Python 数据可视化分析

金林

jinlin@zuel.edu.cn

中南财经政法大学统计与数学学院

2020-01

1 特殊统计图的绘制

2 SEABORN 统计绘图

3 GGPLOT 绘图系统

数据可视化简介

1. 数据可视化旨在借助于图形化手段，清晰有效地传达与沟通信息，
2. 但是，这并不意味着数据可视化就一定因为要实现其功能而令人感到枯燥乏味，或者为了为了看上去绚丽多彩而显得极端复杂。
3. 为了有效地传达思想观念，美学形式与功能需要齐头并进，通过直观地传达关键的方面与特征，来实现对于相当稀疏而又复杂的数据集的深入洞察。
4. 避免没有把握好设计与功能之间的平衡，从而设计出华而不实的数据可视化形式，无法达到其主要目的，也就是传达与沟通信息。
5. 数据可视化与信息图形、信息可视化、科学可视化及统计图形关系密切。
6. “数据可视化”术语实现了成熟的科学可视化领域与较年轻的信息可视化领域的统一。

1 特殊统计图的绘制

2 SEABORN 统计绘图

3 GGPLOT 绘图系统

初等函数图

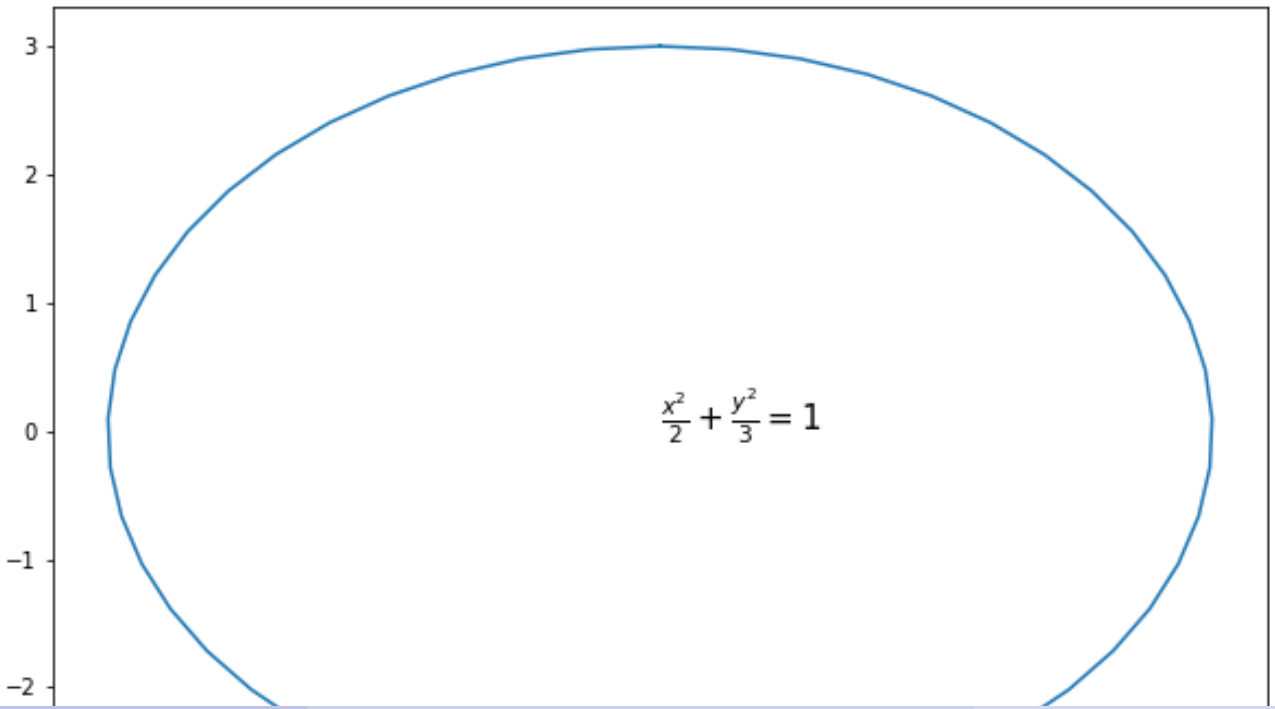
```
import math
import numpy as np
import matplotlib.pyplot as plt
x=np.linspace(0,2*math.pi);x
#fig,ax=plt.subplots(2,2,figsize=(15,12))
```

```
plt.plot(x,np.sin(x))
plt.plot(x,np.cos(x))
plt.plot(x,np.log(x))
plt.plot(x,np.exp(x))
```

极坐标图（加公式）

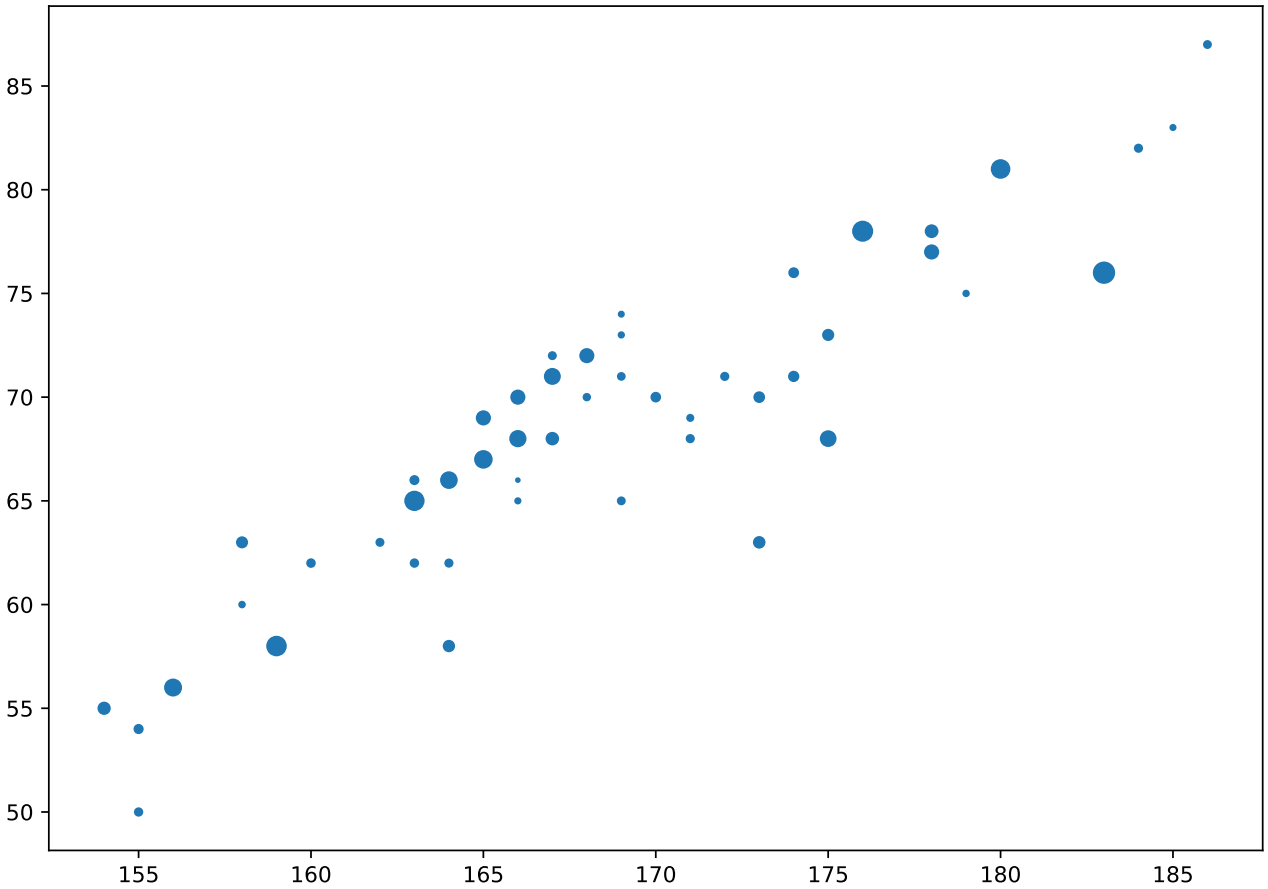
根据函数式的基本绘图，在直角坐标系下可使用参数方程：

```
t=np.linspace(0,2*math.pi)
x=2*np.sin(t)
y=3*np.cos(t)
plt.plot(x,y)
plt.text(0,0,r'$\frac{x^2}{2}+\frac{y^2}{3}=1$',fontsize=15)
```



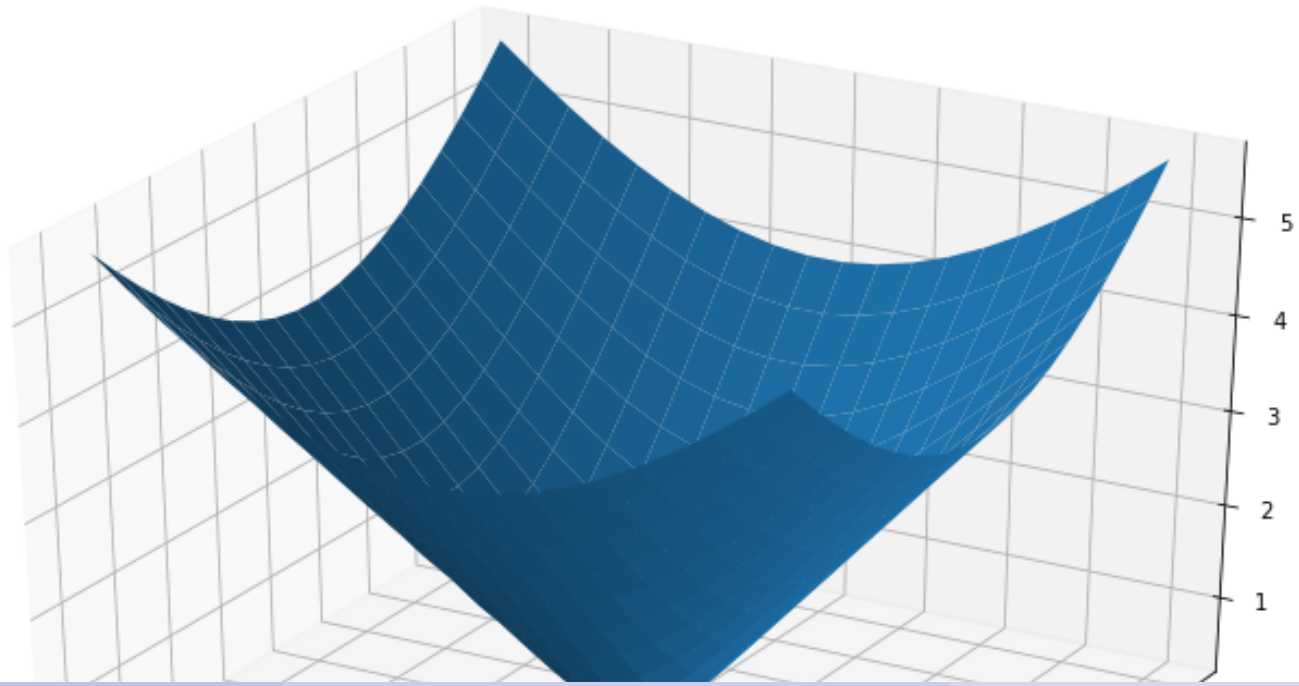
气泡图

```
1 import pandas as pd
2 BSdata = pd.read_csv('../data/BSdata.csv',encoding="utf-8")
3 plt.scatter(BSdata['身高'], BSdata['体重'], s=BSdata['支出'])
```



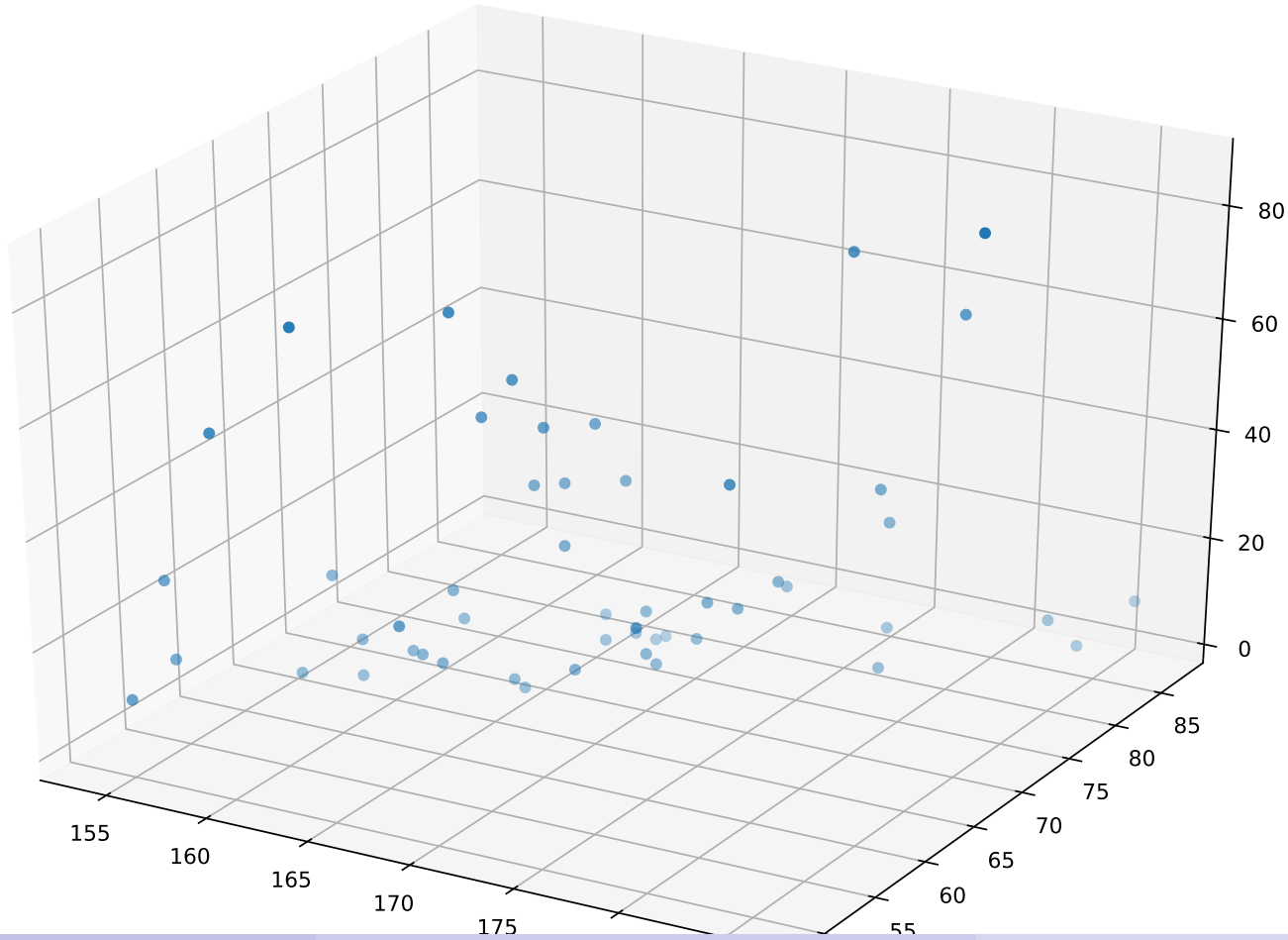
三维曲面图

```
from mpl_toolkits.mplot3d import Axes3D
fig = plt.figure()
ax = Axes3D(fig)
X=np.linspace(-4,4,20) #X = np.arange(-4, 4, 0.5);
Y=np.linspace(-4,4,20) #Y = np.arange(-4, 4, 0.5)
X, Y = np.meshgrid(X, Y)
Z = np.sqrt(X**2 + Y**2)
ax.plot_surface(X, Y, Z);
```



三维散点图

```
from mpl_toolkits.mplot3d import Axes3D
fig = plt.figure()
ax = Axes3D(fig)
ax.scatter(BSdata['身高'], BSdata['体重'], BSdata['支出'])
```



1 特殊统计图的绘制

2 SEABORN 统计绘图

- 基本概念
- 常用统计图

3 GGPLOT 绘图系统

seaborn 简介

1. seaborn 在 matplotlib 的基础上进行了更高级的 API 封装，从而使得作图更加容易，
2. 在大多数情况下，使用 seaborn 就能作出相当具有吸引力的图，而使用 matplotlib 能制作具有更多特色的图。
3. 应该把 seaborn 视为 matplotlib 的补充，而不是替代物。
4. seaborn 针对统计绘图较为方便。

```
1 import seaborn as sns
```

5. 一般来说，seaborn 能满足数据分析 90% 的统计绘图需求。
6. 如果需要复杂的自定义图形，则要用 matplotlit。

- 基本概念
- 常用统计图

分组绘图

1. 比如，需要在一张图上绘制两条曲线，分别是南方和北方的气温变化，用不同的颜色加以区分，这就是分组绘图。
2. 在 seaborn 中用 hue 参数控制分组绘图。

分面绘图

1. 分面绘图其实就是在一张纸上划分不同的区域，比如 2*2 的子区域，在不同的子区域绘制不同的图形，
2. 在 matplotlib 中就是 add-subplot (2 , 2 , 1),
3. 在 seaborn 中用 col 参数控制，col 的全称是 columns 是，不是 color，如果辅助 col-wrap 参数，效果会更好。
4. col 可以控制 columns 的子图，row 可以控制 rows 的子图排列。
5. 如果需要分面绘图，则应该使用 seaborn 的 FacetGrid 对象，seaborn 的一般绘图函数是没有“分面”这个参数的。

统计函数绘图

1. 分组绘图的时候，会对分组变量先用统计函数进行处理，然后绘图，
2. 比如先计算变量的均值，然后绘制该均值的直方图。
3. 统计绘图参数是 estimator，很多情况下默认为 numpy.mean。如果不适用，就需要先用 pandas 进行 groupby 分组汇总，然后用 seaborn 绘图。

- 基本概念
- 常用统计图

箱线图（boxplot）

1. 竖着放的箱线图，也就是将 x 换成 y。
2. 分组绘制箱线图，分组因子是“性别”，在 x 轴不同位置绘制。

```
# 绘制箱线图
sns.boxplot(x=BSdata['身高'])
# 竖着放的箱线图，也就是将x换成y
sns.boxplot(y=BSdata['身高'])
# 分组绘制箱线图，分组因子是性别，在x轴不同位置绘制
```

```
sns.boxplot(x='性别', y='身高',data=BSdata)
# 分组箱线图，分子因子是smoker，不同的因子用不同颜色区分，相当于分组之后又分组
```

```
sns.boxplot(x='开设', y='支出',hue='性别',data=BSdata)
```

小提琴图（violinplot）

```
sns.violinplot(x='性别', y='身高',data=BSdata)
```

```
sns.violinplot(x='开设', y='支出',hue='性别',data=BSdata)
```

点图 (stripplot)

```
sns.stripplot(x='性别', y='身高',data=BSdata)
```

```
sns.stripplot(x='性别', y='身高',data=BSdata,jitter=True)
```

```
sns.stripplot(y='性别', x='身高',data=BSdata,jitter=True)
```

条图 (barplot)

```
1 sns.barplot(x='性别', y='身高',data=BSdata,ci=0,palette=”Blues_d”)
```

计数图（countplot）

```
sns.countplot(x='性别',data=BSdata)
```

```
sns.countplot(y='开设',data=BSdata)
```

```
sns.countplot(x='性别',hue="开设",data=BSdata)
```

分组关系图（catplot）

```
sns.catplot(x='性别',col="开设", col_wrap=3,data=BSdata, kind="count",
    height=2.5, aspect=.8)
```

概率分布图（distplot）

1. 概率分布图包括单变量核密度曲线、直方图、双变量与多变量的联合直方图和密度图。
2. 针对单变量，使用 seaborn 的 distplot() 函数，它集合了 matplotlib 的 hist() 与核函数估计 kdeplot 的功能。
3. kde 控制是否画 kde 曲线，bins 是分组数，rug 控制是否画样本点。
4. 针对双变量，使用 seaborn 中的 jointplot() 函数。
5. 针对多变量，使用 seaborn 中的 pairlot() 函数，默认对角线为直方图（histgram），非对角线为散点图。

```
sns.distplot(BSdata['身高'], kde=True, bins=20, rug=True);
sns.jointplot(x='身高', y='体重', data=BSdata);
sns.pairplot(BSdata[['身高','体重','支出']]);
```

1 特殊统计图的绘制

2 SEABORN 统计绘图

3 GGPLOT 绘图系统

- qplot 快速制图
- ggplot 基本绘图

ggplot 简介

1. ggplot 是用于绘图的 python 扩展包，其理念根植于 Grammay of Graphics 一书。
2. 它将绘图视为一种映射，即从数学空间映射到图形元素空间。例如，将不同的数值映射到不同的色彩或透明度。
3. 该绘图包的特点在于，并不去定义具体的图形（如直方图、散点图），而是定义各种底层组件（如线条、方块）来合成复杂的图形，这使它能以非常简洁的函数构建各类图形，而且默认条件下的绘图品质就能达到印刷精度。
4. cmd 中运行以下代码安装

```
conda install -c conda-forge ggplot
```

- qplot 快速制图
- ggplot 基本绘图

qplot 函数

1. 对于初学者，ggplot 提供了 qplot 函数，可以快捷地绘制多种图形。

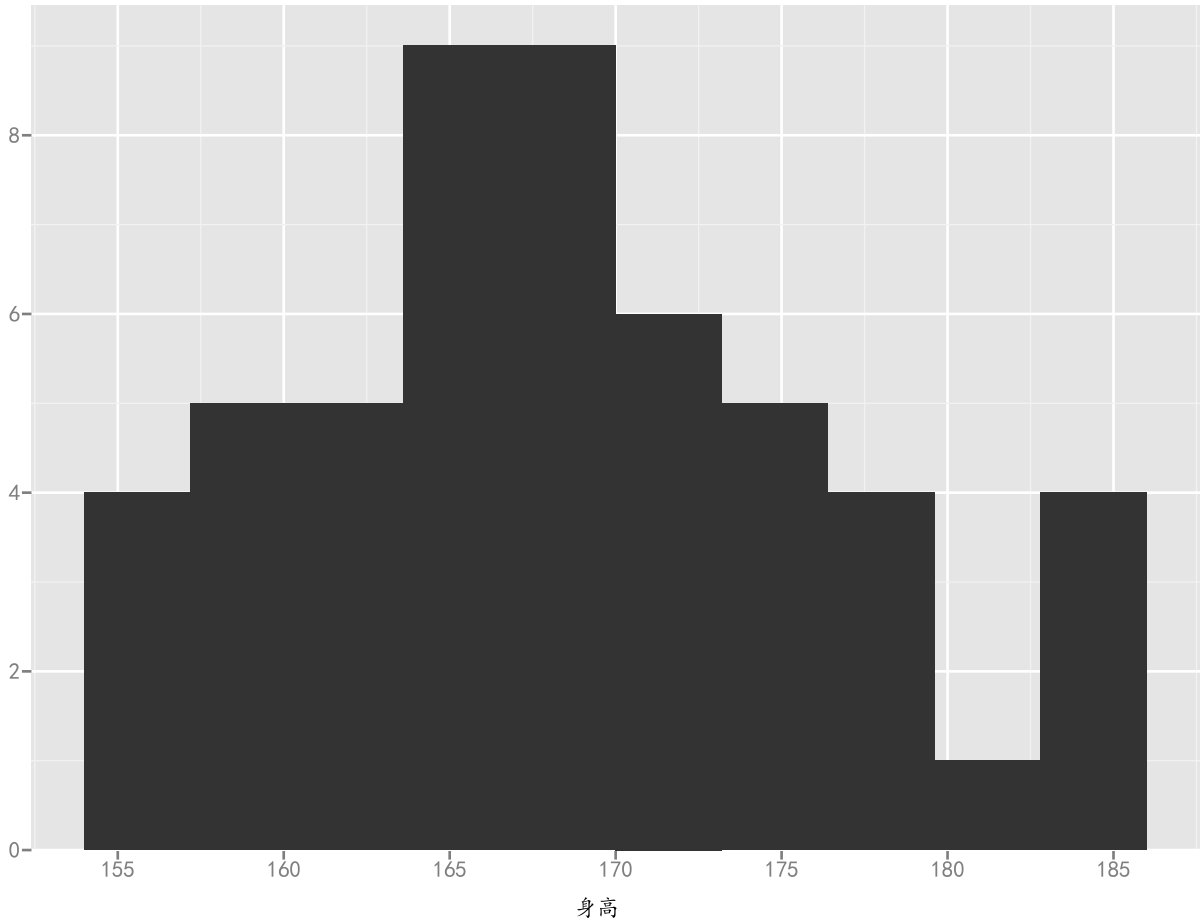
```
from ggplot import *
import matplotlib.pyplot as plt                    #基本绘图包

plt.rcParams['font.sans-serif']=['KaiTi'];    #SimHei黑体

plt.rcParams['axes.unicode_minus']=False;     #正常显示图中负号
```

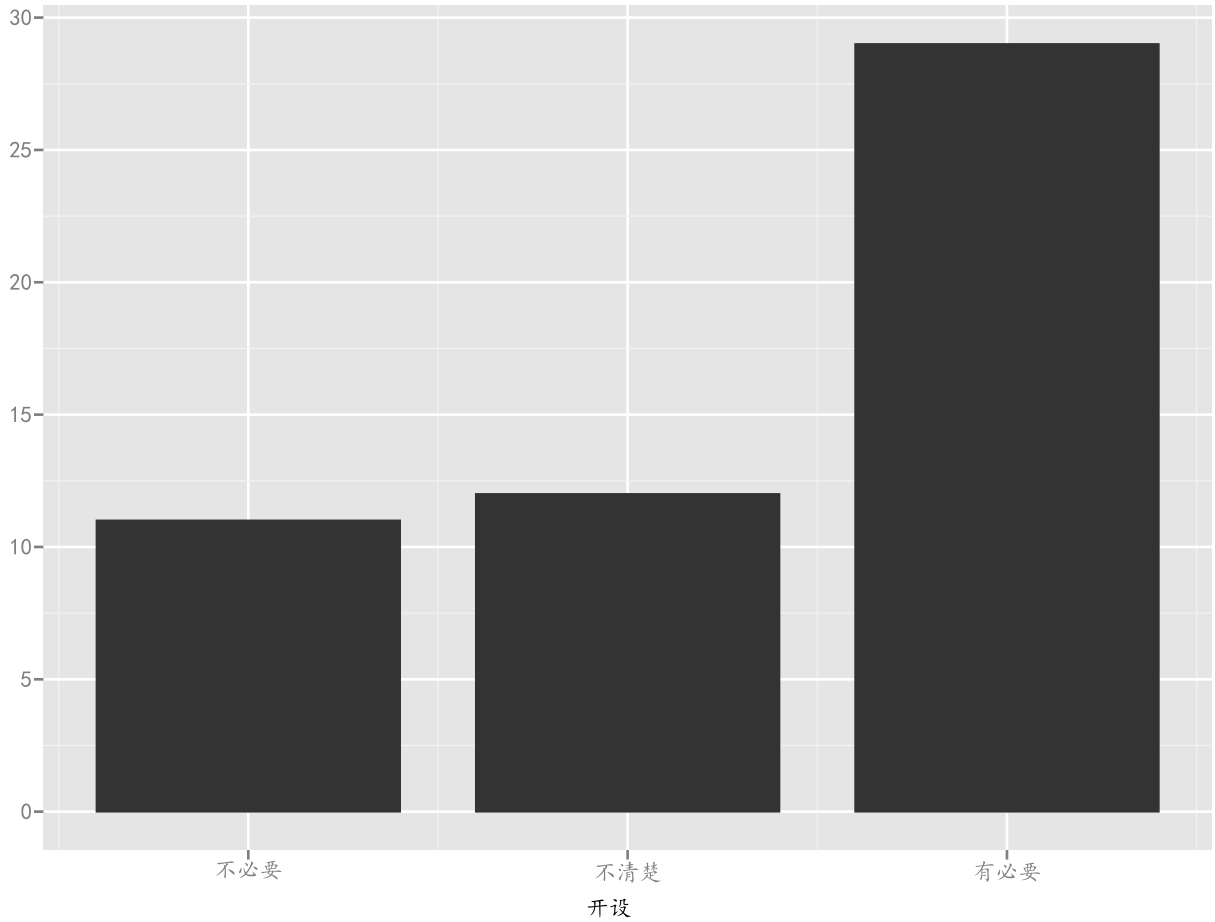
直方图

```
1 qplot('身高',data=BSdata, geom='histogram')
```



条形图

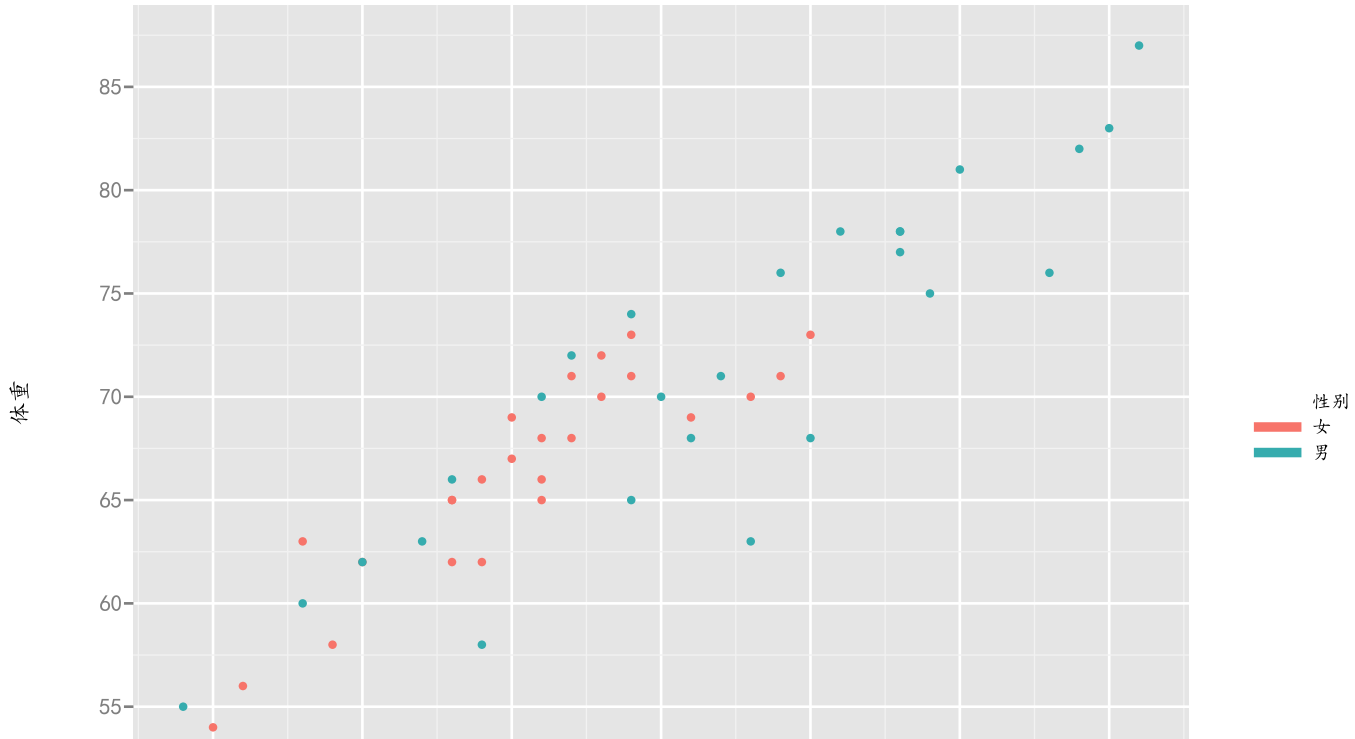
```
1 qplot('开设',data=BSdata, geom='bar')
```



散点图

1. 散点图用来描述两个定量数据变量间的关系，对于多元数据，通常可以用散点颜色和大小来反映不同的属性，
2. 下面对“身高”与“体重”变量进行绘图，其中 color 参数指定不同的性别所显示点的颜色。

```
1 qplot('身高','体重',data=BSdata,color='性别')
```



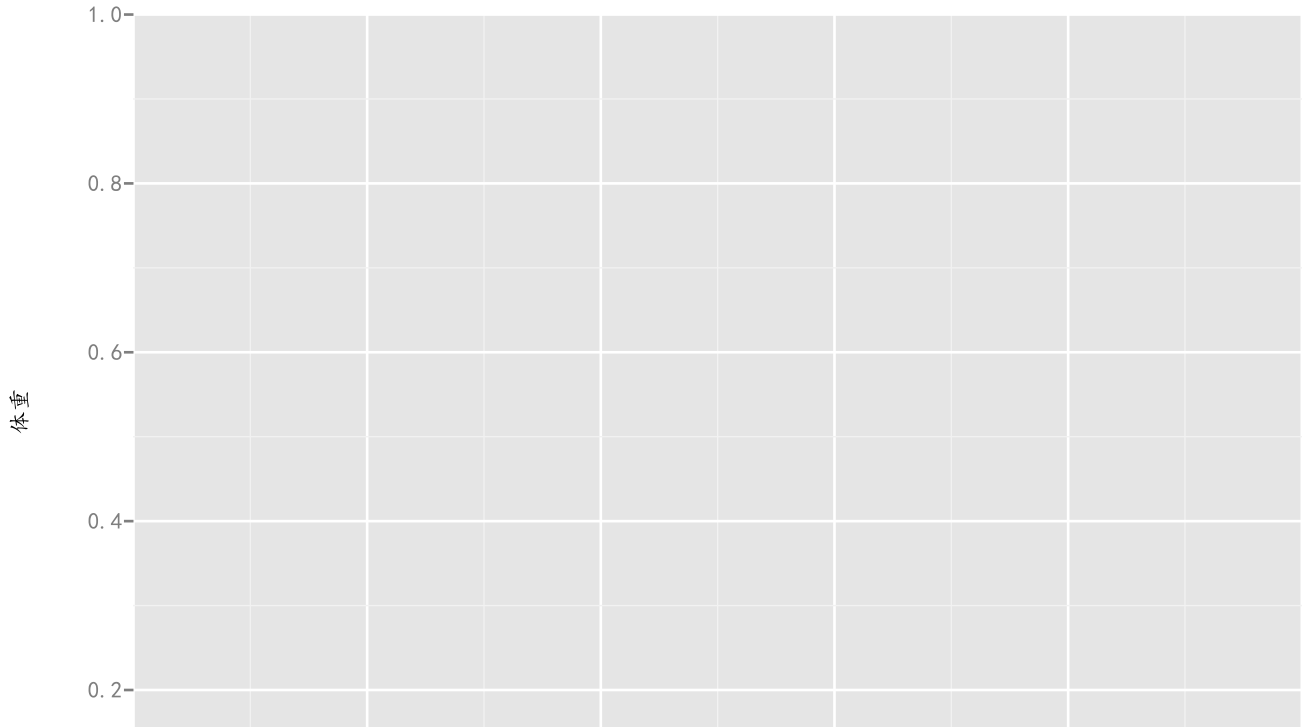
- qplot 快速制图
- ggplot 基本绘图

基本定义

1. 图层（layer）

如果你用过 Photoshop，那么对于图层一定不会陌生。一个图层好比一张玻璃纸，包含各种图形元素，可以分别建立图层，然后叠放在一起，组合成图形的最终效果。图层允许用户一步步地构建图形，方便单独对图层进行修改、增加统计量，甚至改动数据。

```
1 GP=ggplot(aes(x='身高',y='体重'),data=BSdata);GP #绘制直角坐标系
```



基本定义

2. 标度（scale）

标度是一种函数，它控制了数学空间到图形元素空间的映射。一组连续数据可以映射到 X 轴坐标，也可以映射到一组连续的渐变色彩。一组分类数据可以映射成不同的形状，也可以映射成不同的大小。

3. 坐标系统（coordinate）

坐标系统控制了图形的坐标轴并影响所有图形元素，最常用的是直角坐标轴，坐标轴可以进行变换以满足不同的需要，如对数坐标。其他可选的还有极坐标轴。

4. 位面（facet）

很多时候需要将数据按某种方法分组，分别进行绘图，位面就是控制分组绘图的方法和排列形式。

图层概念

1. 下面首先用一个例子展示 ggplot 的绘图功能。
2. 首先加载 ggplot，然后用 ggplot 定义第一层（即数据来源）。
3. 其中 aes 函数的参数非常关键，它将“身高”映射到 x 轴，将“体重”映射到 y 轴，然后使用 + 号添加两个新的图层，第二层加上了散点。

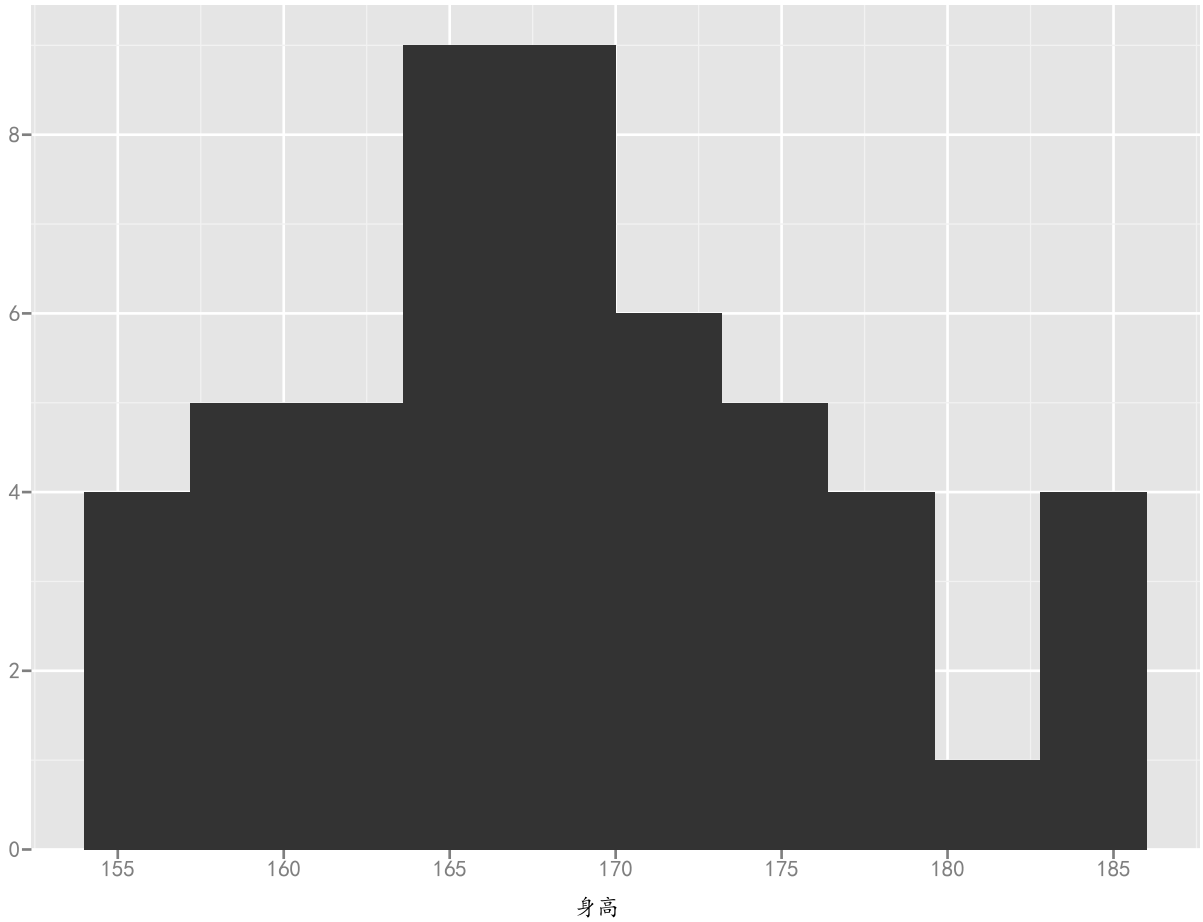
图层的优点

1. 用户可在更抽象层面上控制图形，使创造性绘图更容易；采用图层的设计方式，有利于结构化思维；图形美观，同时避免烦琐细节。
2. 每个点都有自己图像上的属性，比如 x 坐标，y 坐标，点的大小、颜色和形状，这些都叫做 aesthetics，即图像上可观测到的属性，通过 aes 函数来赋值，如果不赋值，则采用 python 的内置默认参数。
3. ggplot 先做 mapping，设定画图对象的 x 坐标和 y 坐标，以及点的颜色、形状，其描述对象的方式都是数据类型（通过 aes 函数来设定参数），然后再做沙岭，把映射的数据转化为图形语言，如转化为像素大小。
4. geom 确定图像的“type”，即几何特征，用点来描述图像，还是用柱或条形。
5. 关于变量问题，ggplot 函数中赋予的值是全局性质的，如果不希望全局生效，则放到后面 + 对应的图层中。

常见图形

1 直方图

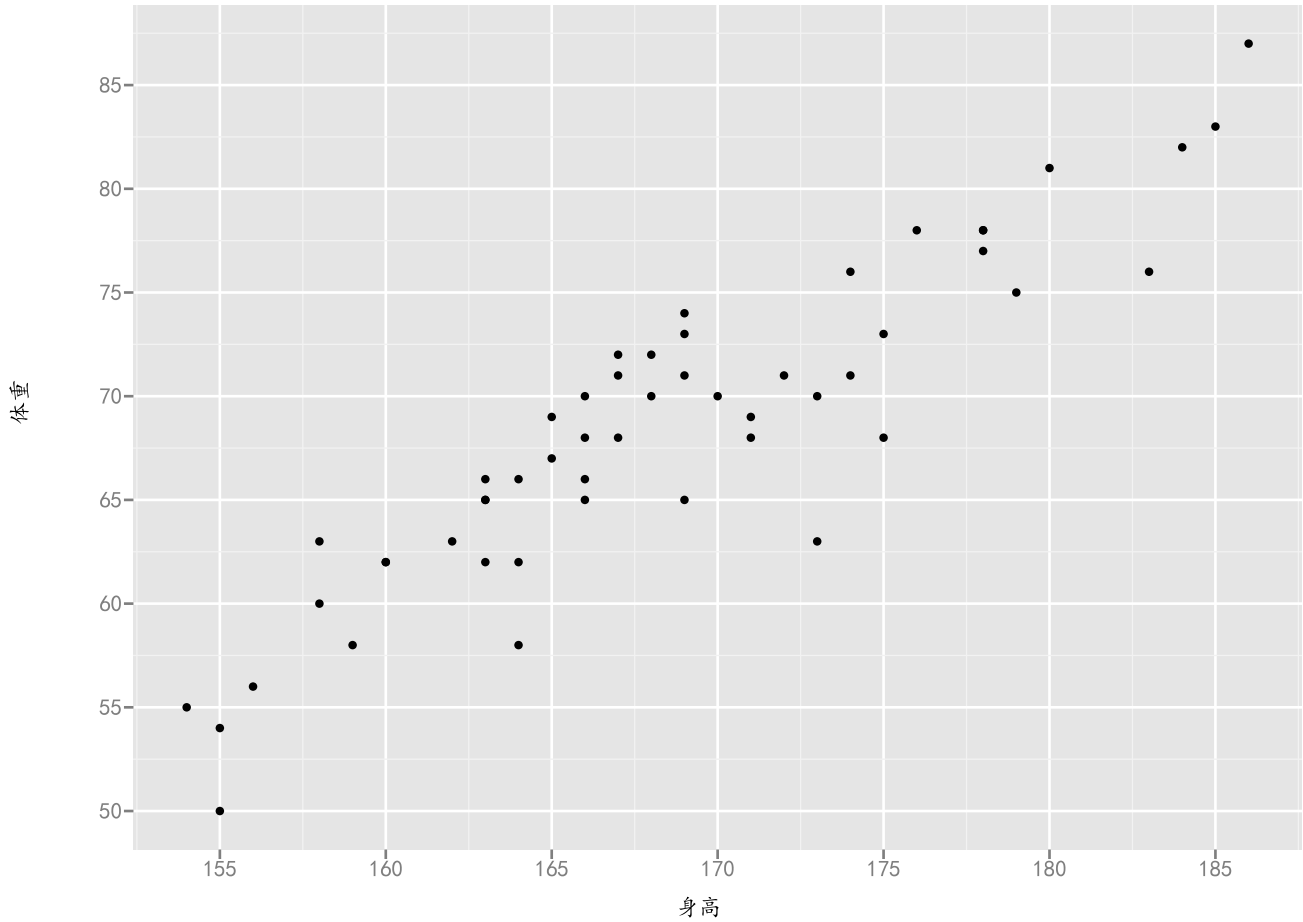
```
1 ggplot(BSdata,aes(x='身高'))+ geom_histogram()
```



常见图形

2 散点图

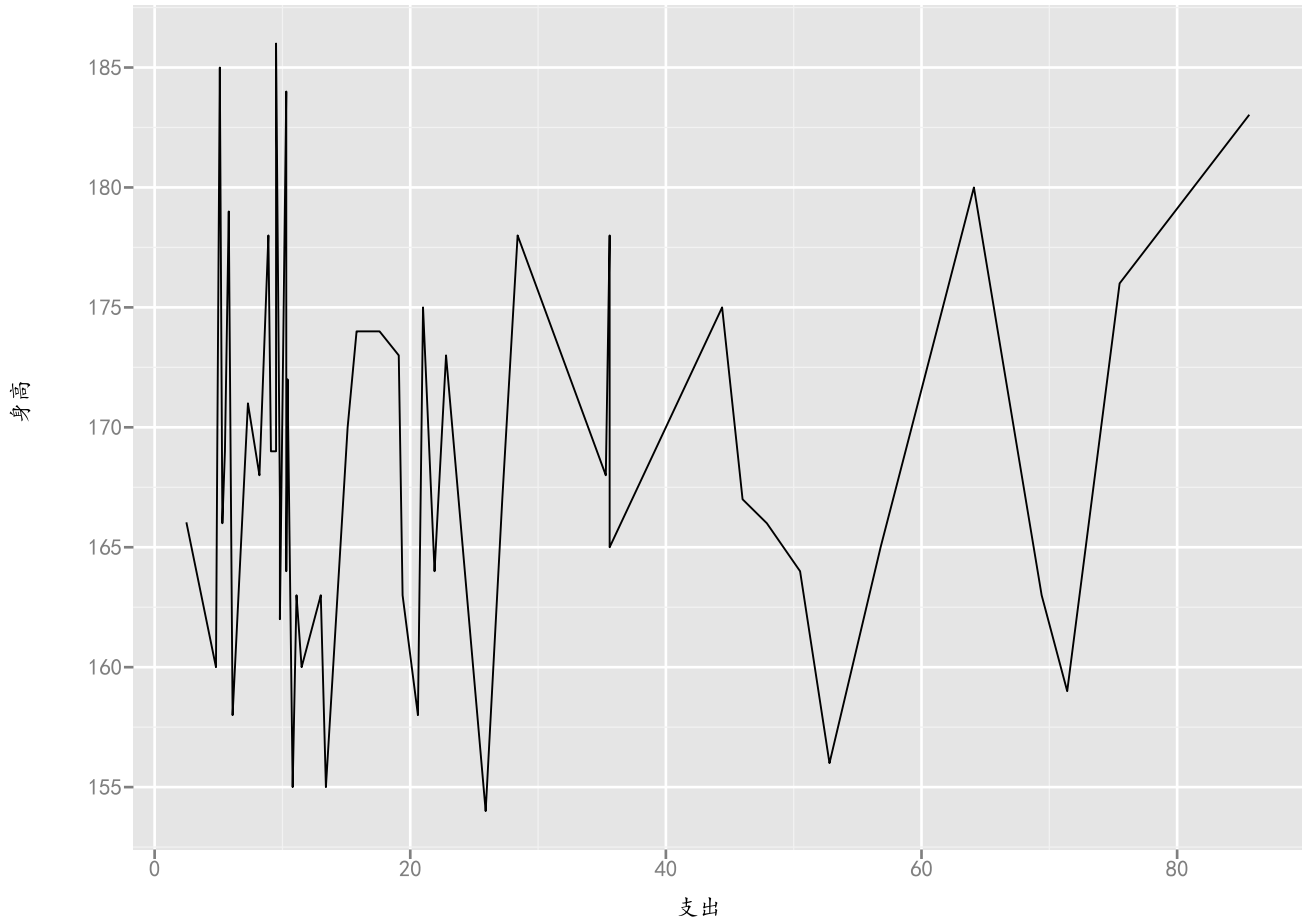
```
1 ggplot(BSdata,aes(x='身高',y='体重')) + geom_point()
```



常见图形

3 线图

```
1 ggplot(BSdata,aes(x='支出'))+geom_line(aes(y='身高'))
```

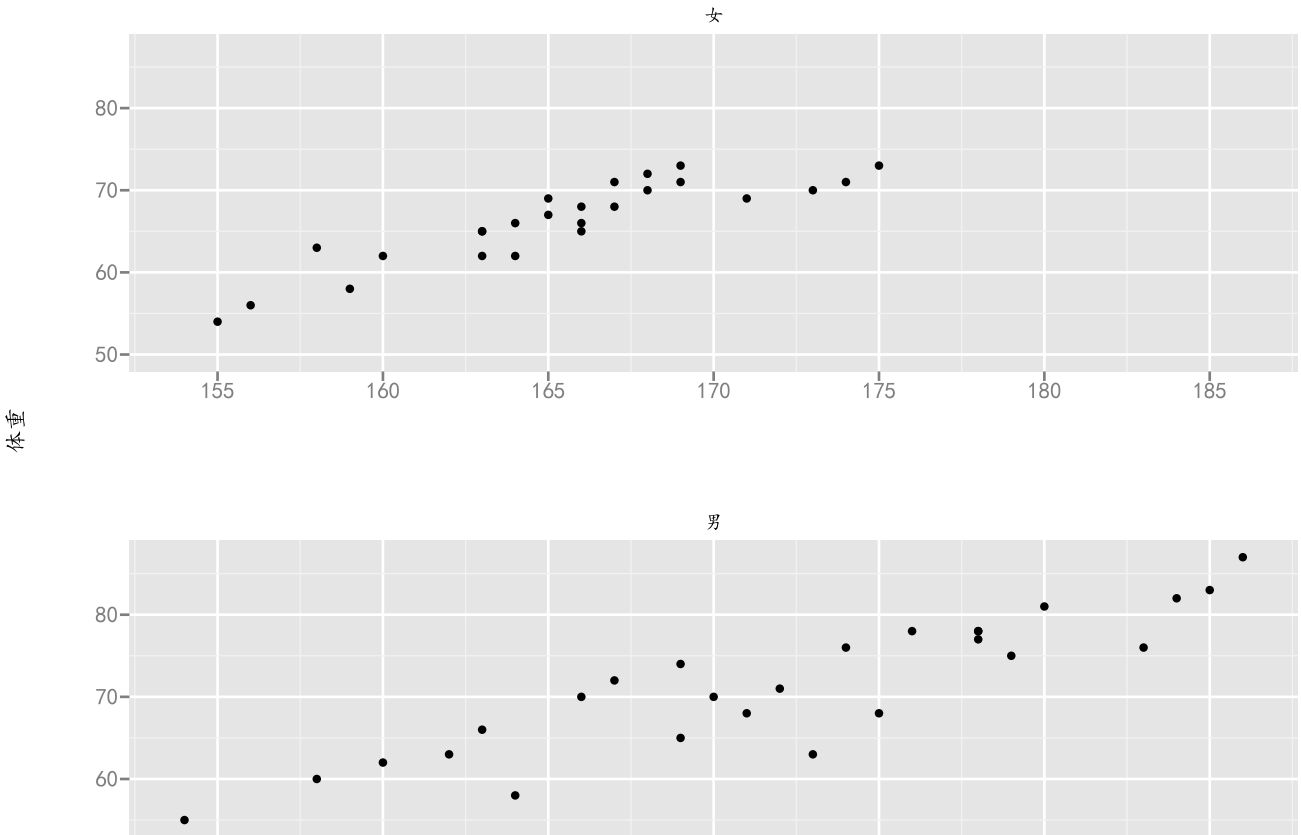


常见图形

4 分面图

使用 facet_wrap 参数可以按照类型绘制分面图。

```
1 ggplot(BSdata,aes(x='身高',y='体重')) + geom_point() + facet_wrap('性别
      ')
```



图形主题

1. ggplot 提供一些已经写好的主题，比如，theme_grey() 为默认主题
2. theme_bw() 为白色背景的主题，
3. theme_classic() 主题，与 python 的基础画图函数类似。

```
1 ggplot(BSdata,aes(x='身高',y='体重',color='性别'))+geom_point()+theme_
      bw()
```