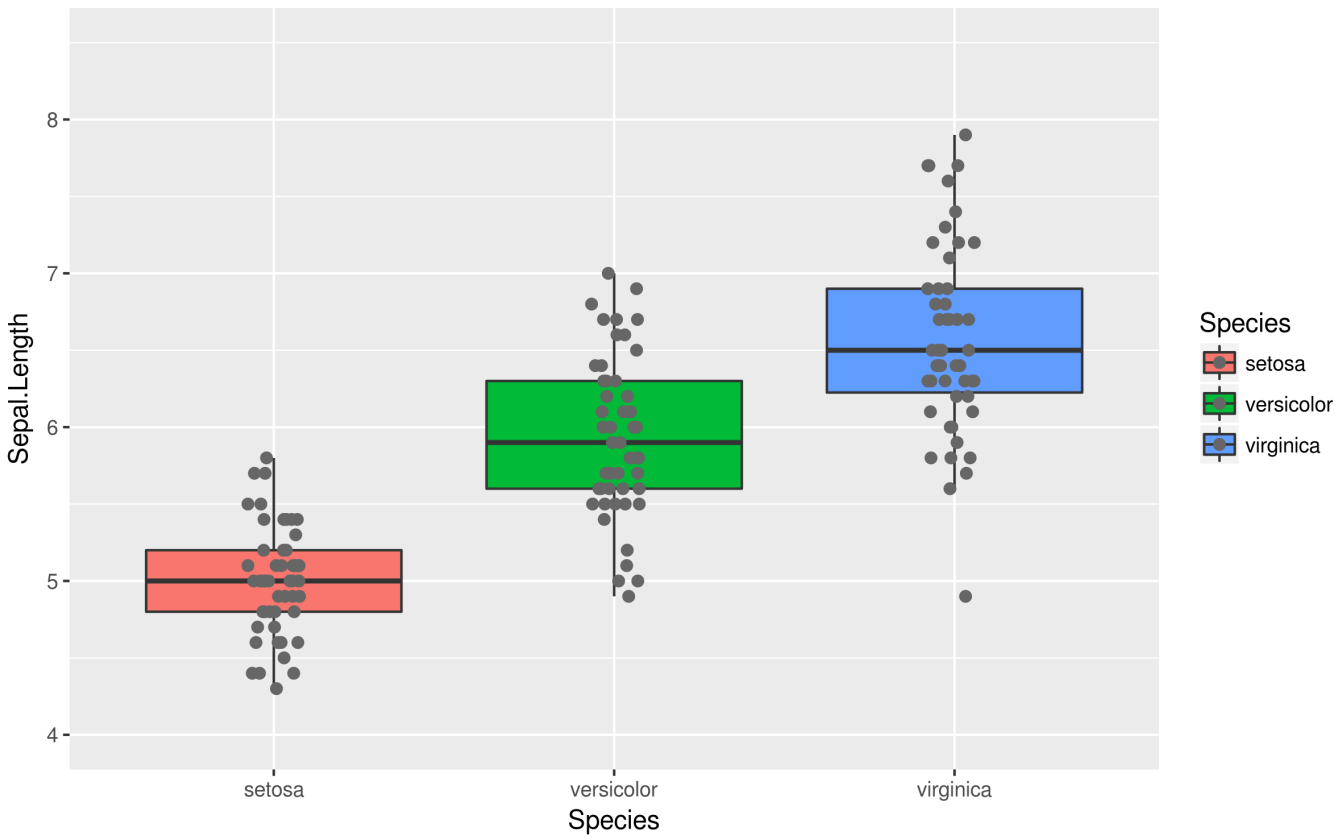
5. 保存图片

ggsave函数可以保存各种常见格式的图片、设置图片大小和分辨率等。

二、各种类型图的实现方法

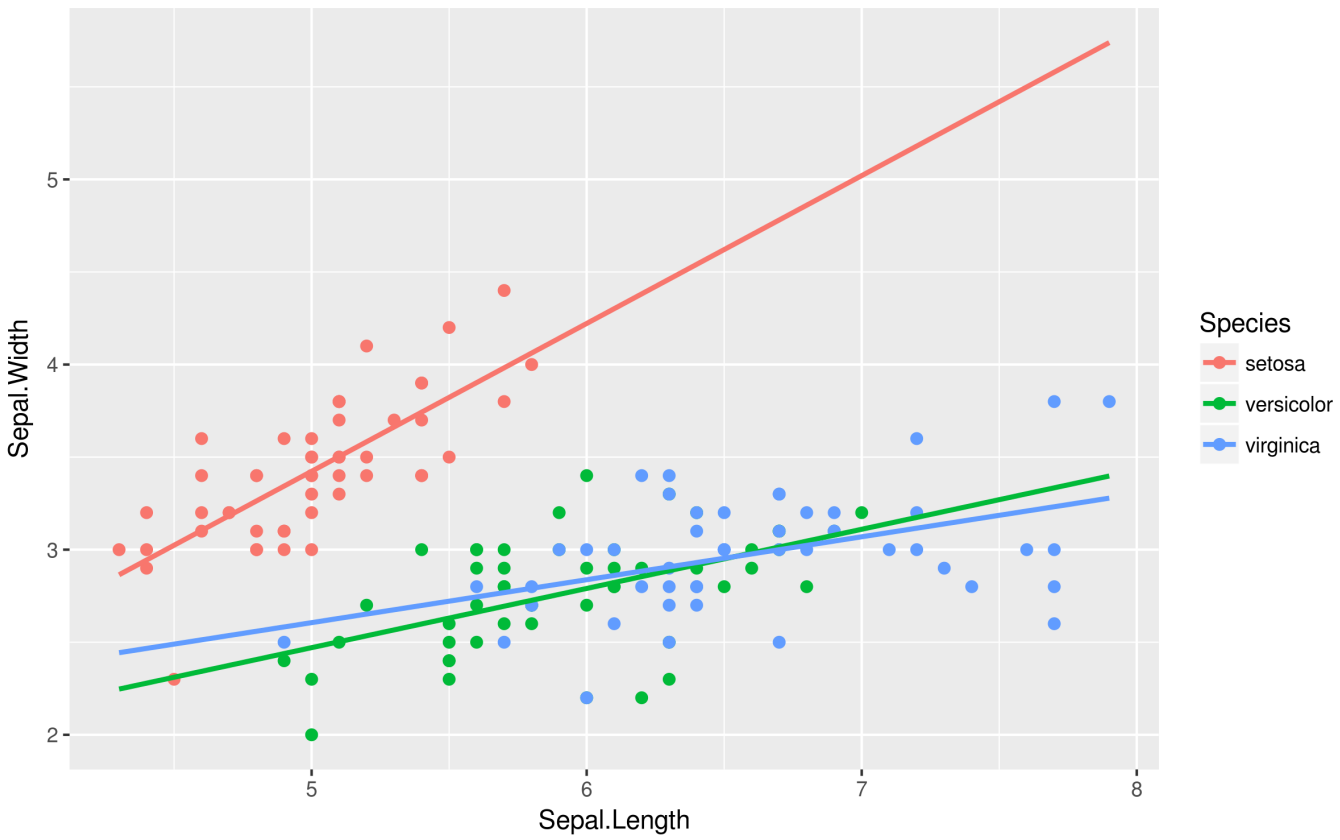
1. 箱线图＋散点图

```
ggplot(iris, aes(x=Species, y=Sepal.Length, fill=Species)) +
  expand_limits(y=c(4,8.5))+
  geom_boxplot(outlier.shape =NA, alpha =1, position=position_dodge(1.2)) +
geom_point(position=position_jitterdodge(dodge.width=0.85),color=rgb(0.4,0.4,0.
4,1), shape=20, size=3)
```



2. 散点图＋最小二乘拟合线

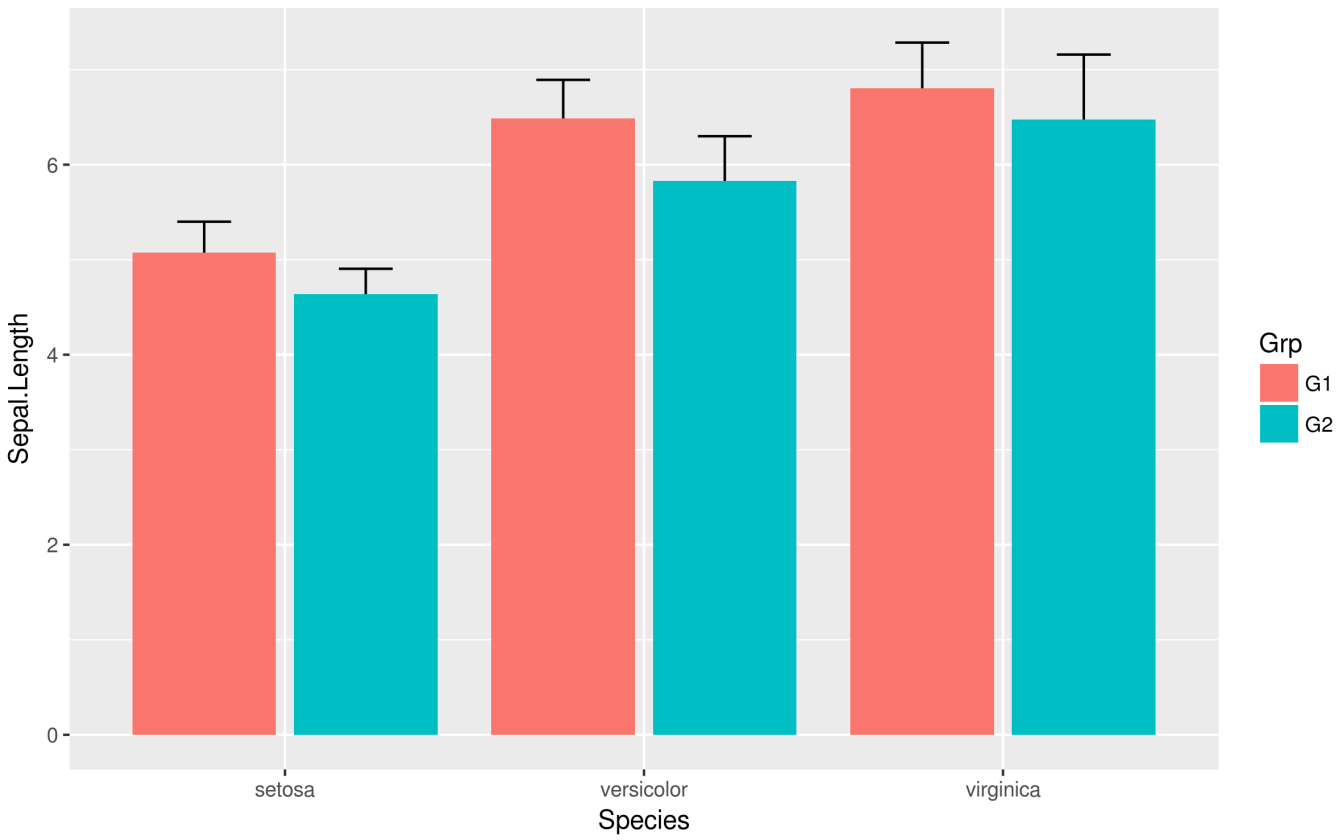
```
ggplot(iris, aes(x=Sepal.Length, y=Sepal.Width, color=Species)) +
  geom_point(shape=20, size=3) +
  geom_smooth(method = "lm", se=FALSE, fullrange = TRUE)
```



3. 柱状图

```
Grp <- as.factor(ifelse(iris$Sepal.Width > 3, 'G1','G2'))
ggplot(iris, aes(x=Species, y=Sepal.Length, fill=Grp)) +
  stat_summary(position = position_dodge(width = 0.9),fun.ymin = function(x)
mean(x)-sd(x), fun.ymax = function(x) mean(x) + sd(x), geom = "errorbar",
width=0.3)+
  stat_summary(fun.y = mean, geom = "bar",position = position_dodge(width =
0.9), width=0.8)
```

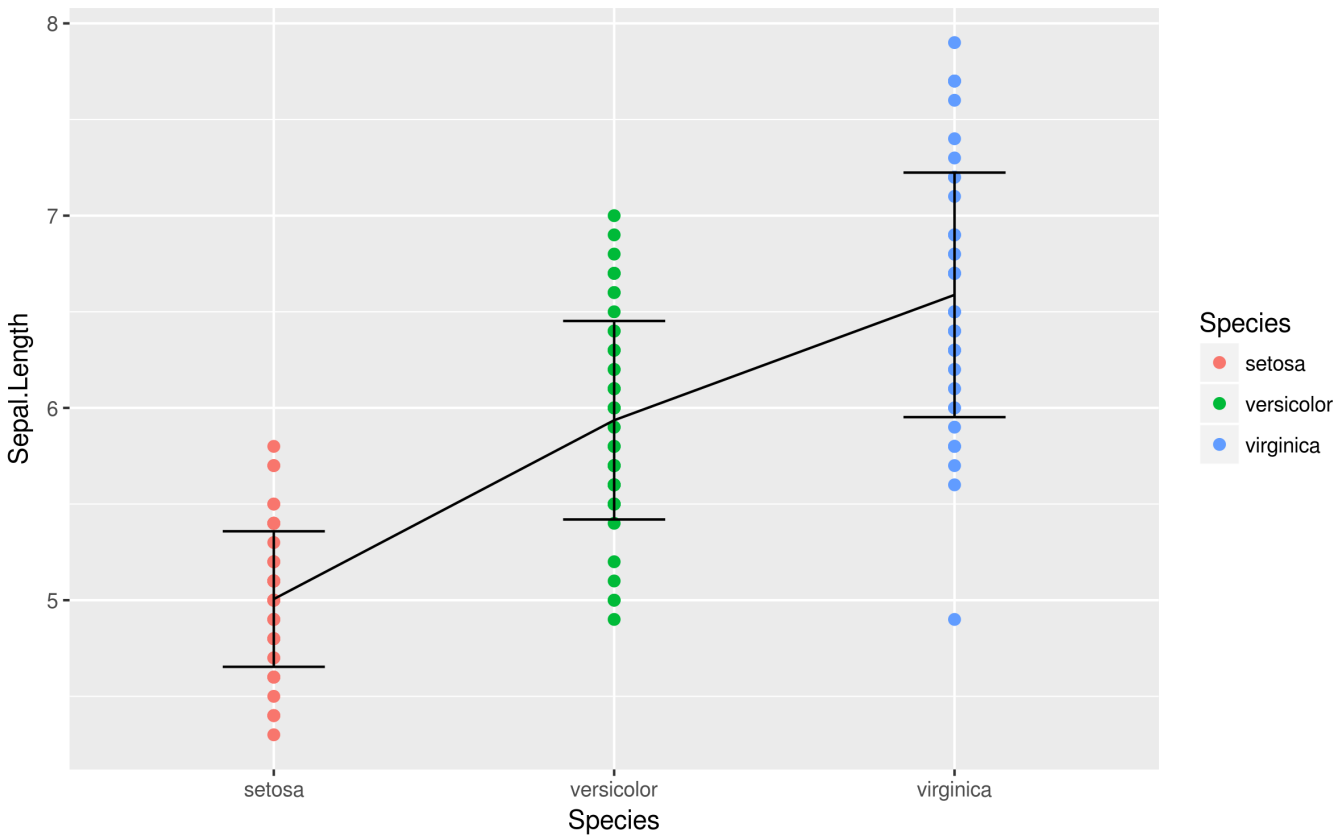
由于iris数据集里只有一个分类变量，这里我自己生成了一个新的分类变量。



4. 散点图＋连接线

```
ggplot(iris, aes(x=Species, y=Sepal.Length, group=1))+
  geom_point(aes(color=Species), shape=20, size=3)+
  stat_summary(fun.ymin = function(x) mean(x) - sd(x), fun.ymax = function(x)
mean(x) + sd(x), geom = "errorbar", width=0.3) +
  stat_summary(fun.y = mean, geom = "line")
```

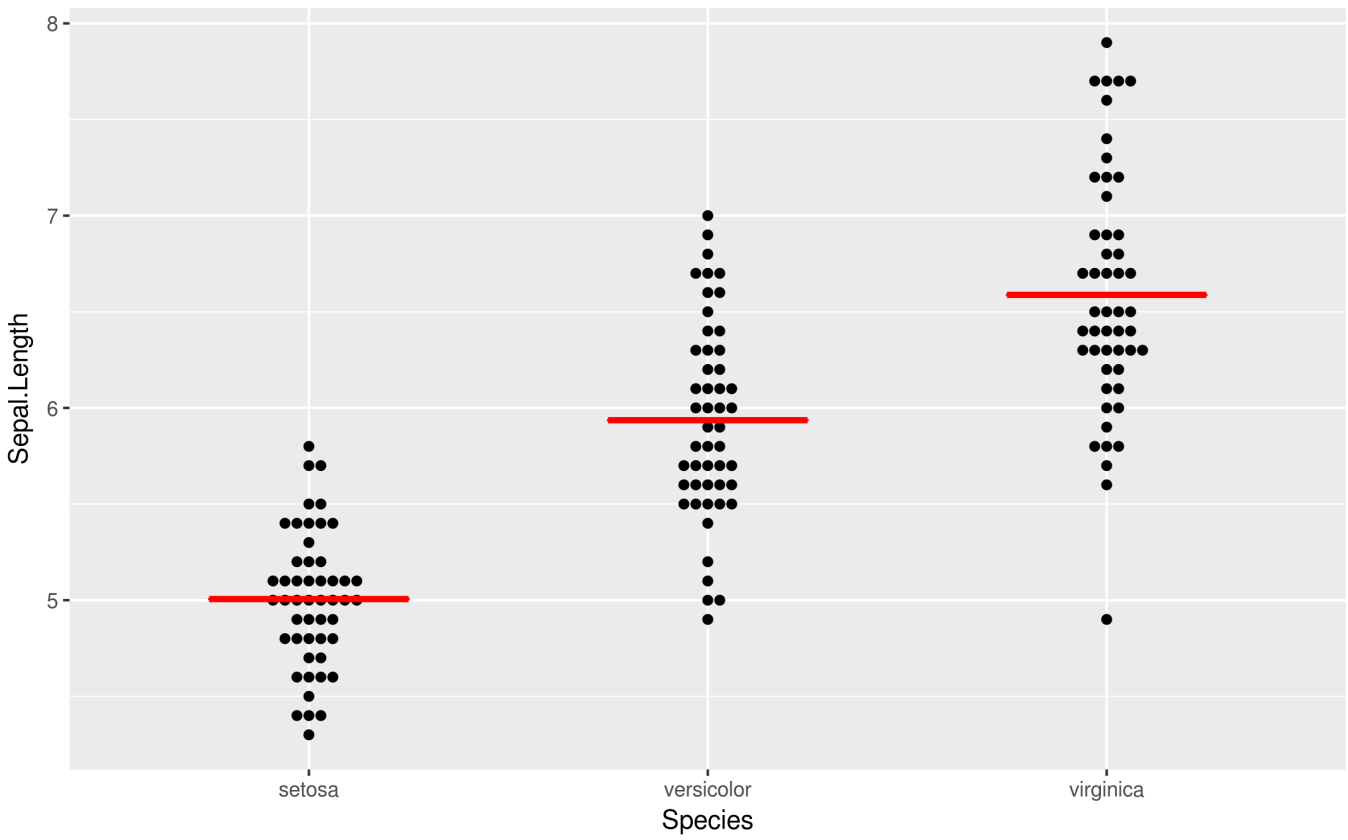
这里的连接线就是将每组的均值连起来，以表示一种趋势，适用于分类变量是顺序变量的情况。



5. 蜂群图（Beeswarm plot）

蜂群图是一种散点图，可以避免数据重叠而且可以显示数据分布密度，因此相当于结合了散点图和小提琴图（violin plot）。除了ggplot2，还需要ggbeeswam包。代码如下：

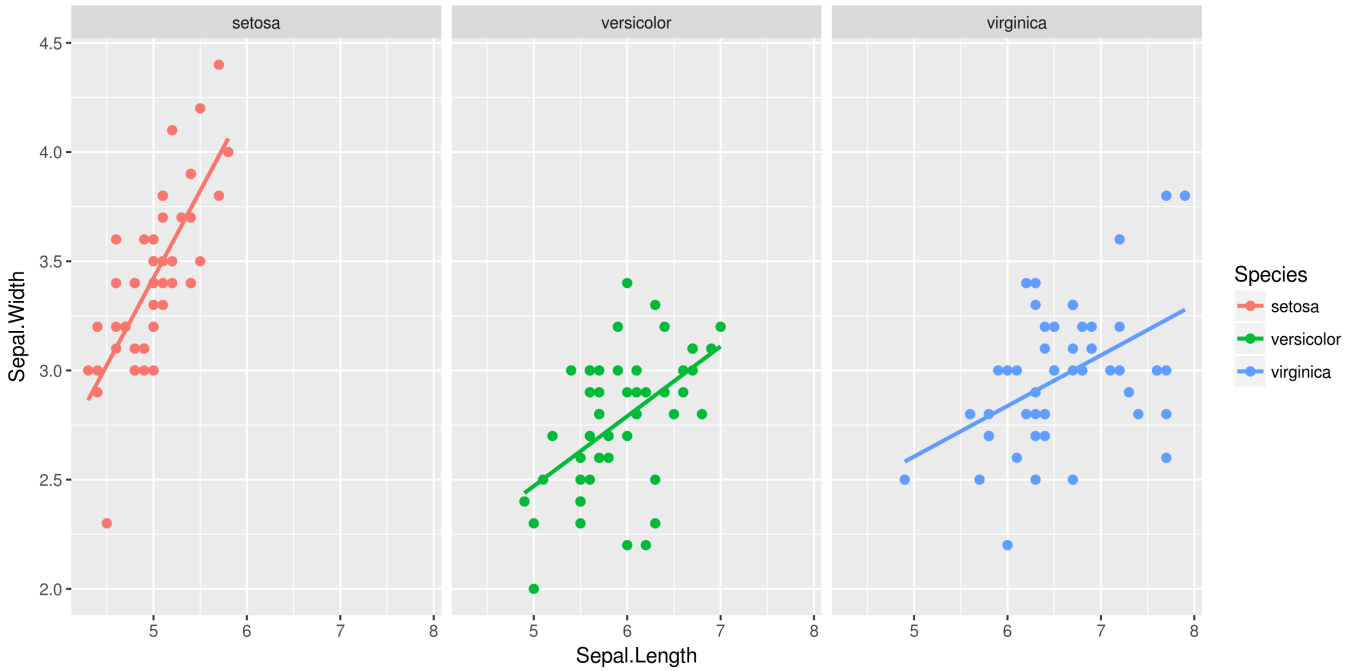
```
ggplot(iris, aes(Species, Sepal.Length)) +
  geom_beeswarm()+
  stat_summary(fun.y= mean, fun.ymin=mean, fun.ymax=mean, geom="crossbar",
width=0.5, color="red")
```



6. multi-facet图

multi-facet图指的是一幅图由多张子图构成：

```
ggplot(iris, aes(x=Sepal.Length, y=Sepal.Width, color=Species))+
  geom_point(shape=20, size=3)+
  geom_smooth(method = "lm", se=FALSE)+
  facet_grid(.~Species)
```



三、其他参数设置

要实现自己理想中的可视化效果，往往需要做很多微调，以下是我自己用到过的一些参数（或函数）：

1. 查看配色

有时候为了保持配色一致，需要查看使用的颜色，可以使用`ggplot_build`实现：

```
p <- ggplot() + ...
ggplot_build(p)$data
```

2. 设置坐标轴等

```
+scale_x_continuous(expand=c(0,0)) ## 调整与x轴的间距
+scale_y_continuous(expand=c(0,0)) ## 调整与y轴的间距
+scale_x_discrete(limits=c('G2','G1','G3')) ## 改变x轴标签的顺序
+scale_x_discrete(breaks=c("G1","G2","G3"),labels=c("Control","Treat1","Treat2"
)) ## 设置标签文本
+theme(axis.ticks.x=element_blank(), axis.text.x=element_blank()) ## 去掉x轴刻度
和标签文本
+xlab(NULL)+ylab(NULL) ## 去掉x轴和y轴的标题
+theme(legend.position="none") ## 去掉图例
+theme(aspect.ratio=0.5) ## 调整长宽比例
+theme_class() ## 去掉ggplot2画图中的网格线、背景色、右和顶部的边框
```