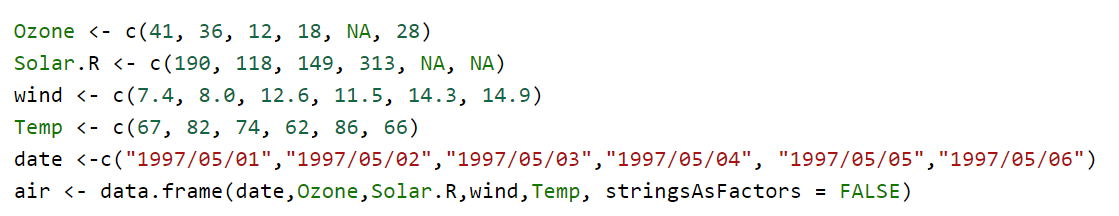
实验二 R语言与描述性统计

【实验目的】

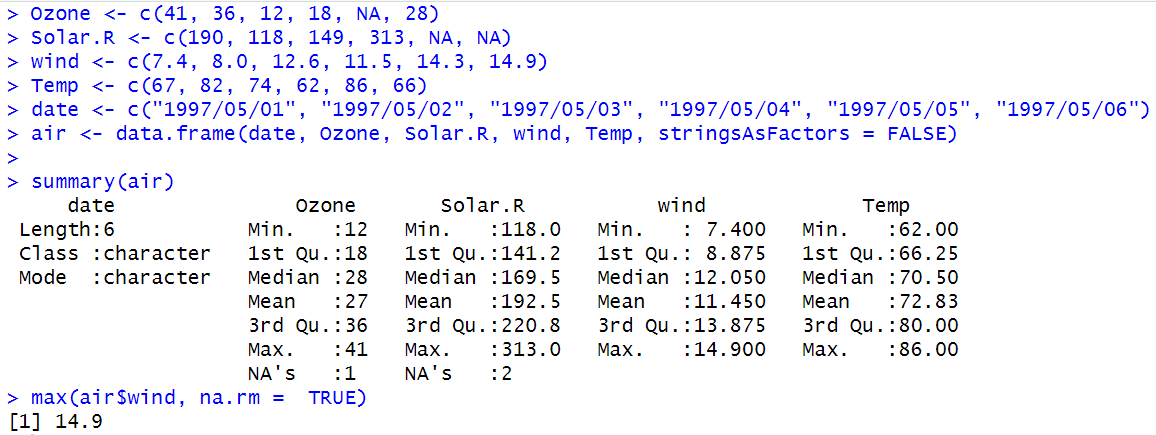
1. 掌握数据预处理的相关操作，包括查看数据集的缺失值，对缺失值进行处理，对异常变量重新赋值。
2. 掌握使用可视化图表方法进行探索性统计分析。
3. 掌握常用的概率分布，并学会用R语言编程求解概率分布问题。
4. 掌握常用的抽样方法，并学会用R语言编程实现抽样。

【实验内容与实现】

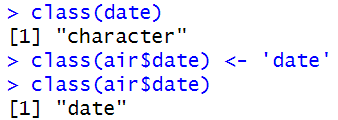
1. 根据下述代码，生成数据，并按要求做统计分析。



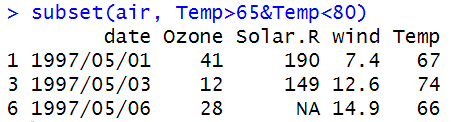
1. 使用summary()函数查看哪一列含有缺失值的个数，并查看‘wind’列的最值。



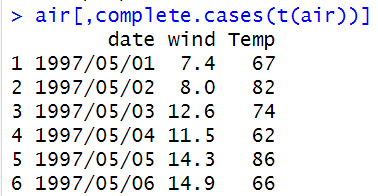
1. 使用class()函数查看date的类型，并将其转化为date型，格式为年/月/日。



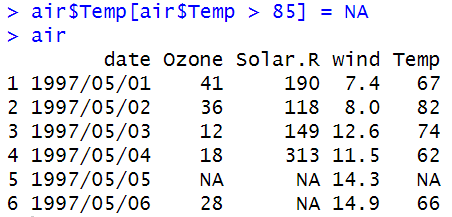
1. 使用subset()根据‘Temp’列获取air中‘Temp’大于65小于80的数据。



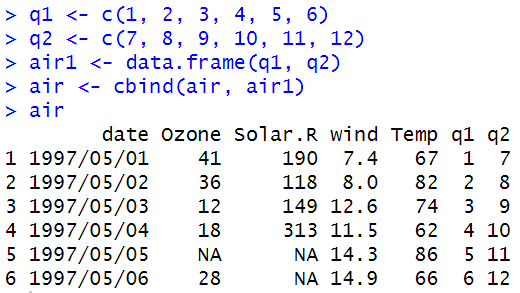
1. 去掉air中含有缺失值的列。



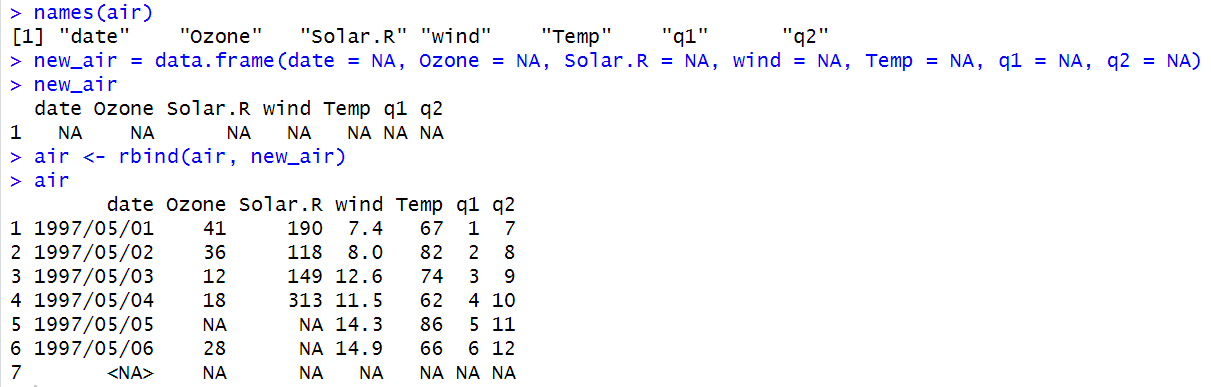
1. 找出 ‘Temp’ 中大于 85 的值将其设置为缺失值。



1. 创建一个数据框 air1 包含两个列 q1，q2。用 cbind 函数添加到 air 中。



1. 使用 names（） 函数查看 air 的列名，再自己创建一个新的数据框，新数据框的列名要与air的列名相同。使用 rbind（） 函数添加到 air 中。



1. 科学家正在开展一项模拟车窗着色对侧窗视觉影响的实验，测得的数据在data.txt，数据中含有以下几列：

case：观测序号；

id：研究车辆的识别号（1-26）；

age：受测者年龄；

sex：表示受测者性别的因子变量，具有f（女性）和m（男性）两个水平；

tint：表示着色程度的有序因子变量，no<lo<hi；

target：表示对比度的因子变量，水平为：locon（低对比度），hicon（高对比度）；

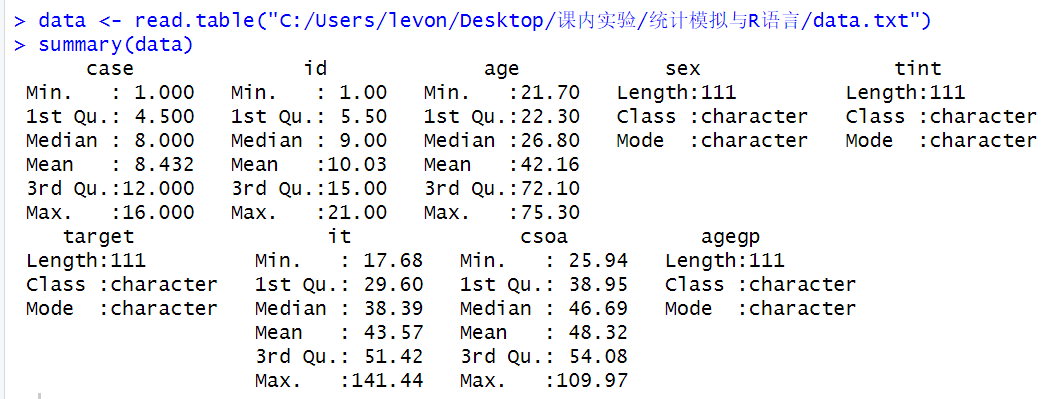
it：检查时间，执行简单识别任务所需的时间（单位为毫秒）；

csoa：临界刺激开始异步，识别字母数字目标的时间（单位为毫秒）；

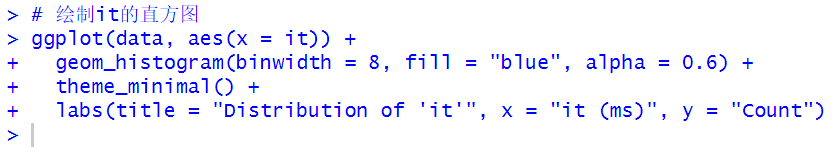
agegp：年龄阶段因子变量，水平为younger（21-27岁）、older（70-78岁）；

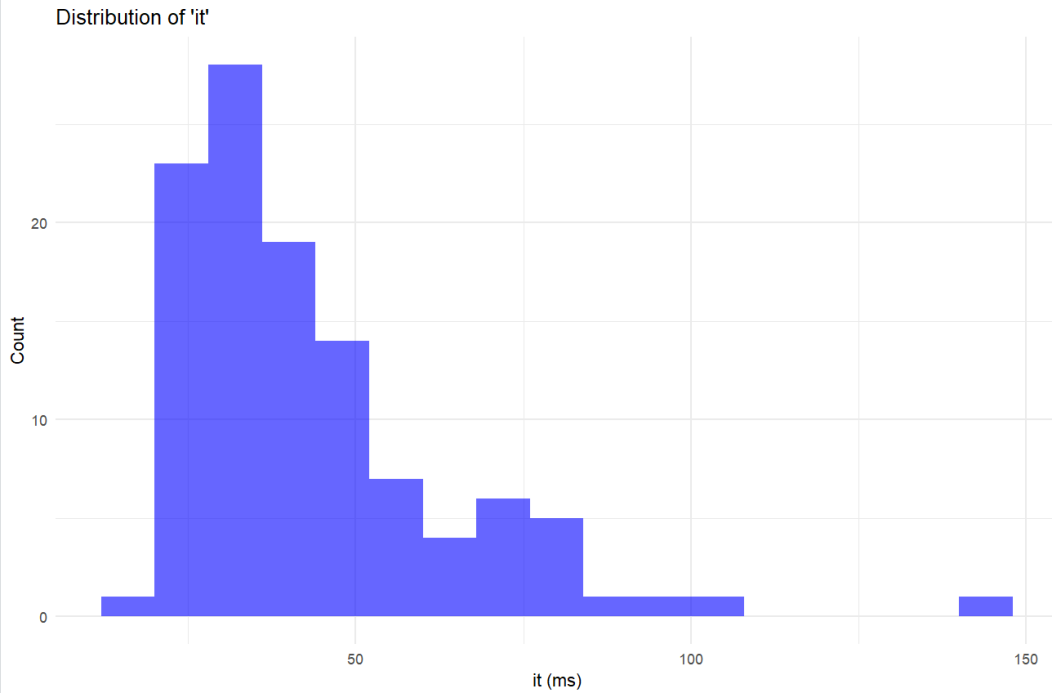
完成下述任务：

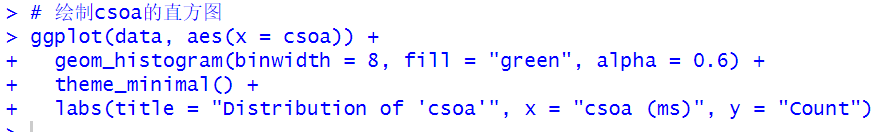
1. 读取数据，并逐列对数据进行描述性统计分析；

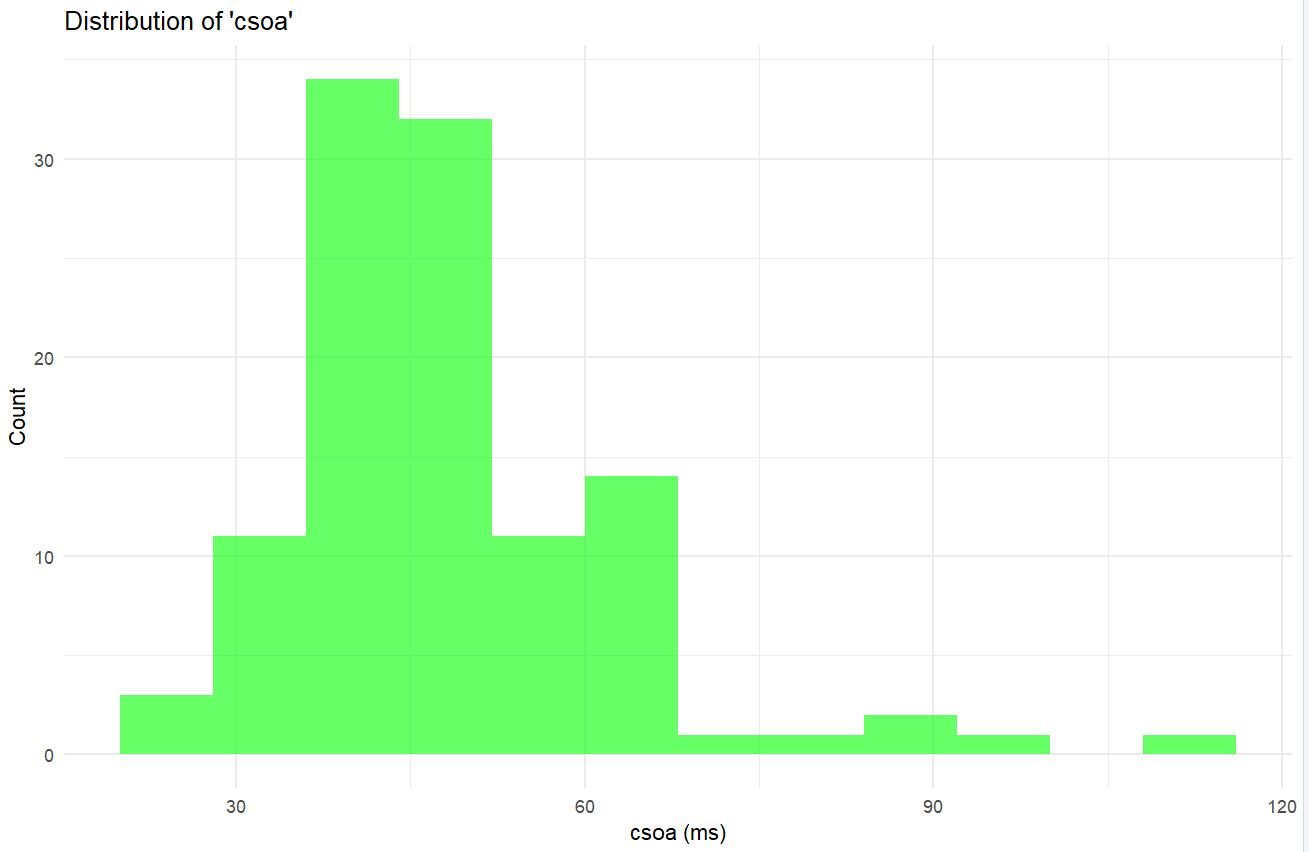


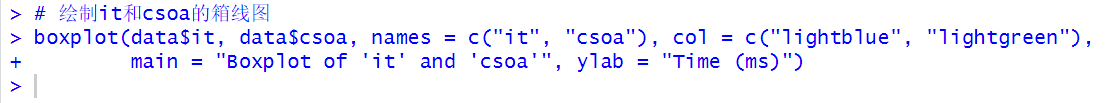
1. 分别探索it和csoa的数据分布情况；

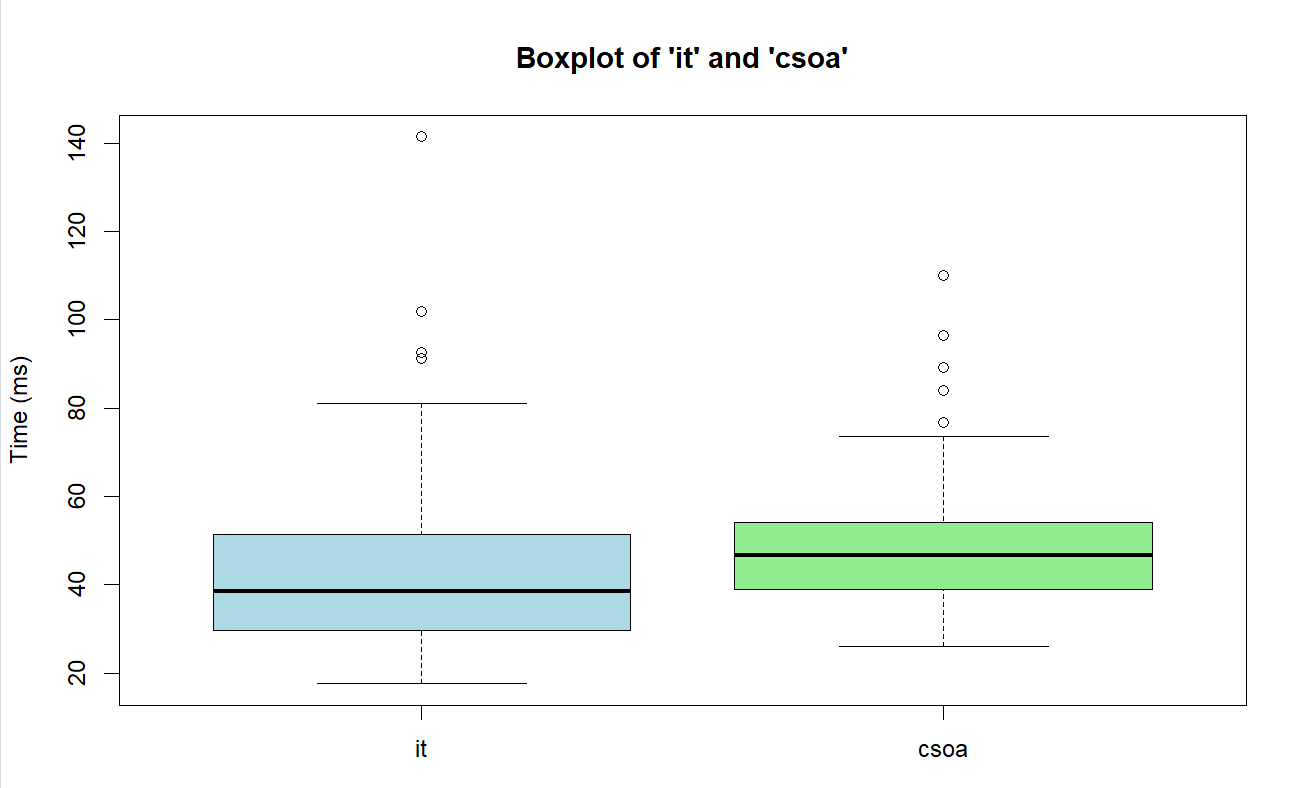


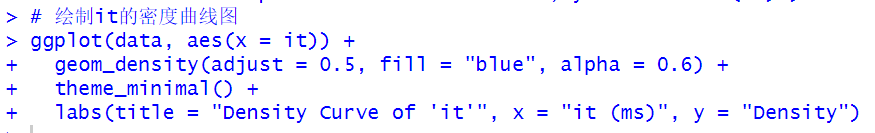


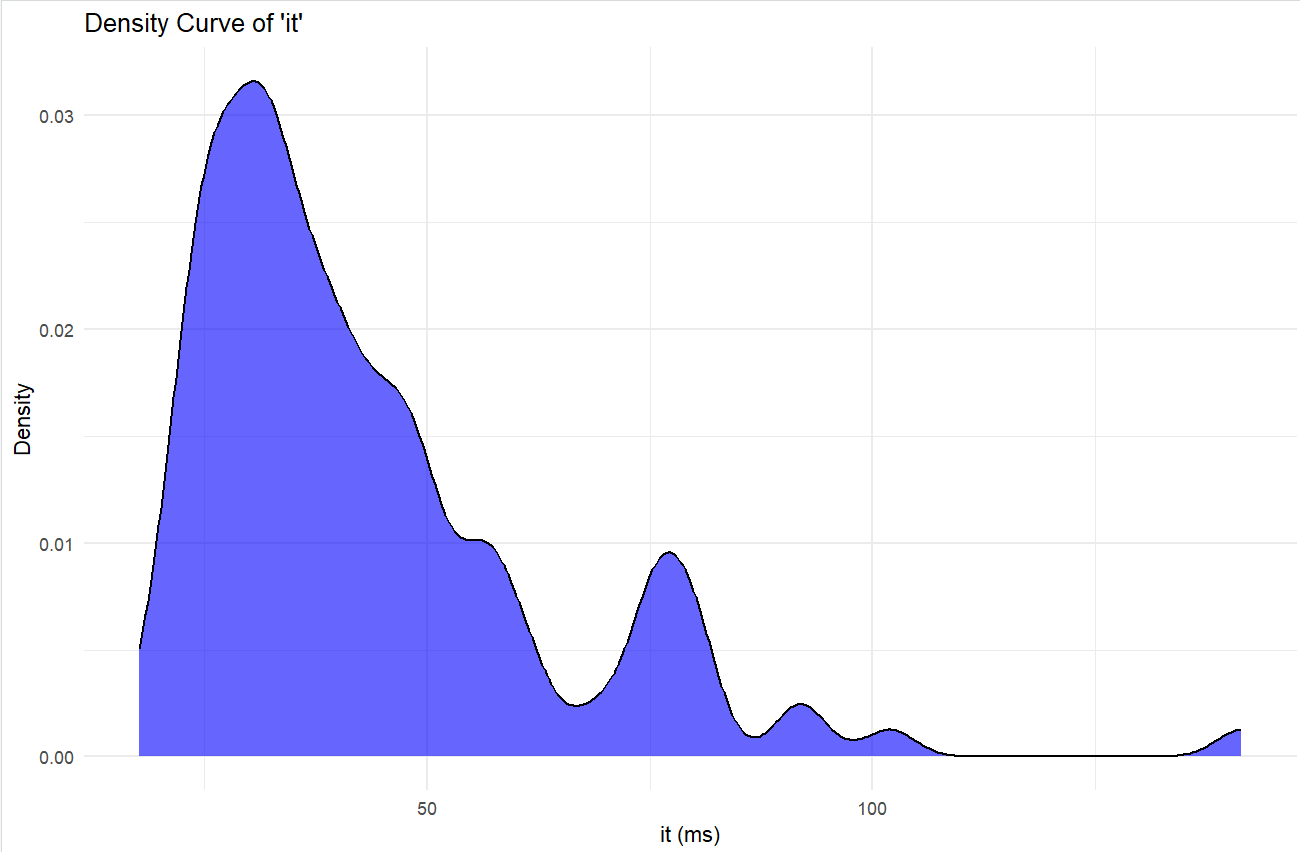


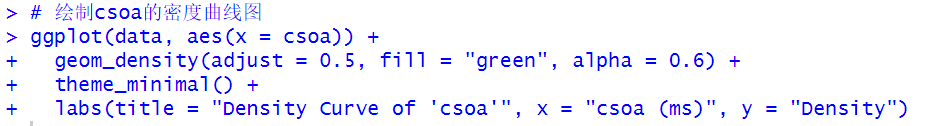


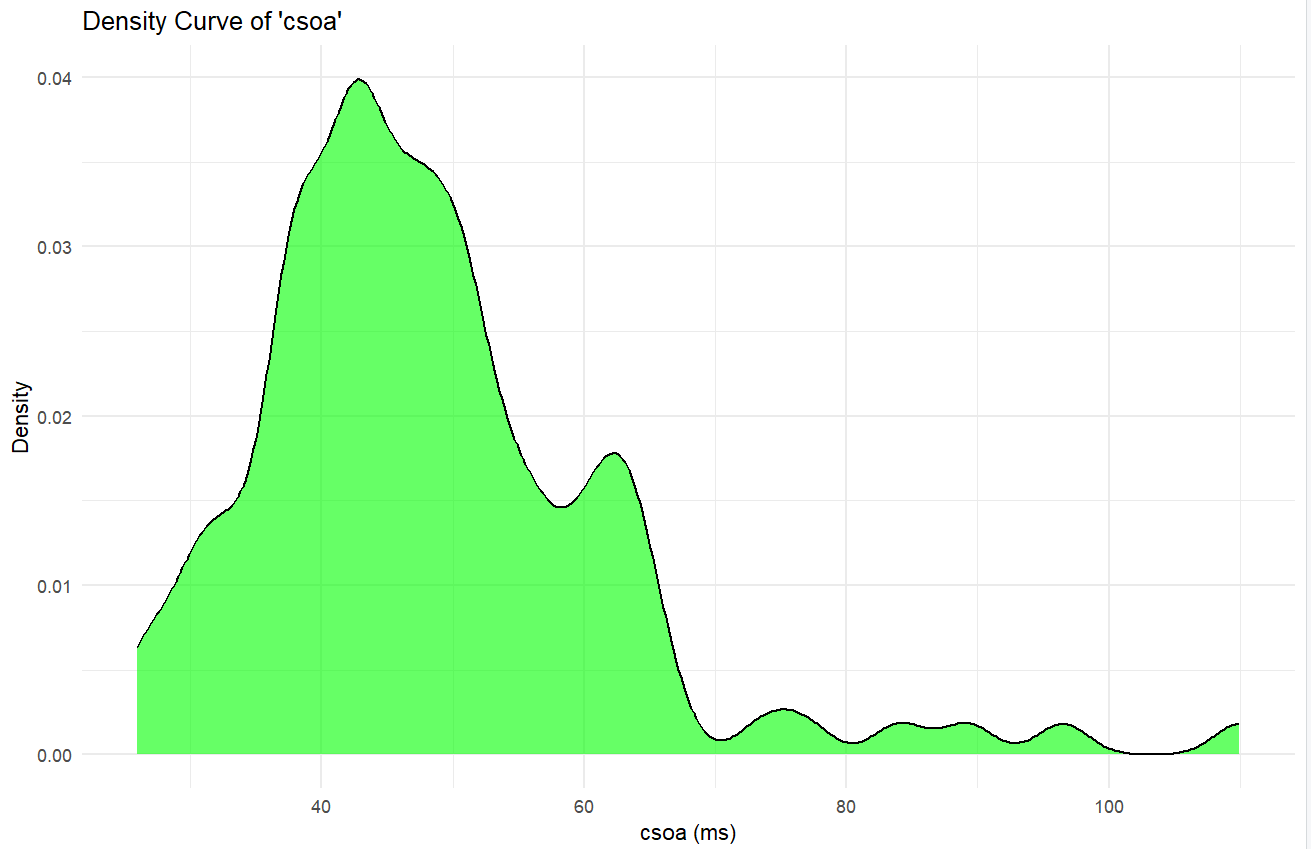


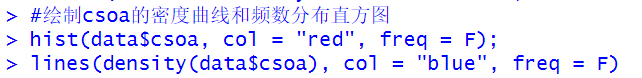


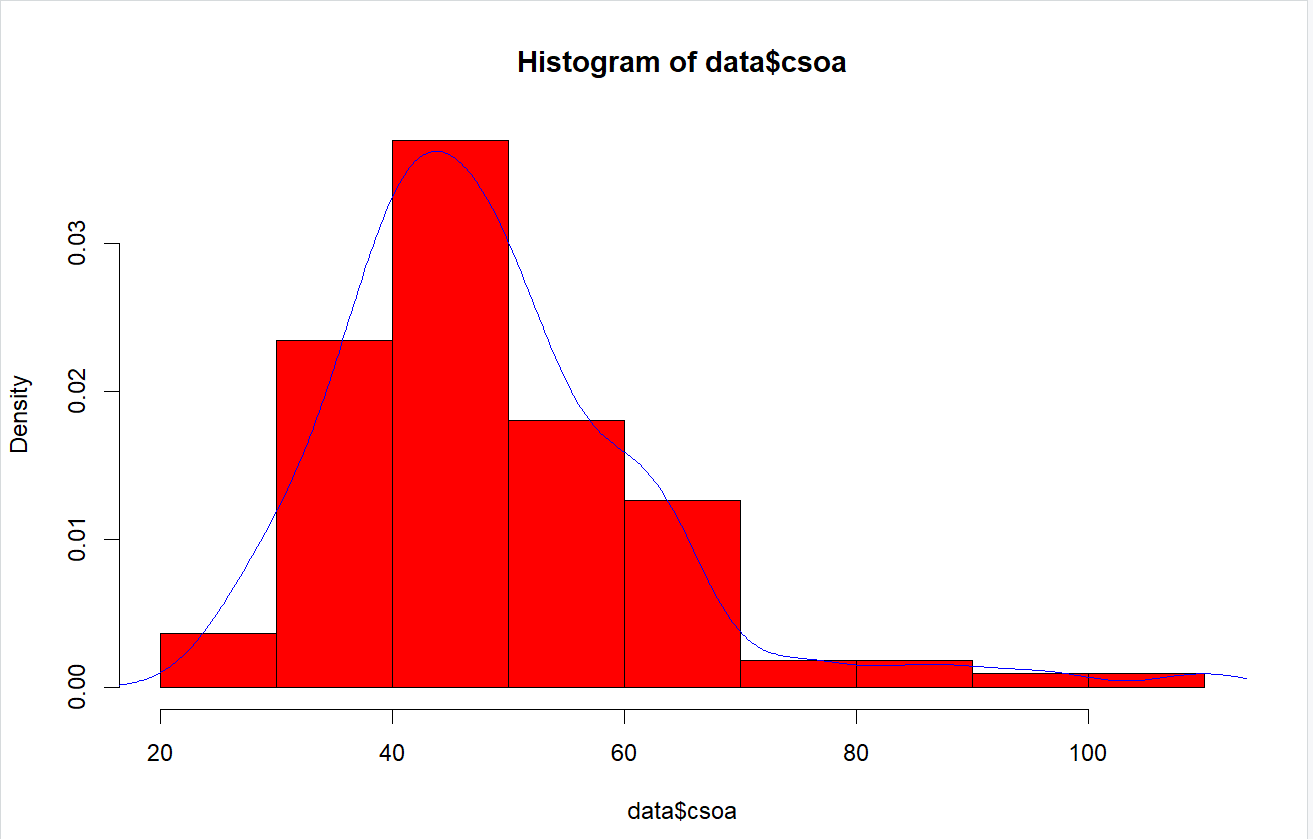




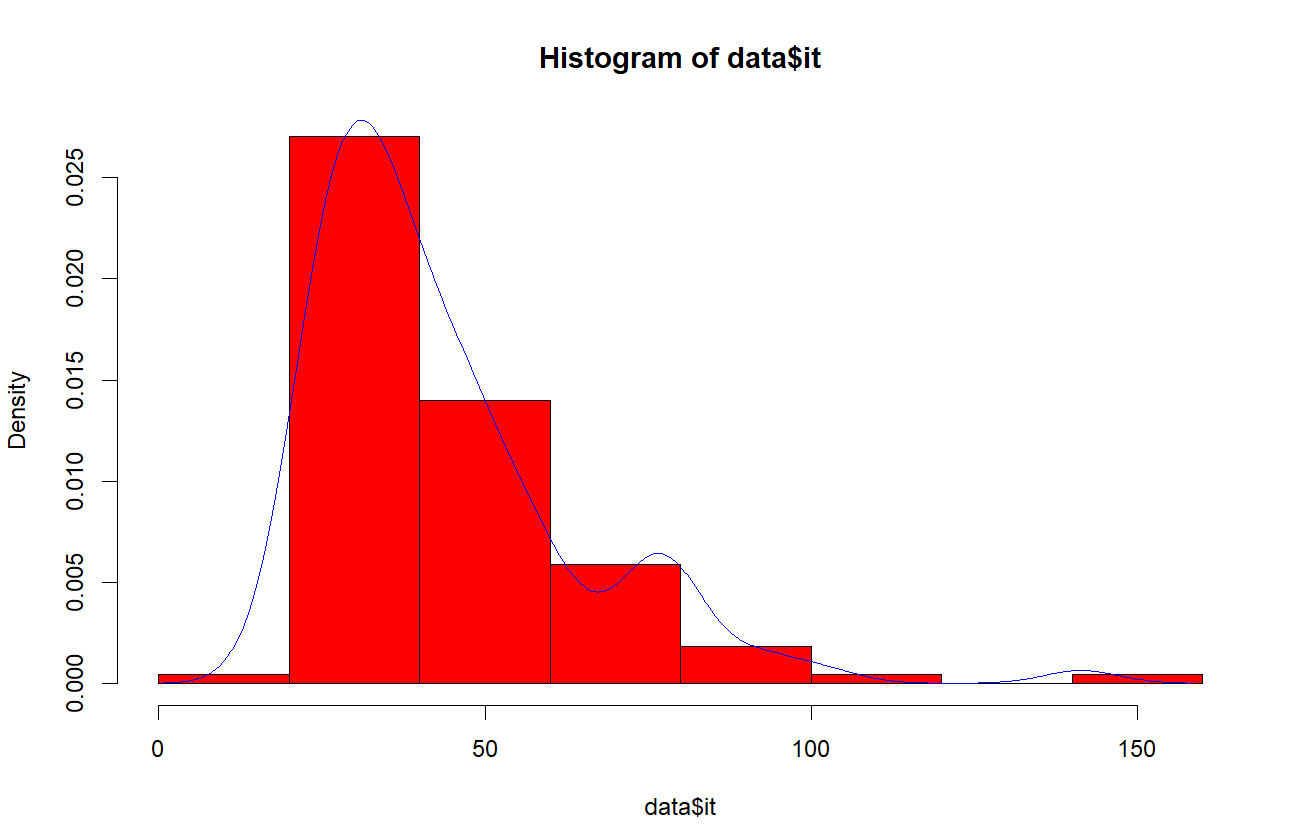




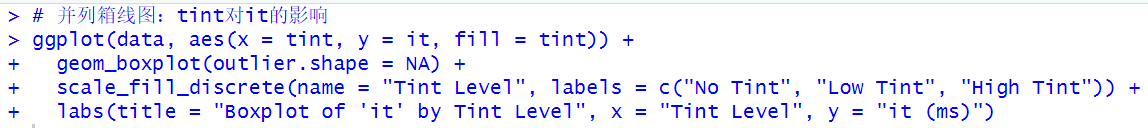


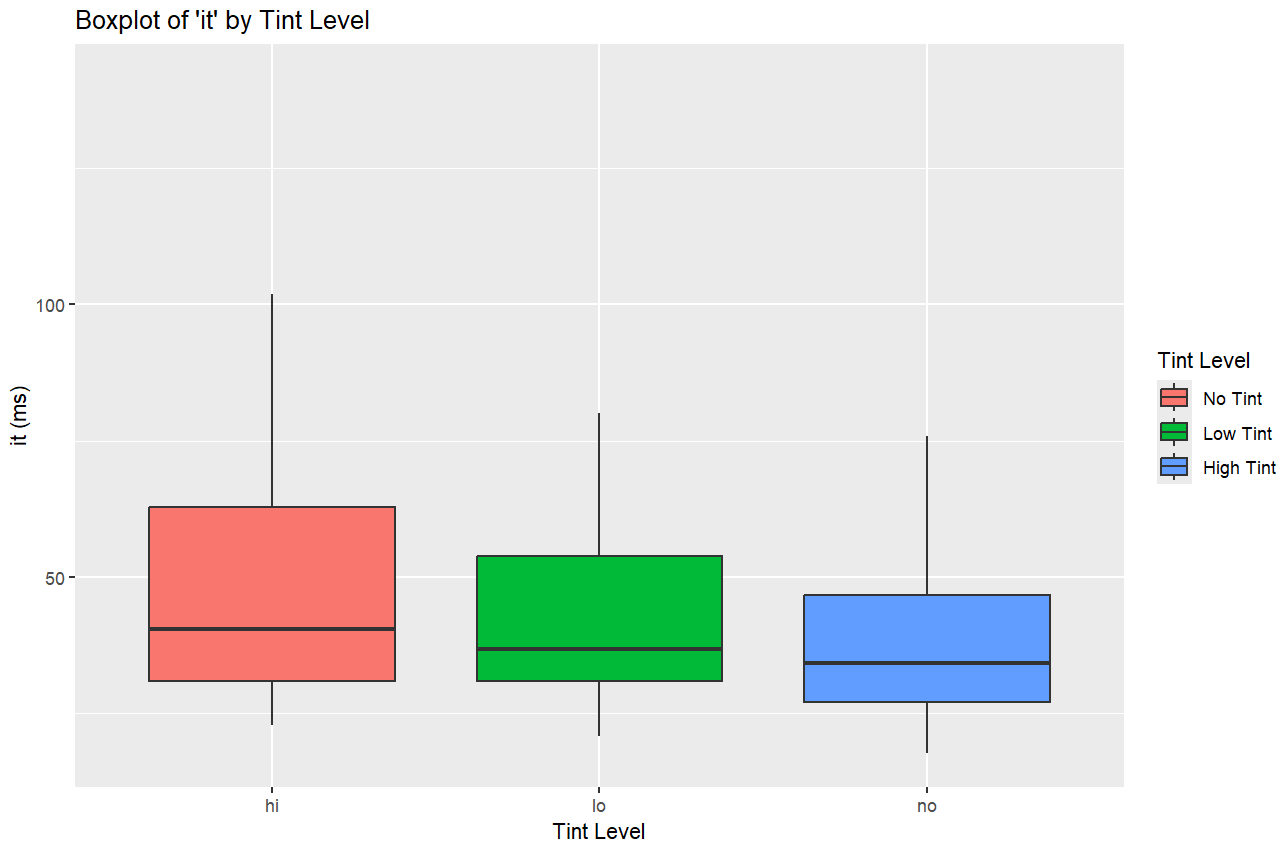


IMG_256



1. 利用图表探索变量tint对it和csoa的影响，并给出相应的文字解析；





(a)图表概述

·图表类型：箱线图（Boxplot）

·变量：着色程度（tint）对检查时间（it）的影响

·箱线图显示了三个不同着色程度（无着色、低着色、高着色）下，执行简单识别任务所需的时间（单位为毫秒）的分布情况。

（b）箱线图解析

·无着色（No Tint）：

中位数（箱线图中的横线）显示在30毫秒左右，表明在无着色条件下，识别任务的平均完成时间大约为30毫秒。箱形图的“箱子”部分较宽，说明数据分布较为分散，但大部分数据集中在30毫秒左右。

·低着色（Low Tint）：

中位数略高于无着色组，大约在35毫秒左右，表明低着色条件下，识别任务的平均完成时间略长。箱子的形状与无着色组相似，但整体上移，显示数据分布略有变化，但仍然相对集中。

·高着色（High Tint）：

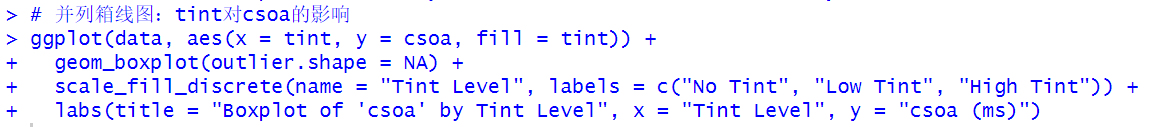
中位数明显高于前两组，大约在40毫秒左右，表明高着色条件下，识别任务的平均完成时间显著增加。箱子部分非常宽，表明数据分布更加分散，且存在极端值（由于outlier.shape = NA，这些极端值未在图中显示）。

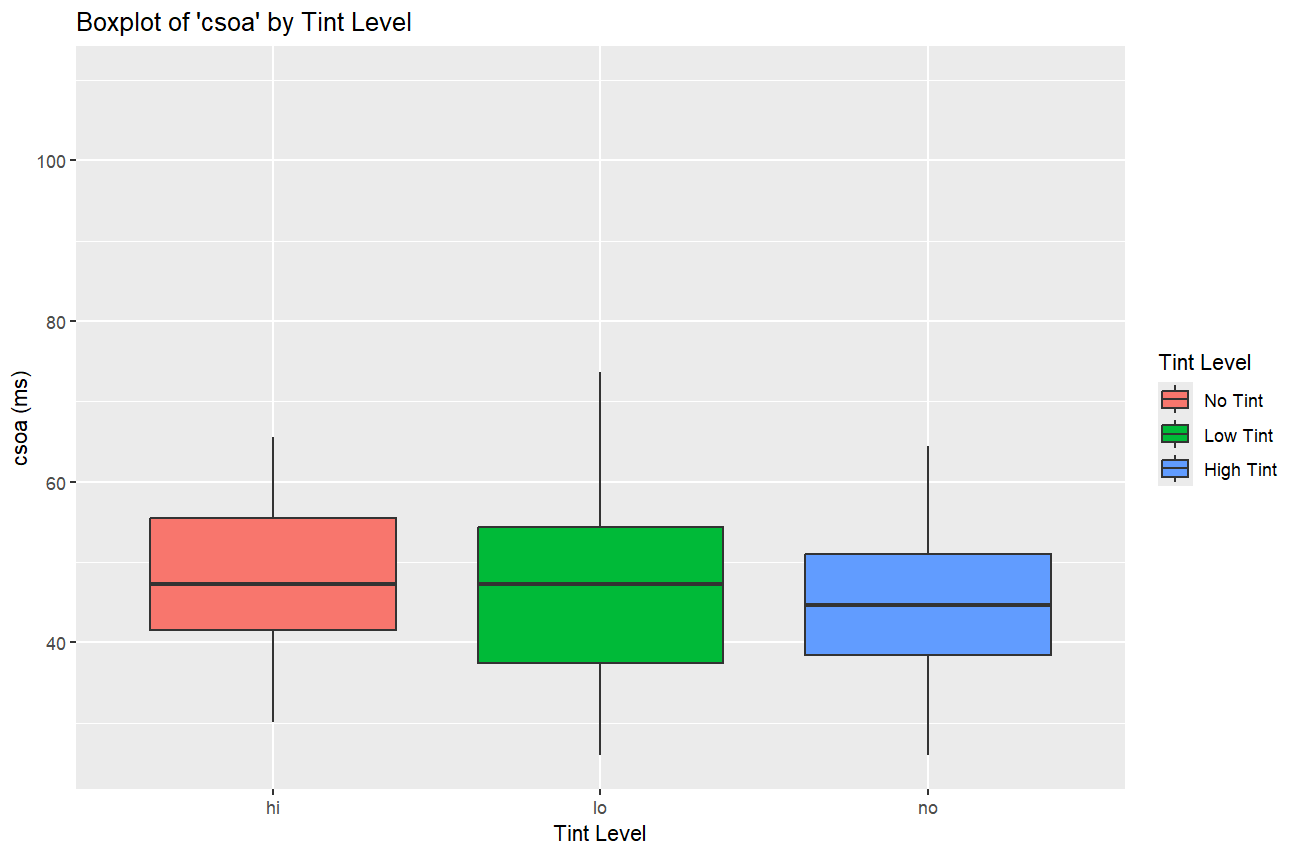
（c）结论

·着色程度对执行简单识别任务所需的时间有显著影响。

·随着着色程度的增加，识别任务的平均完成时间增加，表明着色可能降低了识别效率。

·高着色条件下的数据分布更加分散，可能意味着高着色对个体的视觉识别能力影响更大，导致完成时间的变异性增加。





1. 图表概述

·图表类型：箱线图（Boxplot）

·数据：展示了不同着色程度（tint）对临界刺激开始异步（csoa，单位毫秒）的影响。

·箱线图特征

无着色（No Tint）：箱线图显示无着色组的csoa值中位数在45毫秒左右。

低着色（Low Tint）：低着色组的csoa值中位数在48毫秒左右，比无着色组的中位数略高。

高着色（High Tint）：高着色组的csoa值中位数在48毫秒左右，与低着色组的中位数基本一致。

·结论

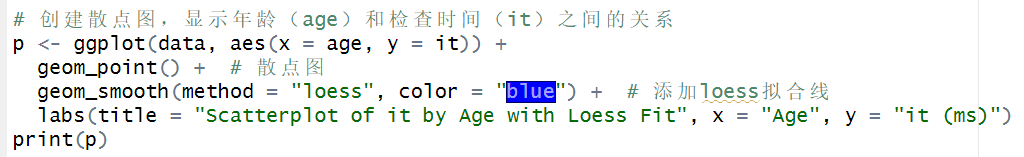
从箱线图可以看出，随着着色程度的增加，csoa值的中位数呈现增加趋势，虽然高着色组与低着色组的中位数基本一致，但是高着色组的箱子宽度更小，这表明着色程度越高，识别字母数字目标所需的平均时间越长。

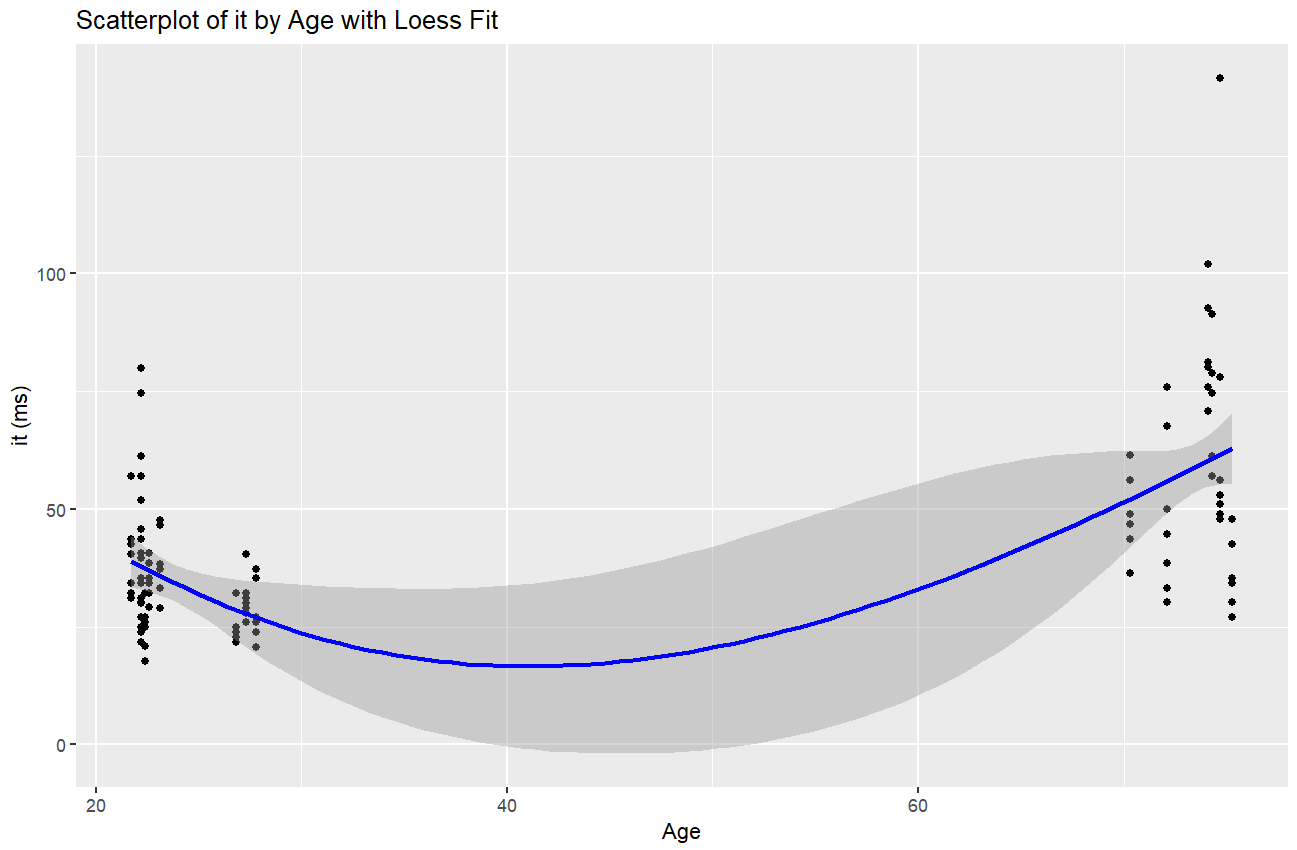
无着色组的csoa值分布较为集中，而到了低着色组的时候，数据的分布开始变得更加分散，这表明低着色条件下识别任务的变异性增加。

无着色和高着色组的箱子宽度（即IQR，四分位间距）相对较小，表明数据点比较集中，变异性较小。

从无着色到低着色再到高着色，箱子的宽度由窄变宽再变窄，这表明识别任务的变异性有一个先增大后减小的过程。

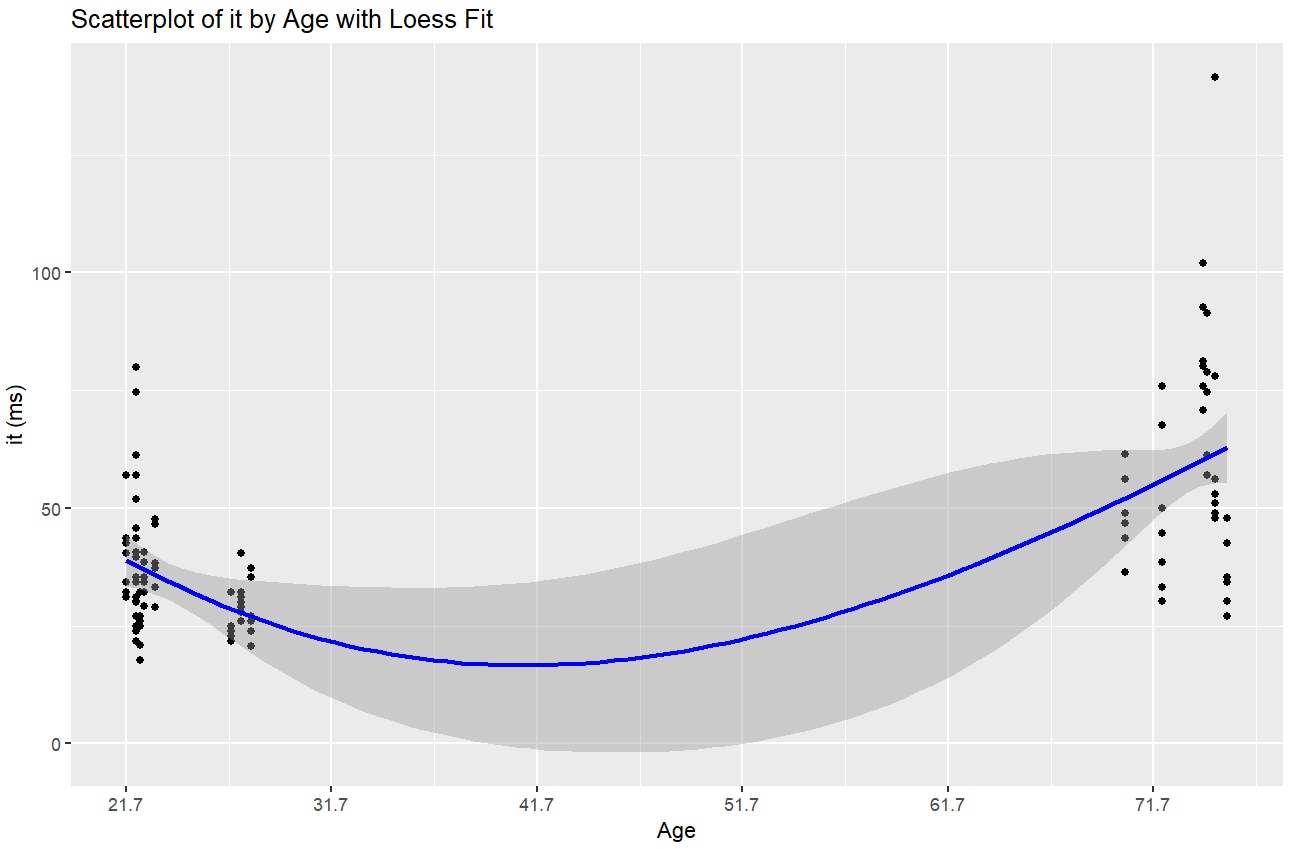
1. 作出age和it的散点图, 并进一步完成下面的操作:
2. 用loess()函数作出拟合线;



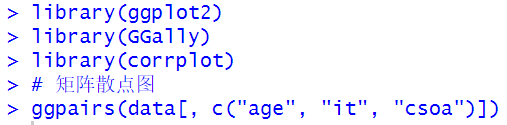


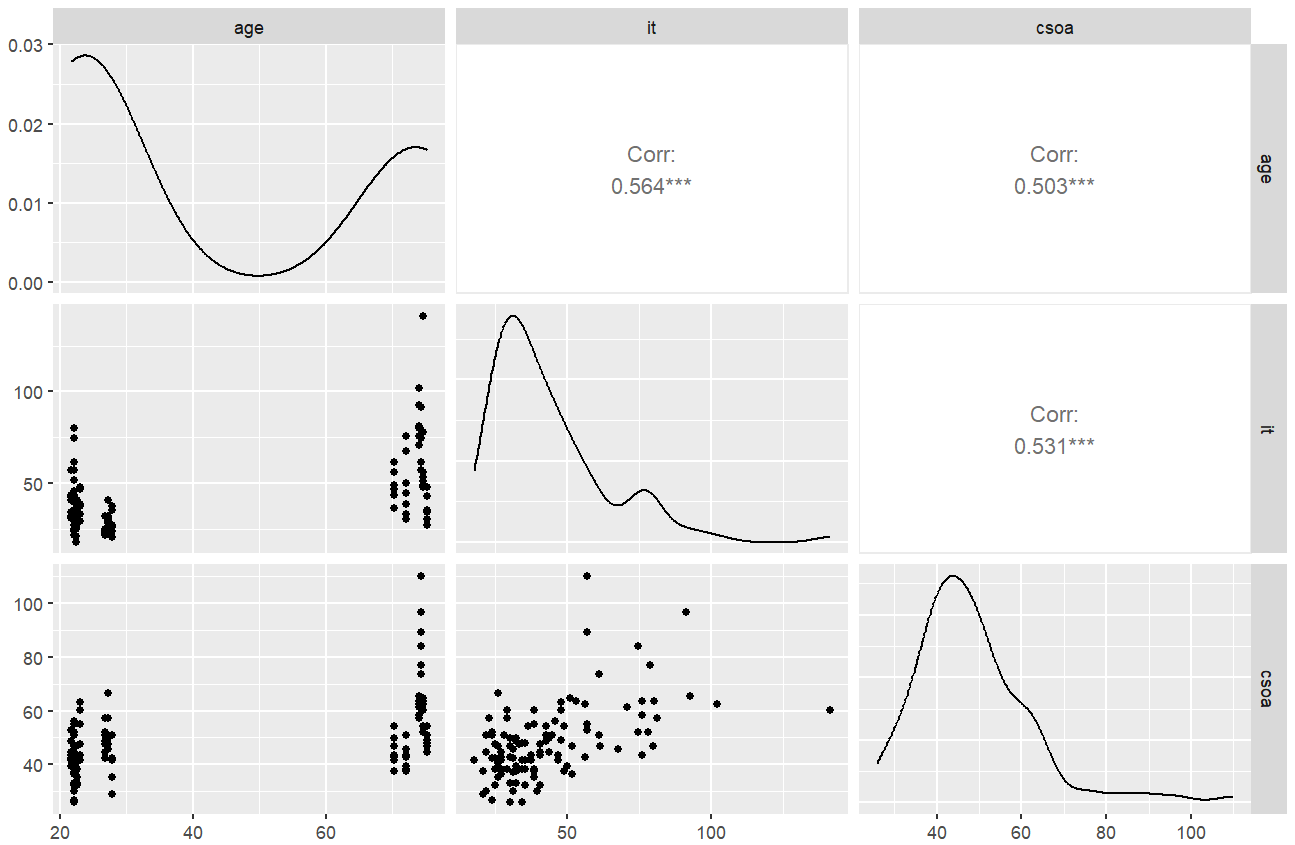
1. 对图的坐标刻度进行调整;

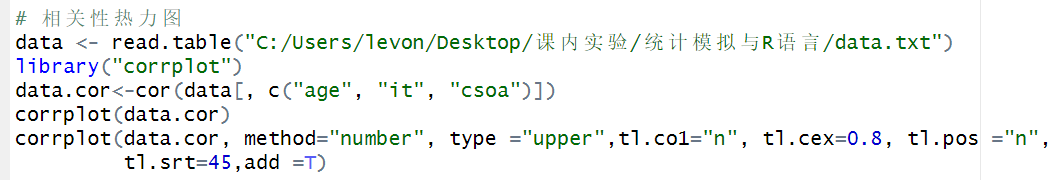
IMG_256

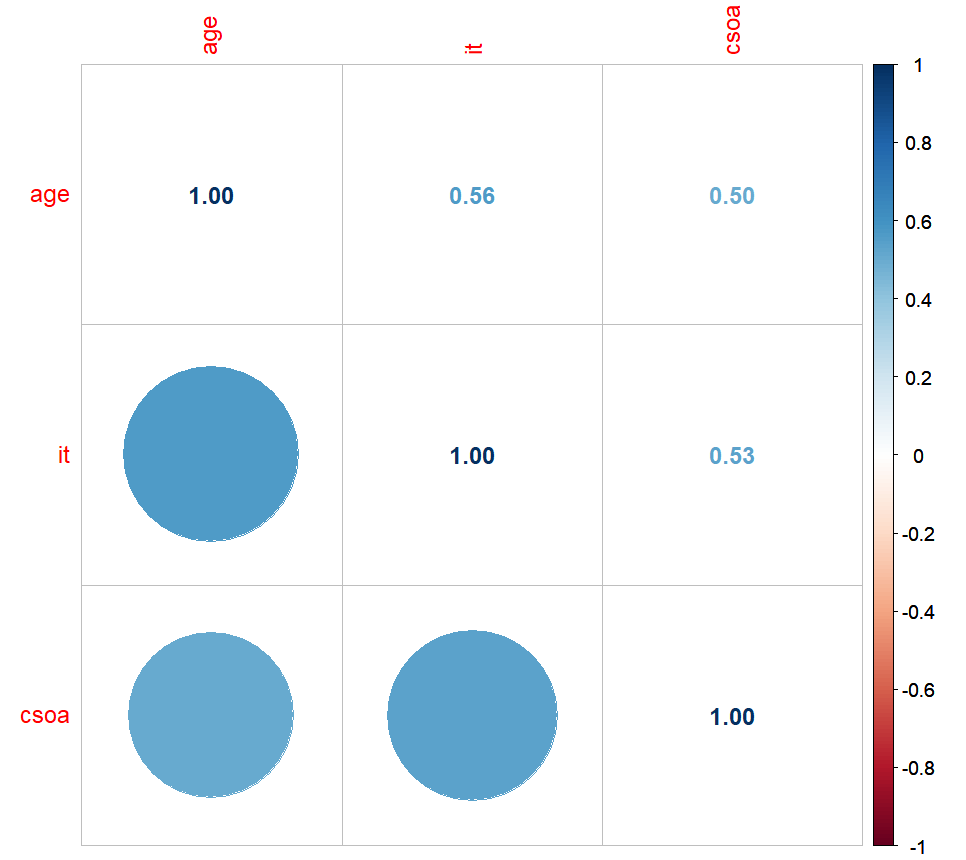


1. 探索描述变量age、it和csoa两两之间是否存在线性关系。



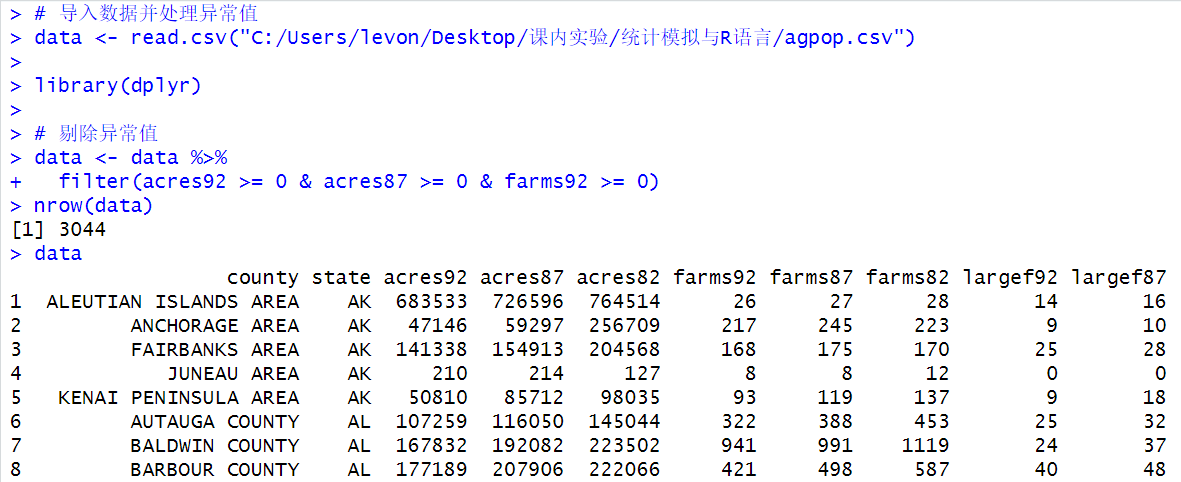






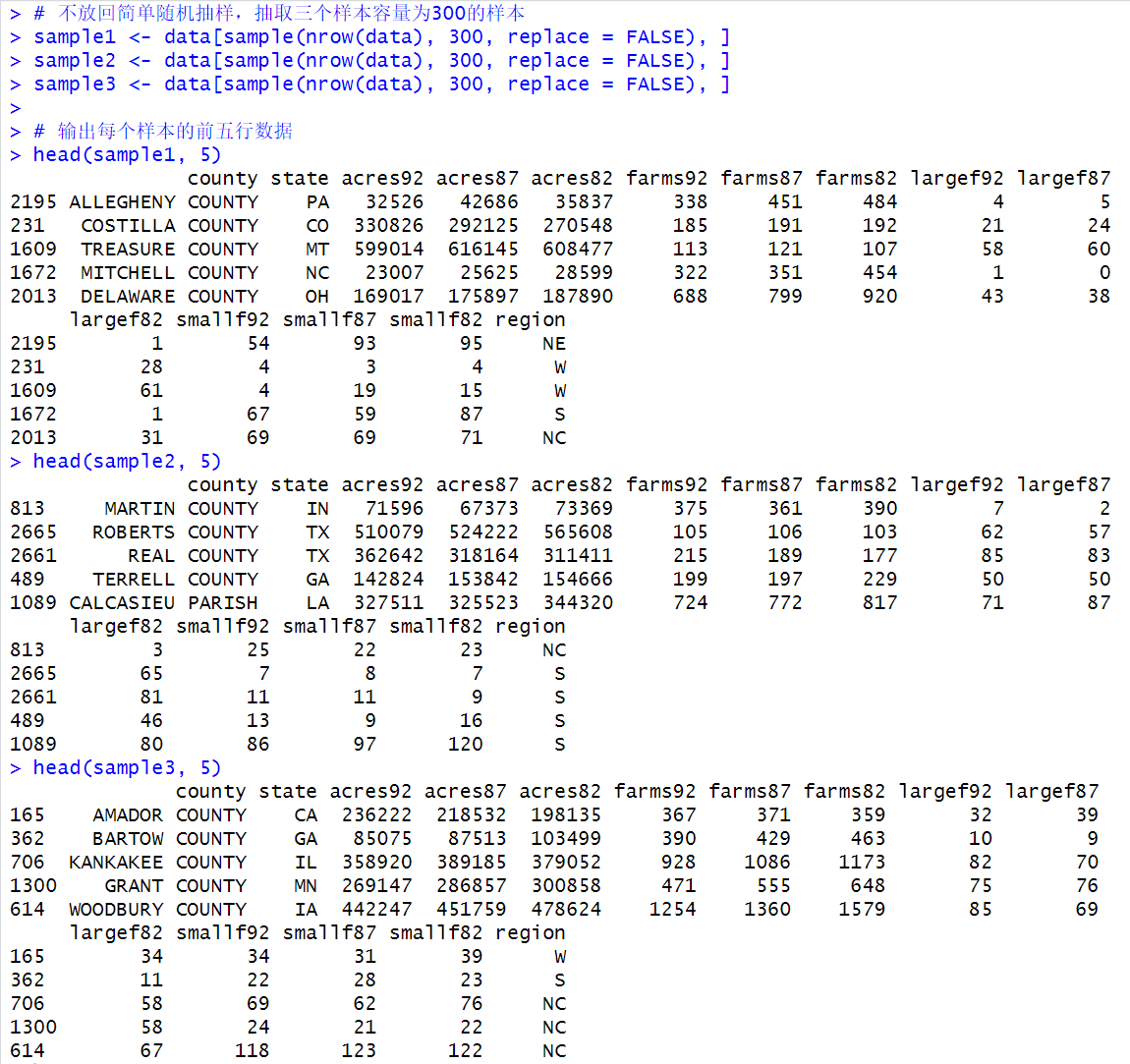
观察矩阵散点图可以发现，age与it, age与csoa的数据点在散点图中呈垂直于x坐标轴分布，但相关系数显示三个星号\*\*\*，这可能意味着这两组变量各自之间存在一种非线性关系或强烈的非线性相关性。而it与csoa的数据点在散点图中整体呈现线性分布，并且皮尔逊相关系数较大，三个“\*”也表明相关性较强，故而可以初步得出结论：it与csoa存在线性关系，age与it, age与csoa不存在线性关系。

1. agpop.csv文件是美国政府每五年做一次有关农业的普查，收集50个州的所有农场的有关数据，共3078个美国县级或与县级规模相当的农场数据。共有4个区域（region/rnum），50个州（state/snum）,以及3041个县（county/cnum）。我们用到的变量有:县（county/cnum）, 州（state/snum）, 区域（region/rnum）,1992年每个县的耕地面积（acres92）, 1987年每个县的耕地面积（acres87）,1992年每个县拥有的农场个数（farms92）。完成下述任务：
   * 1. 导入原始文件并对进行异常值做剔除处理；

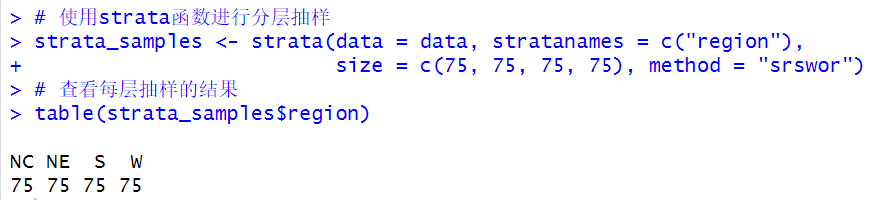


* + 1. 利用以下抽样方法，获取抽样样本：

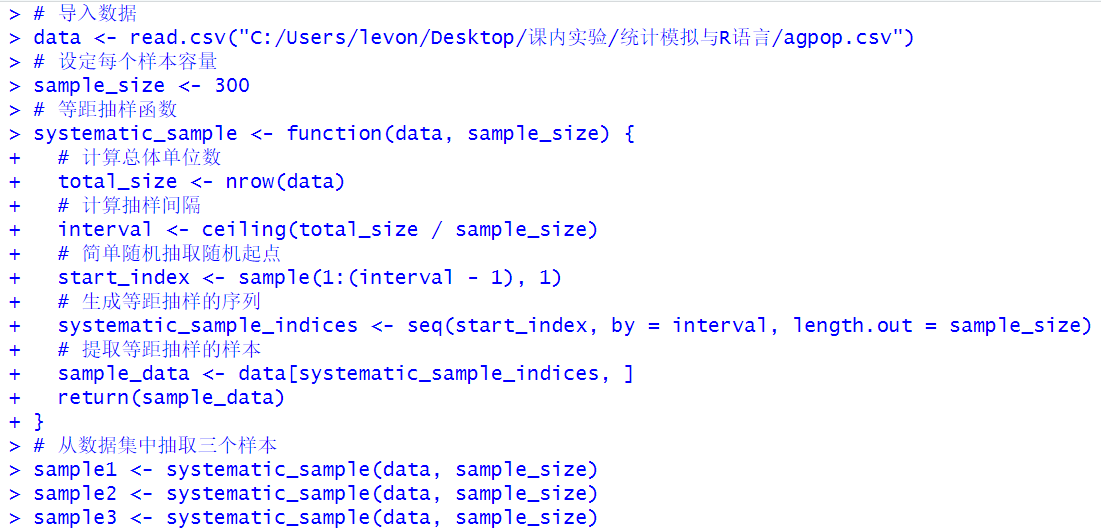
i. 利用不放回简单随机抽样，抽取三个样本容量为300的样本，并分别输出每个样本的前五行数据；

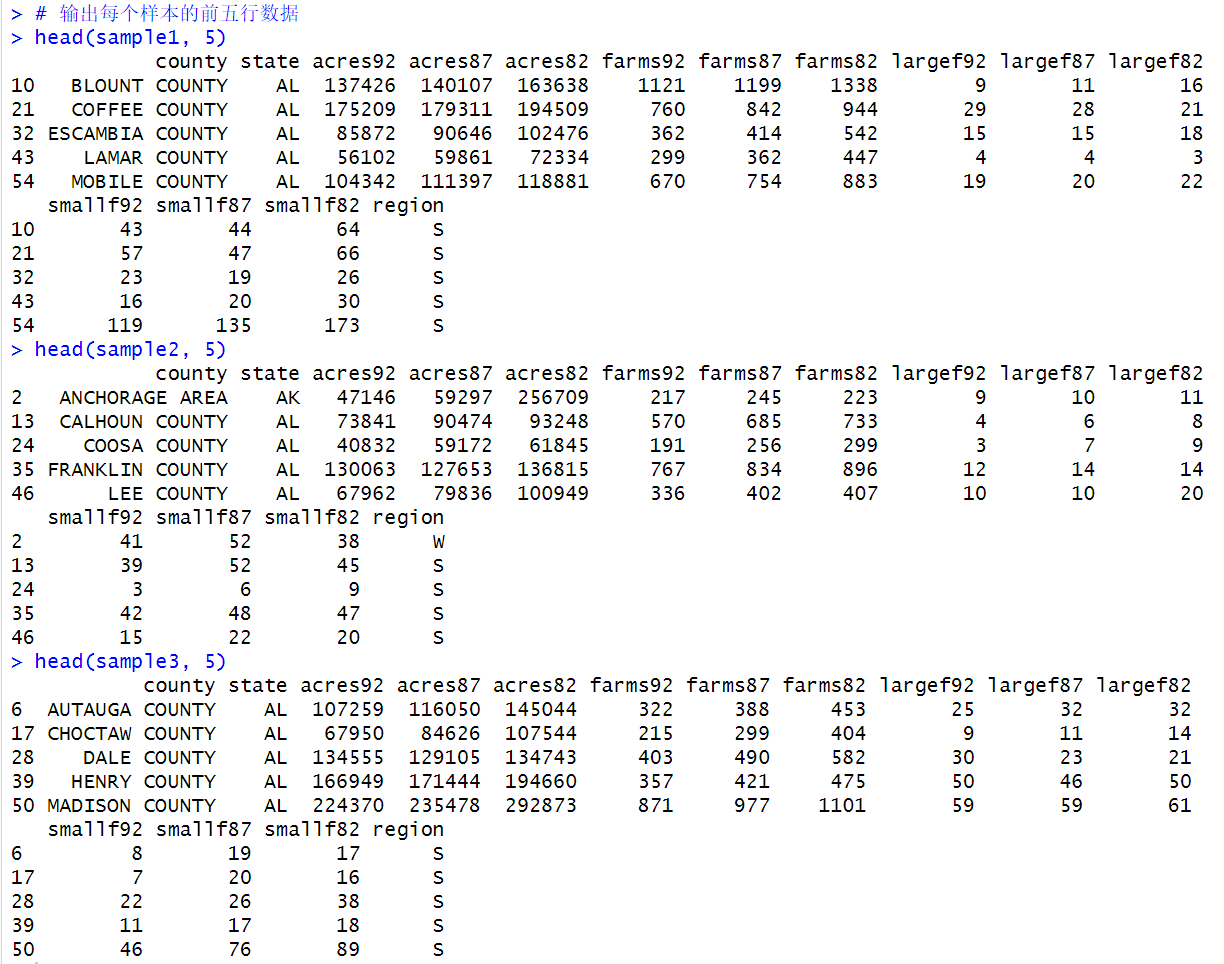


ii. 使用分层抽样方法，以‘region’为分层变量，从每层简单随机抽取75个样本单元，构成一个样本，并查看样本中‘region’每层的样本数量；



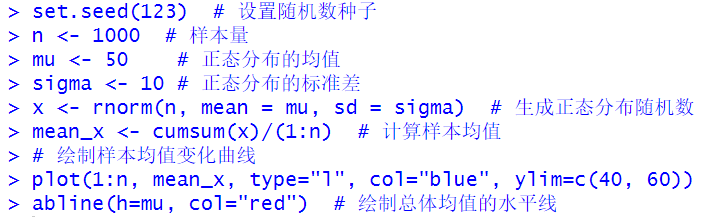
iii. 采用等距抽样方法抽取三个样本容量为300的样本，并分别输出每个样本的前五行数据；（注：等距抽样方法，也称系统抽样法，先将总体的全部单元按照一定顺序排列，采用简单随机抽样抽取第一个样本单元(或称为随机起点)，再顺序抽取其余的样本单元）。

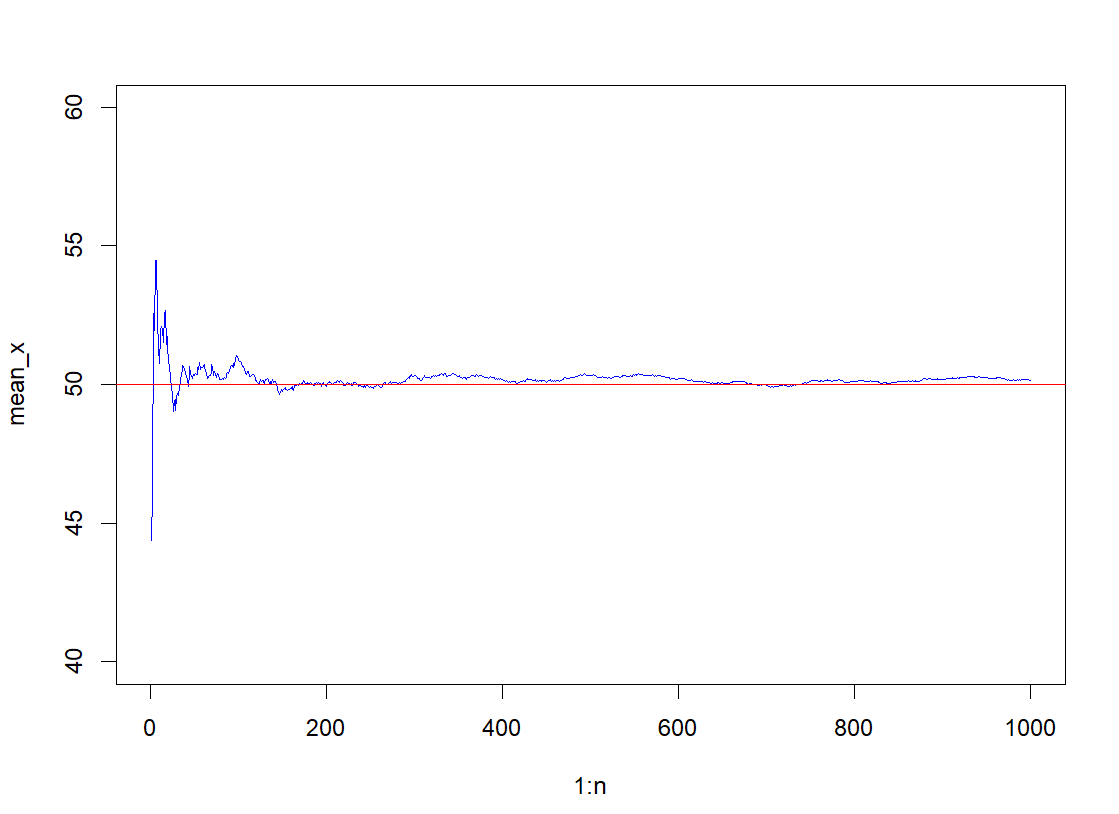




1. 蒙特卡洛模拟是一大类计算算法的集合，依靠重复的随机抽样来获得数值结果。在统计学中，蒙特卡洛模拟也被广泛应用，可以用来验证各种统计定理和方法的正确性。请自学下面内容https://zhuanlan.zhihu.com/p/619959058，实现用蒙特卡洛模拟验证大数定理和中心极限定理的基本操作。

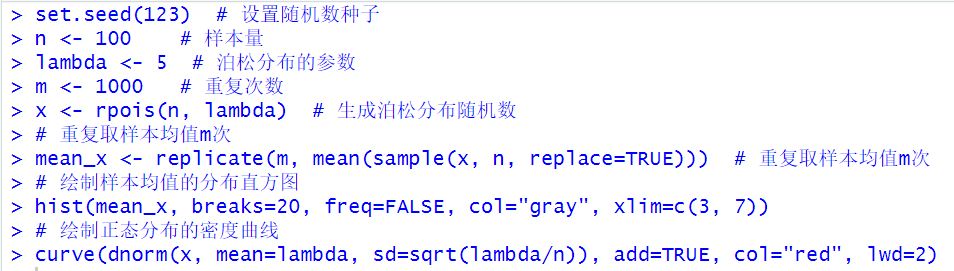
**·大数定律**

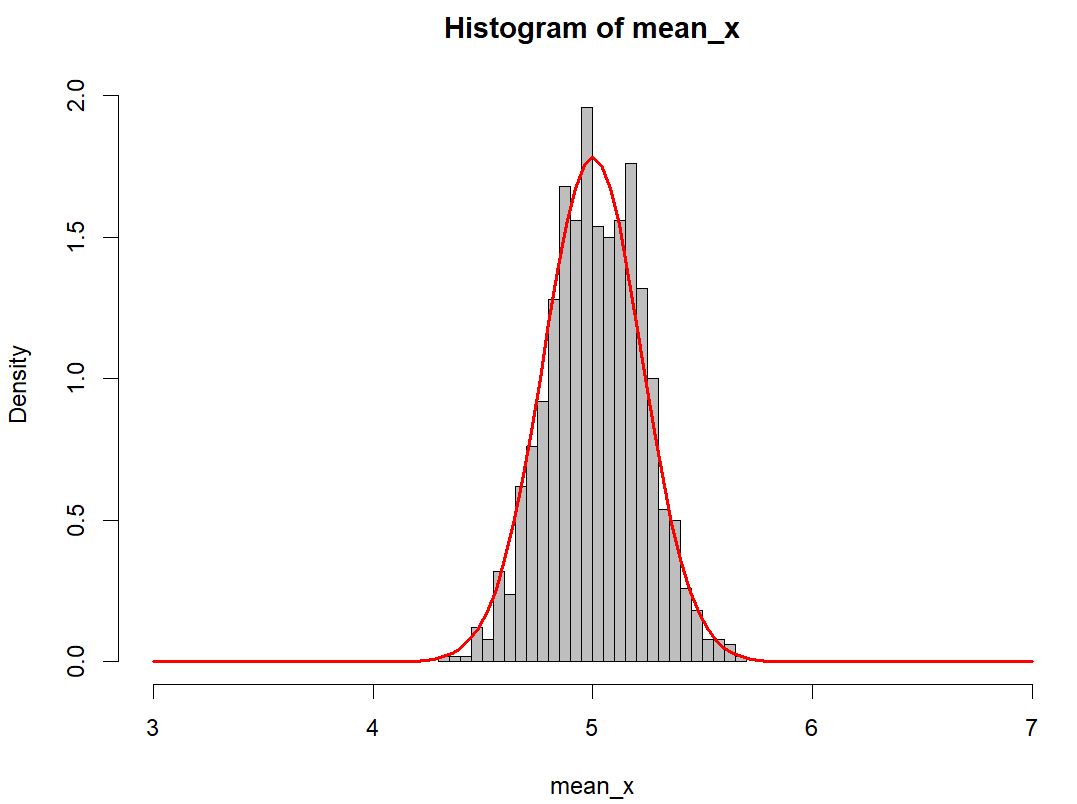




本例中生成了一组服从正态分布的随机变量，可以观察到，样本均值随着样本量的增加而趋近于该正态分布的均值。

**·中心极限定理：**

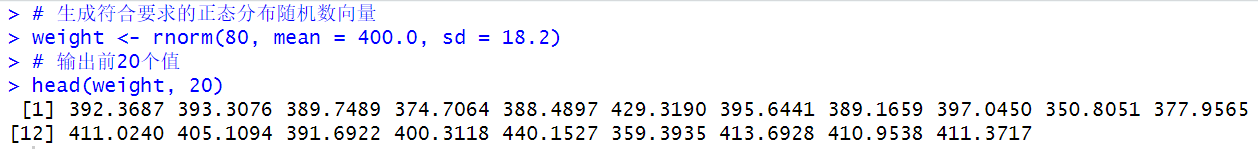




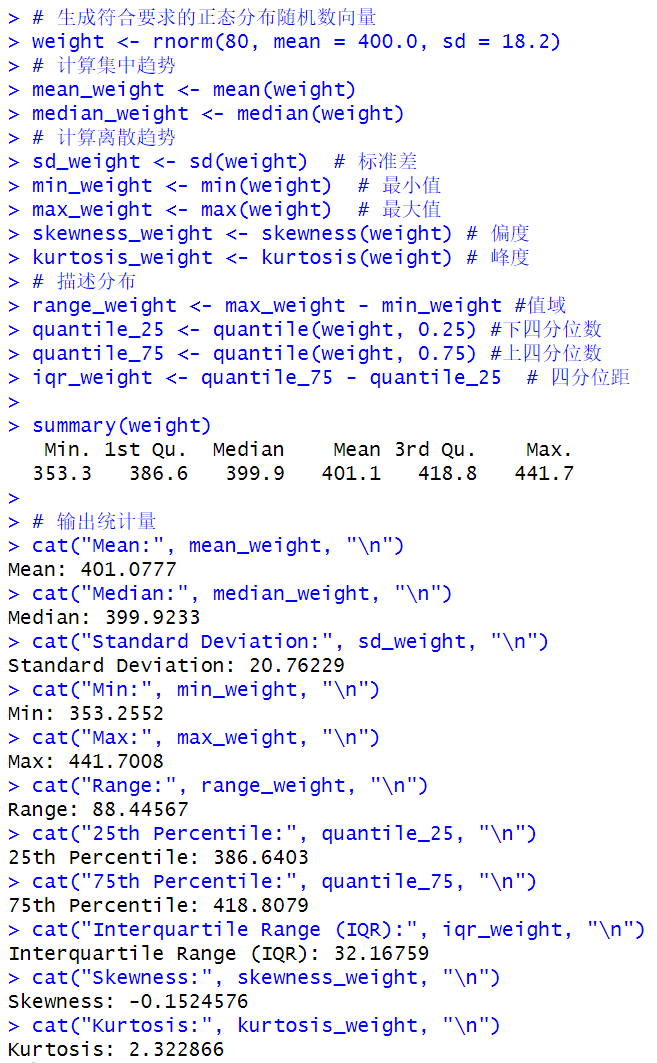
本例中生成了一组服从泊松分布的随机变量，可以观察到样本均值的分布趋近于正态分布。

【作业】

1. 我们以奶牛为例，执行下列操作：
   1. 生成如下的数据：80个正态分布的随机数（即长度为80的向量），该正态分布满足均值为400.0以及标准差为18.2的条件。将该向量命名为‘weight’并输出他的前20个值。

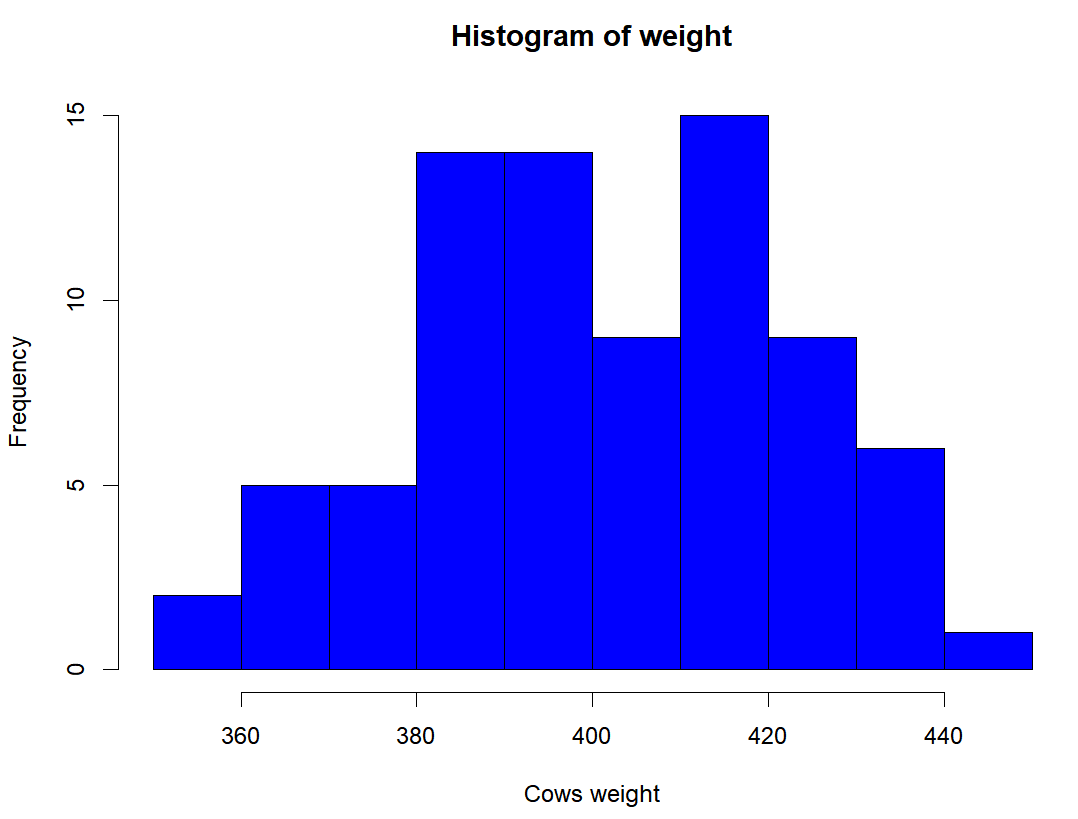


* 1. 计算并输出反映‘weight’向量数据的集中趋势、离散趋势和描述分布的各统计量。



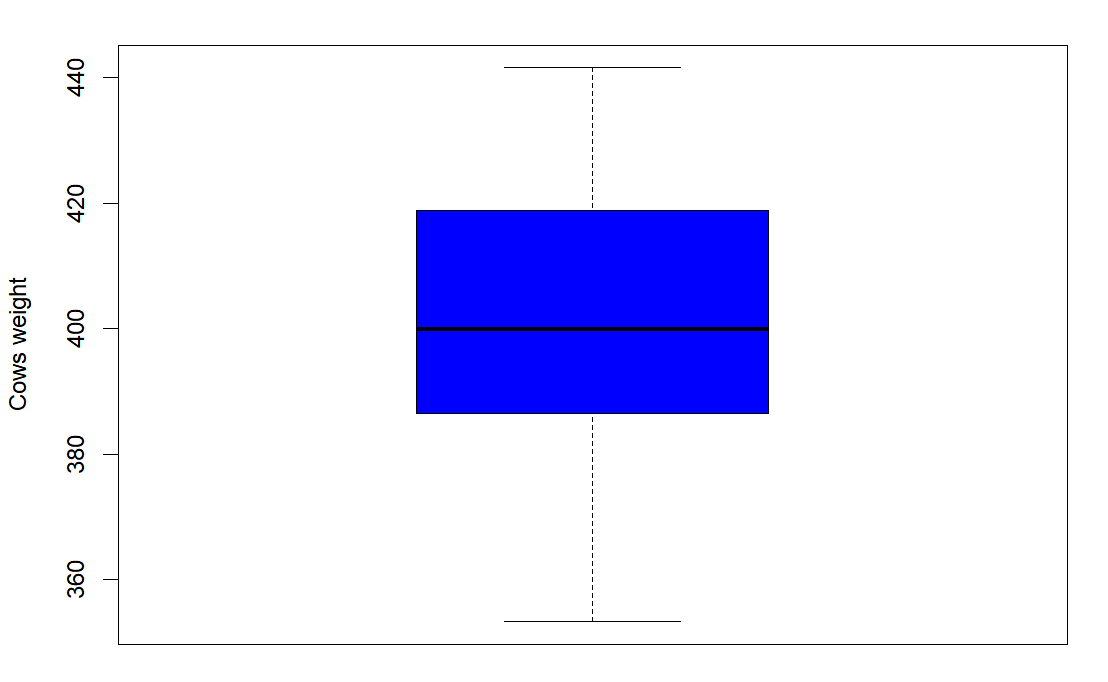
* 1. 通过‘weight’向量生成一个直方图。要求x轴标签为‘Cows weight’且直方图的柱状体为蓝色。

IMG_256



* 1. 生成‘weight’向量的箱线图。要求y轴对应标签为‘Cows weight’，设置颜色为蓝色。

IMG_256



* 1. 分析c,d绘制的两种图。通过你的分析，那种图生成了更多的信息？

直方图：

显示了数据的分布情况，可以观察到数据的集中区域、分布的形状以及是否有异常值。直方图通过柱状图的形式，直观地展示了数据的频率分布。

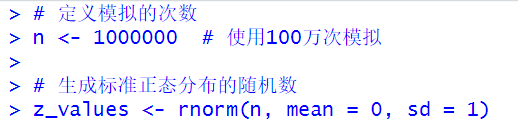
箱线图：

提供了数据的五数概括（最小值、第一四分位数、中位数、第三四分位数、最大值），并且通过箱体和须（whiskers）的长度，可以直观地看出数据的离散程度和异常值。

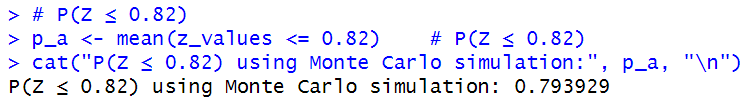
结论：

箱线图在识别异常值和数据集中趋势方面提供了更多信息，而直方图则在展示数据分布形状方面提供了更多信息。

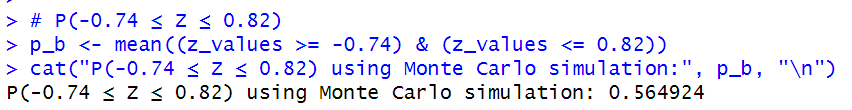
1. 设Z〜N（0,1）。请使用R模拟(通过产生的随机数求得) 来计算以下概率。



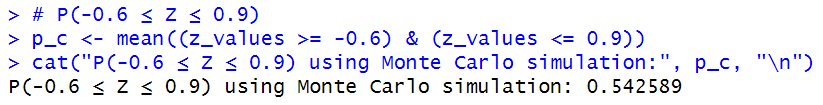
1. P(Z ≤ 0.82)



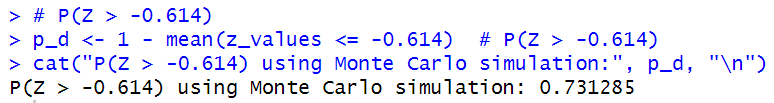
1. P(-0.74 ≤ Z ≤ 0.82)



1. P(-0.6 ≤ Z ≤0.9)

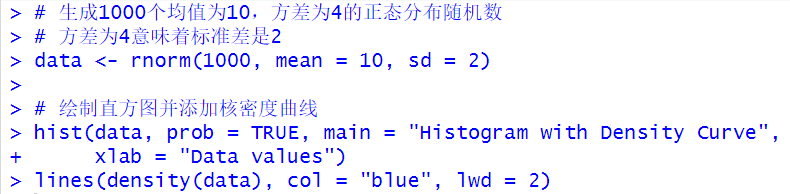


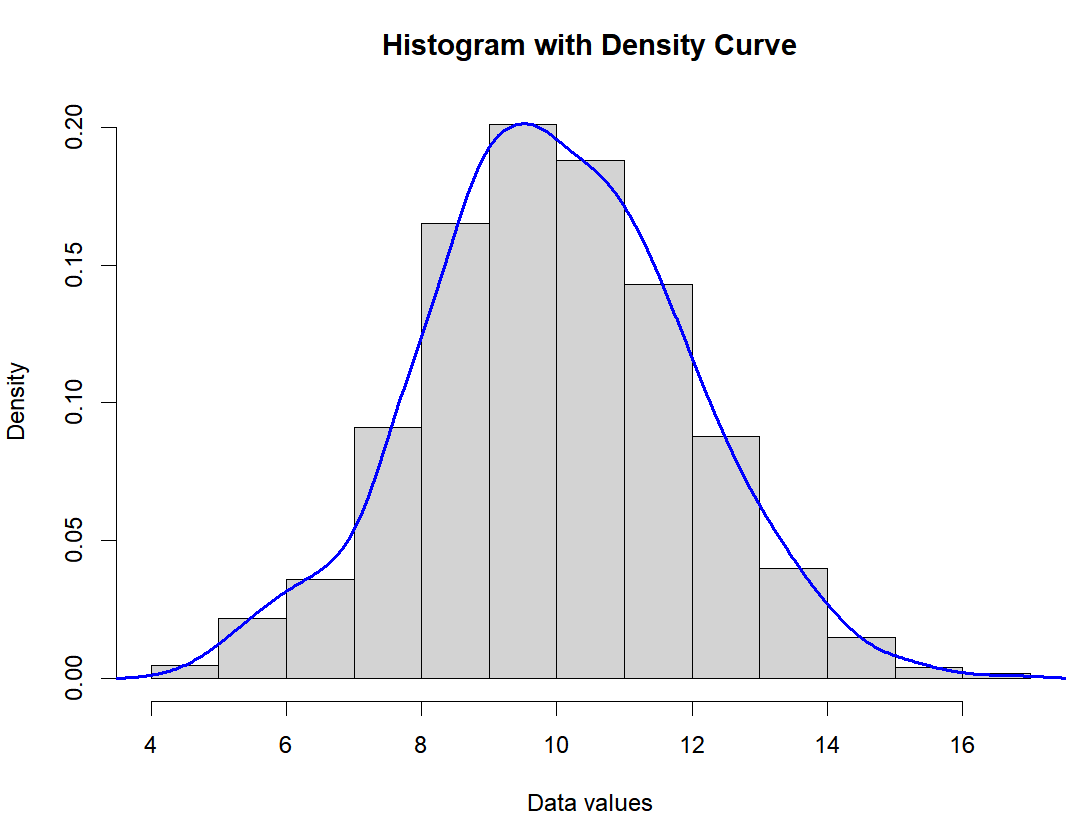
1. P(-0.614 ≤ Z)

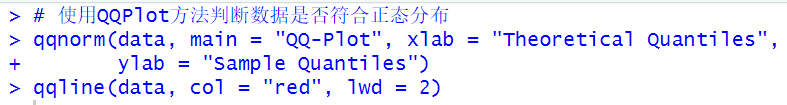


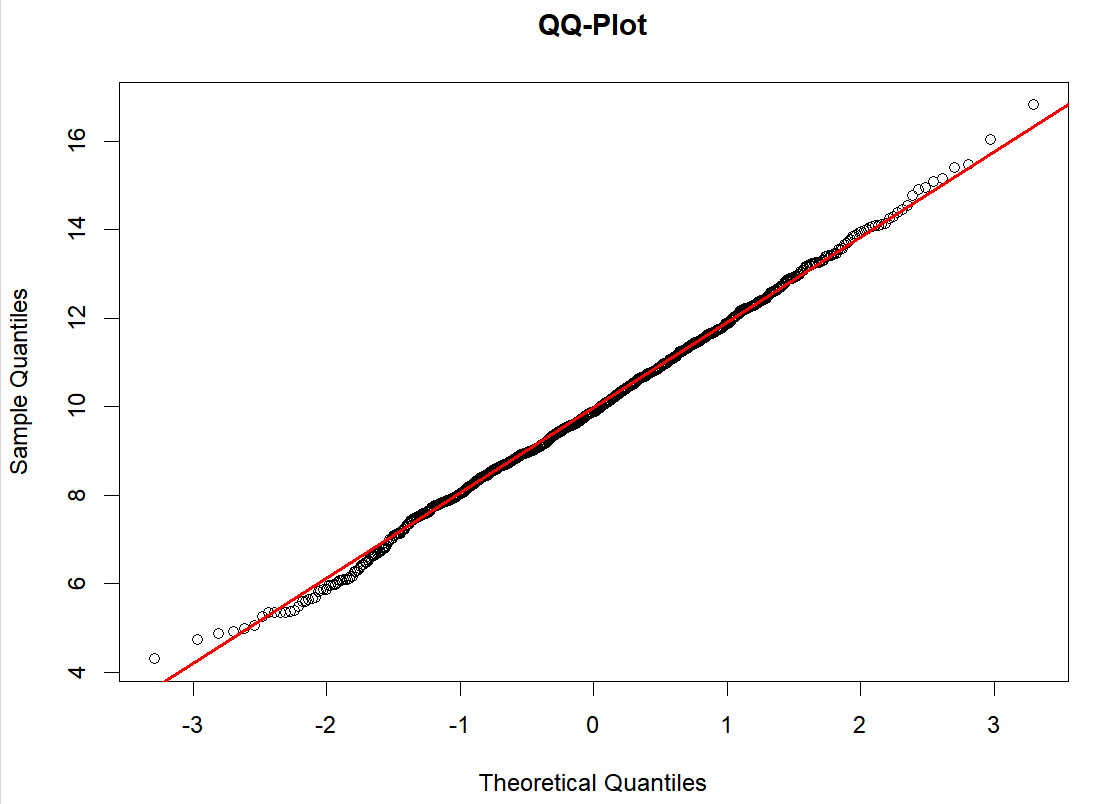
1. 用命令rnorm( )命令产生1000个均值为10, 方差为4的正态分布随机数,

用直方图呈现数据的分布并添加核密度曲线，并用QQPlot方法判断该组数据是否符合正态分布。





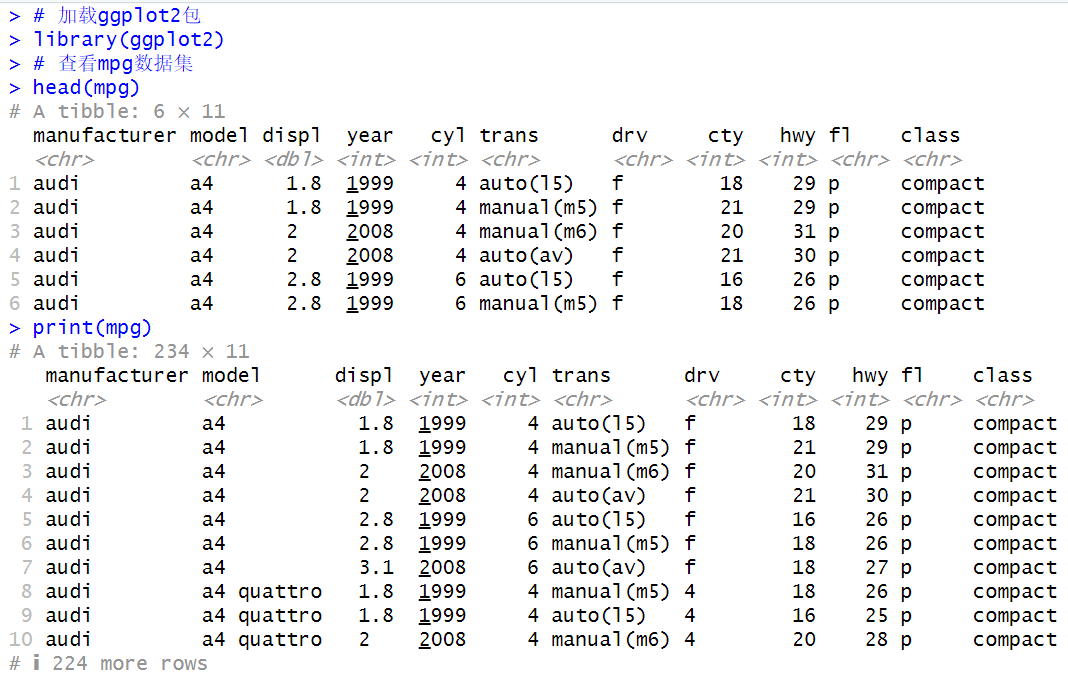




观察QQ图后发现，大部分的数据点都落在参考线上或者接近这条线，因此可以认为数据近似符合正态分布。

1. 对ggplot2 package里面自带的 mpg 数据进行描述性统计分析:

(1) 理解并描述数据集：请用R代码查看该数据集、并用简练的语言描述出该数据集的基本信息（包括但不限于观察值和变量数目、类型，各变量代表什么含义）；



·观察值数目与类型：数据集包含234个观察值（即234辆车的数据）。

·变量数目：数据集共有11个变量。

·变量类型与含义：

manufacturer: 汽车制造商的名称，属于字符类型。

- model: 汽车的型号，属于字符类型。

- displ: 发动机排量，单位为升，属于数值类型。

- year: 汽车生产的年份，属于数值类型。

- cyl: 发动机的气缸数，属于数值类型。

- trans: 变速器类型，属于字符类型，可能的值包括自动（auto）、手动（manual）和半自动（semi-auto）。

- drv: 驱动类型，属于字符类型，可能的值包括前驱（f）、后驱（r）和四驱（4）。

- cty: 城市行驶时的燃油效率，单位为每加仑英里数，属于数值类型。

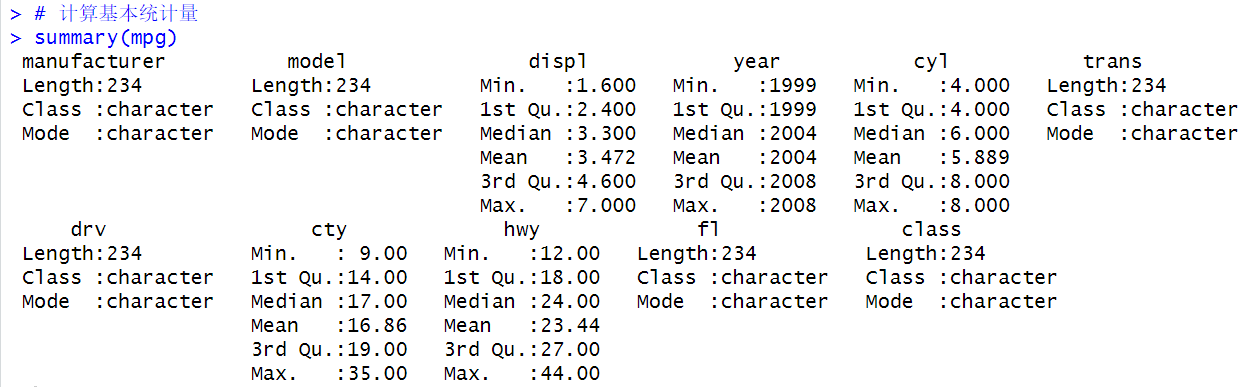
- hwy: 高速公路行驶时的燃油效率，单位为每加仑英里数，属于数值类型。

- fl: 燃油类型，属于字符类型，可能的值包括柴油（diesel）、汽油（gas）、乙醇（ethanol）等。

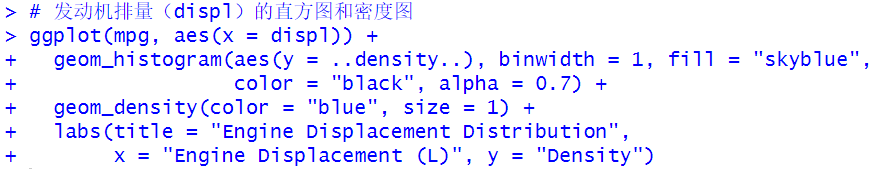
- class: 汽车的类别，属于字符类型，可能的值包括紧凑型（compact）、中型（midsize）、SUV（suv）等。

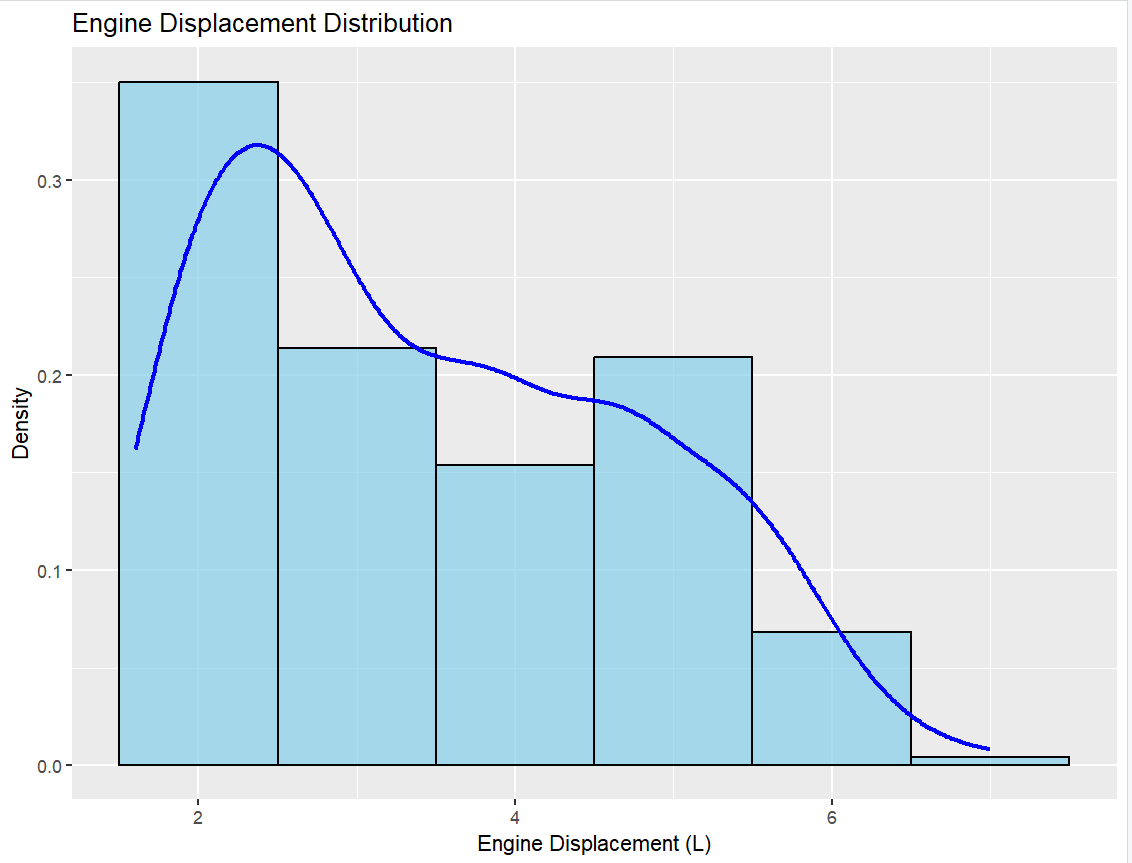
1. 计算每个变量的基本统计量，数据分布，用图、表形式展示（注：描述离散变量和连续变量的统计量和数据分布是不一样的）；

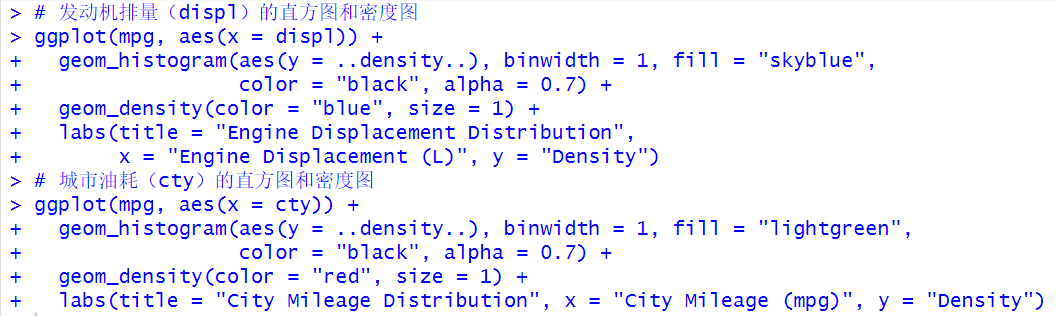
·统计量：

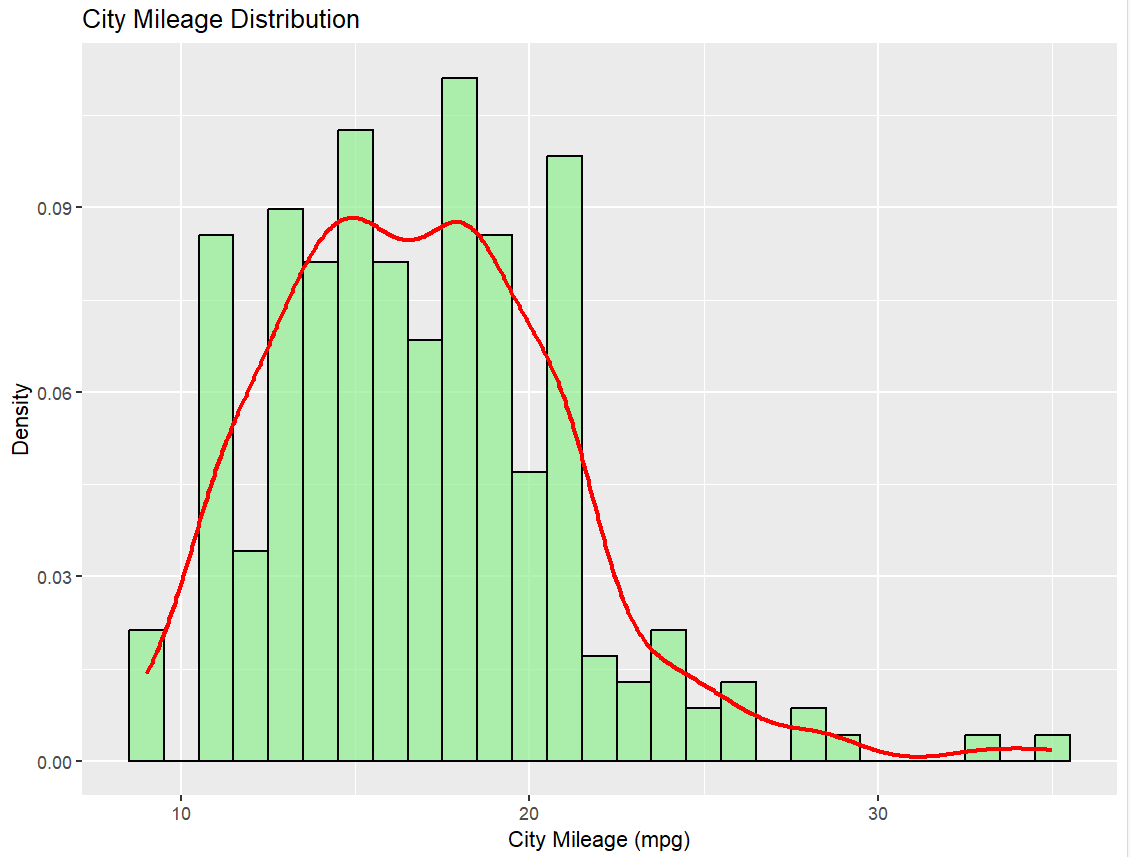


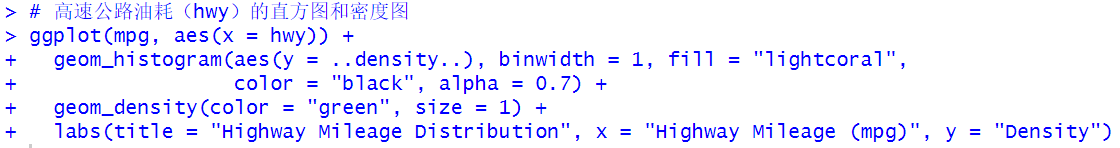
·数据分布：

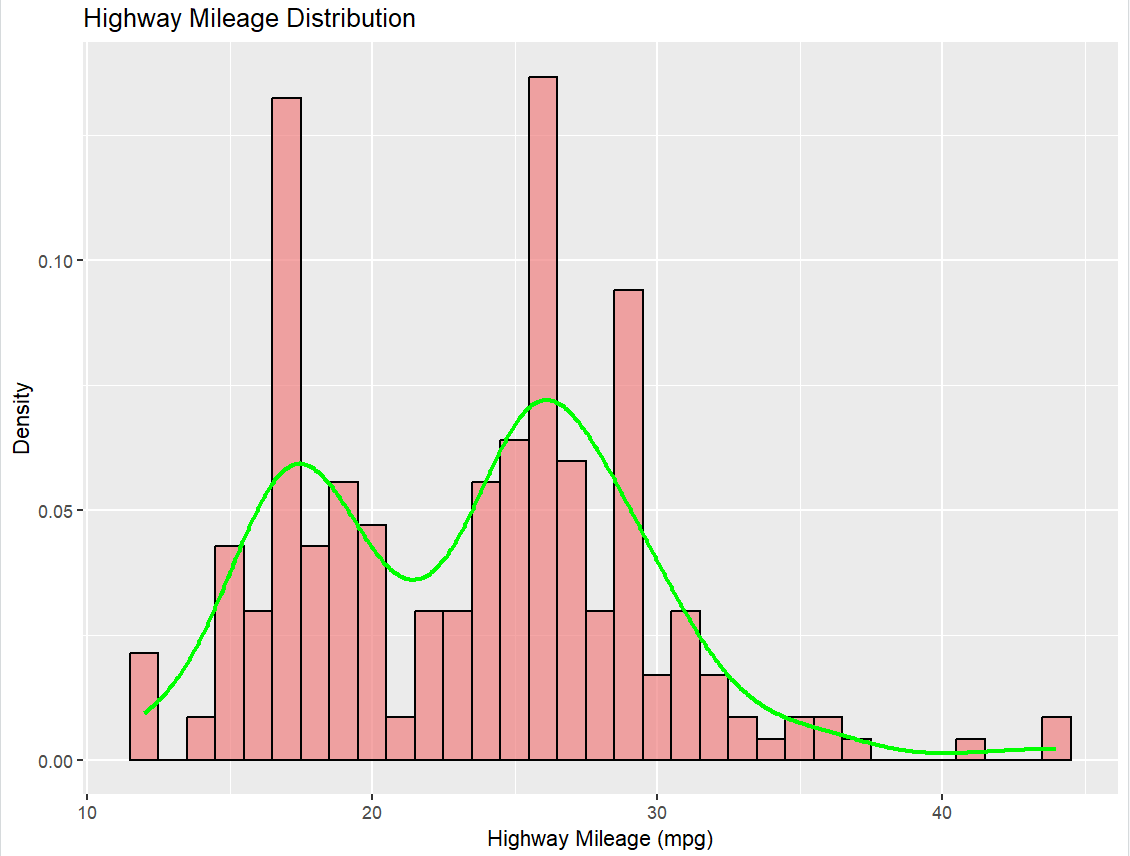


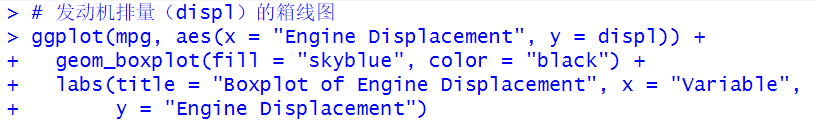


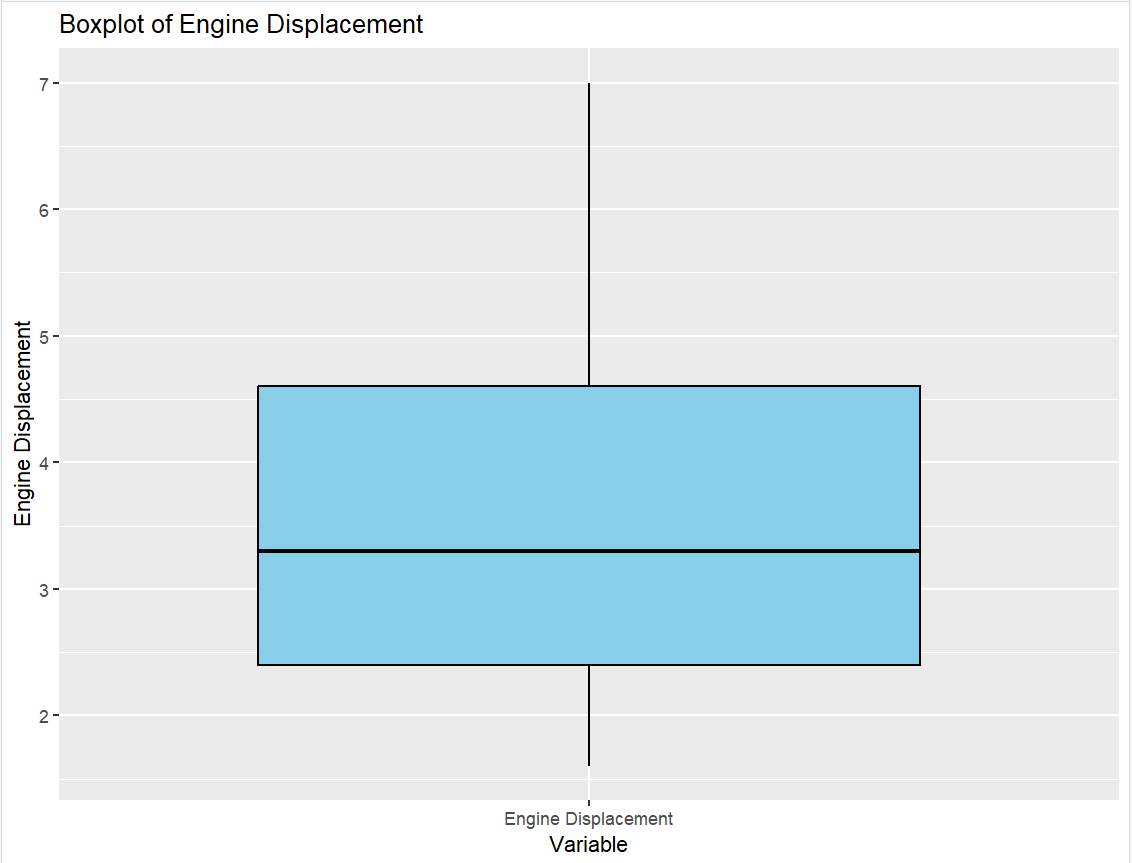


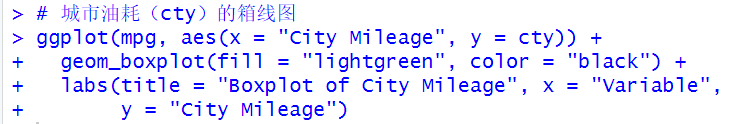


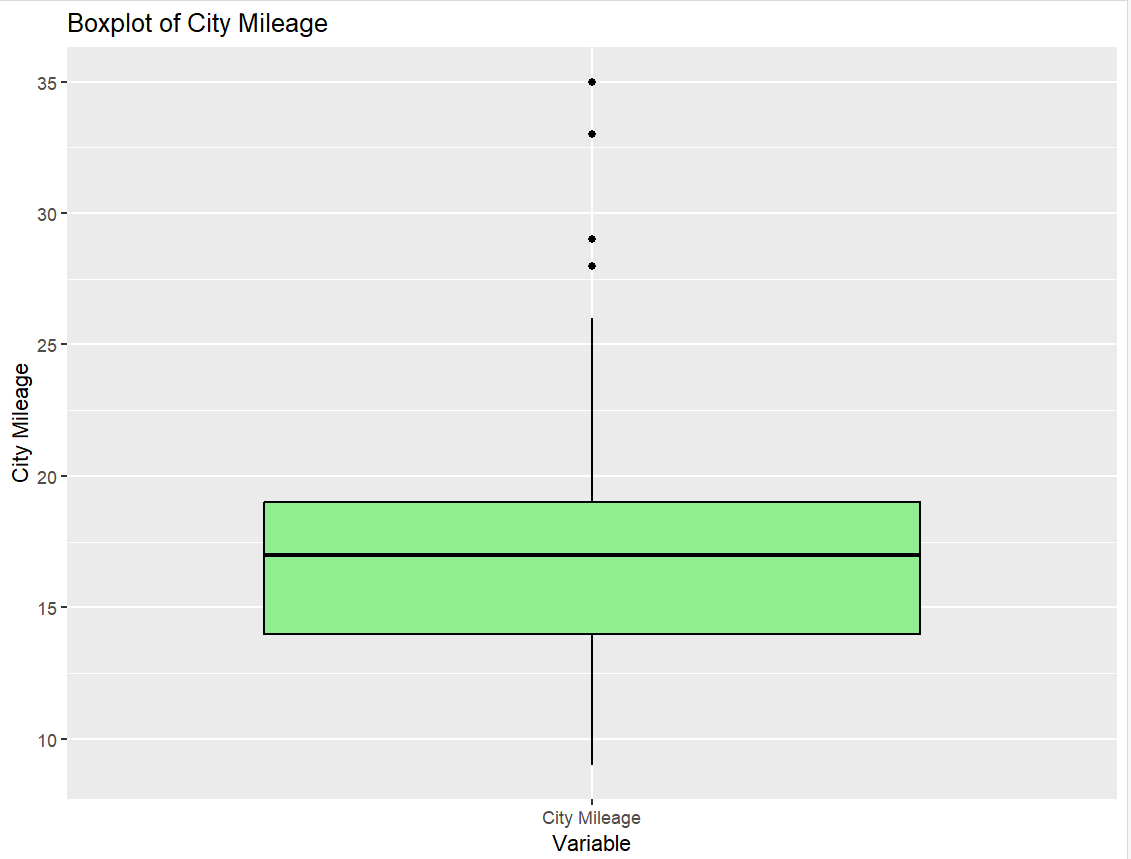


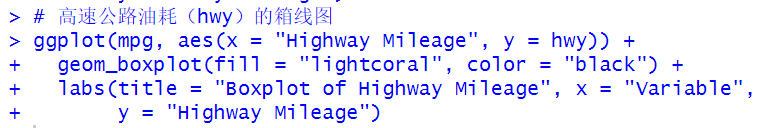


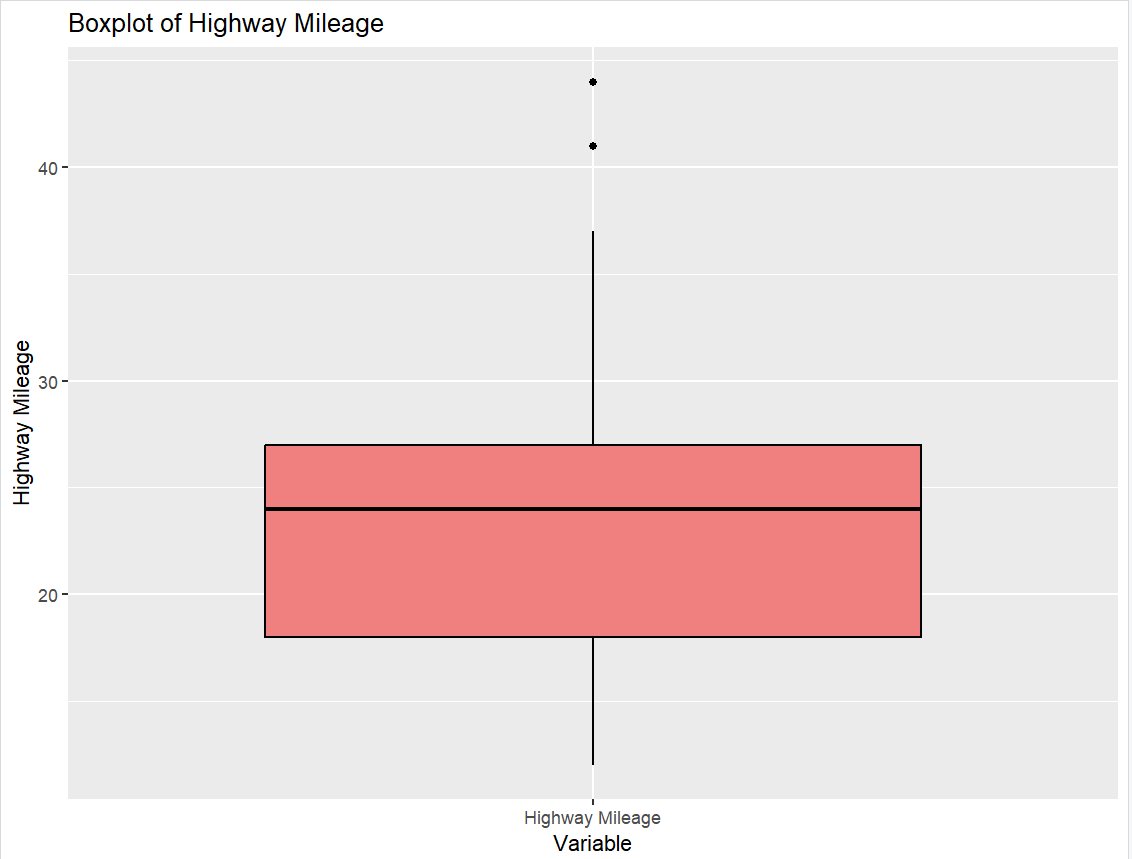


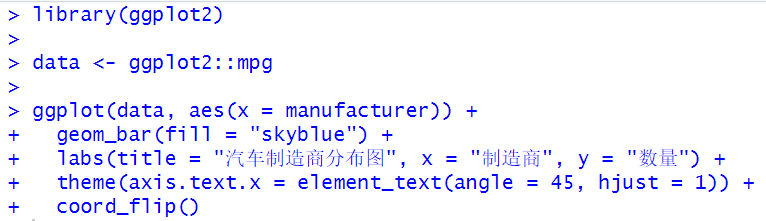


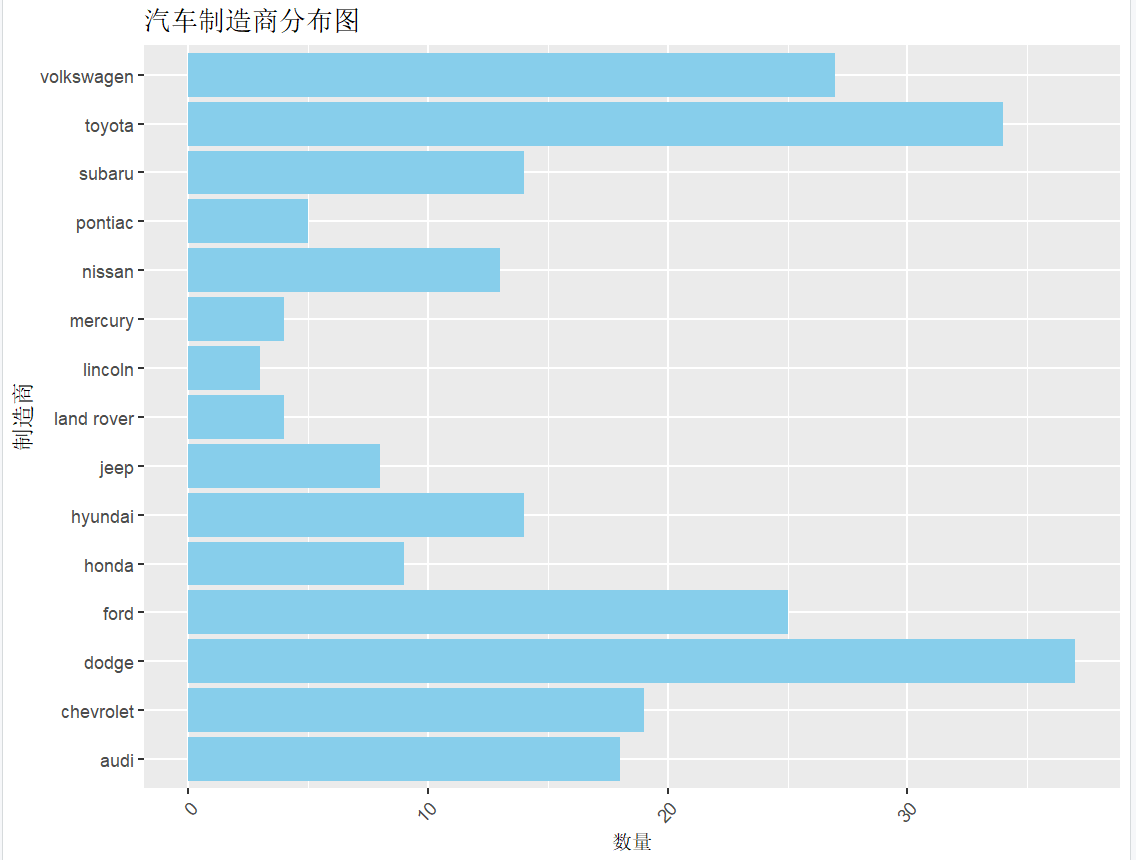


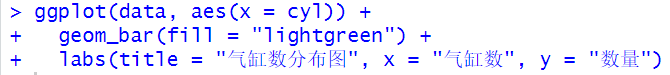


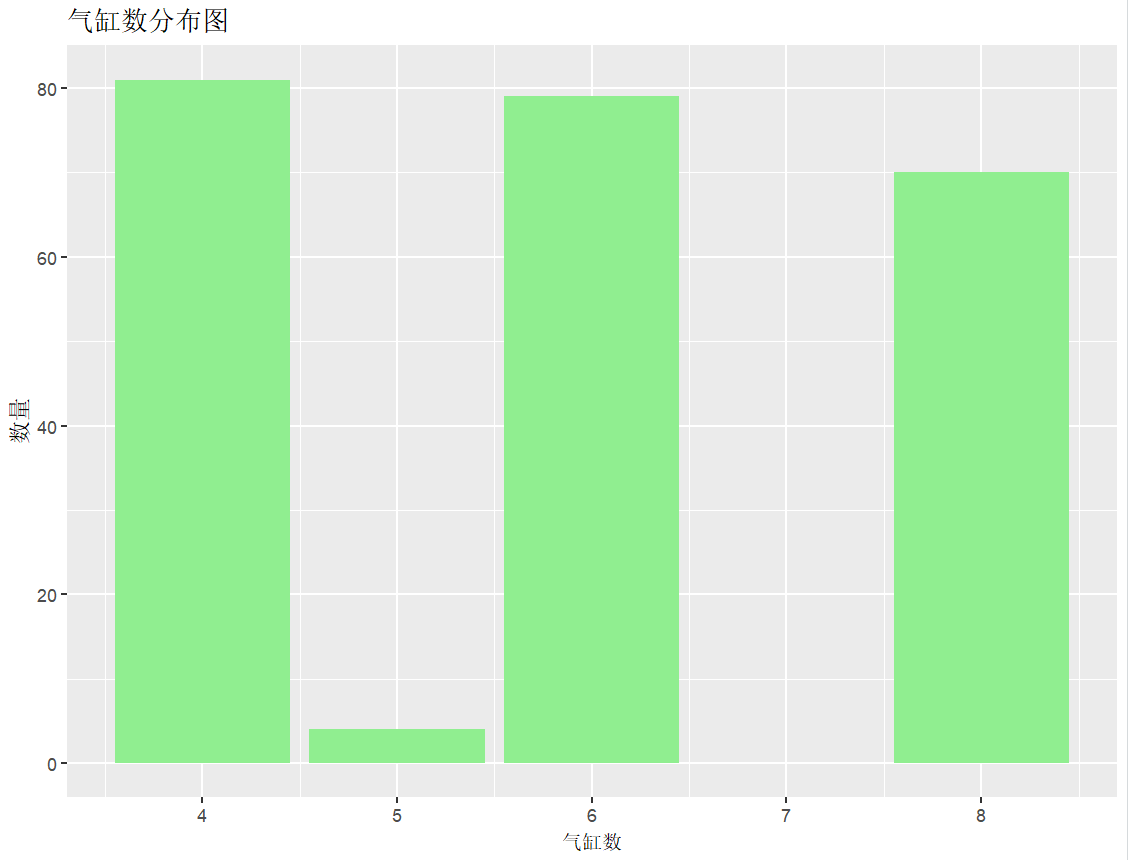


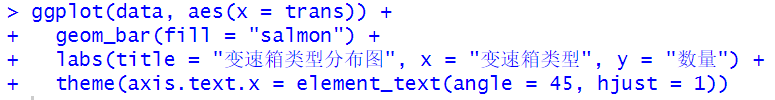


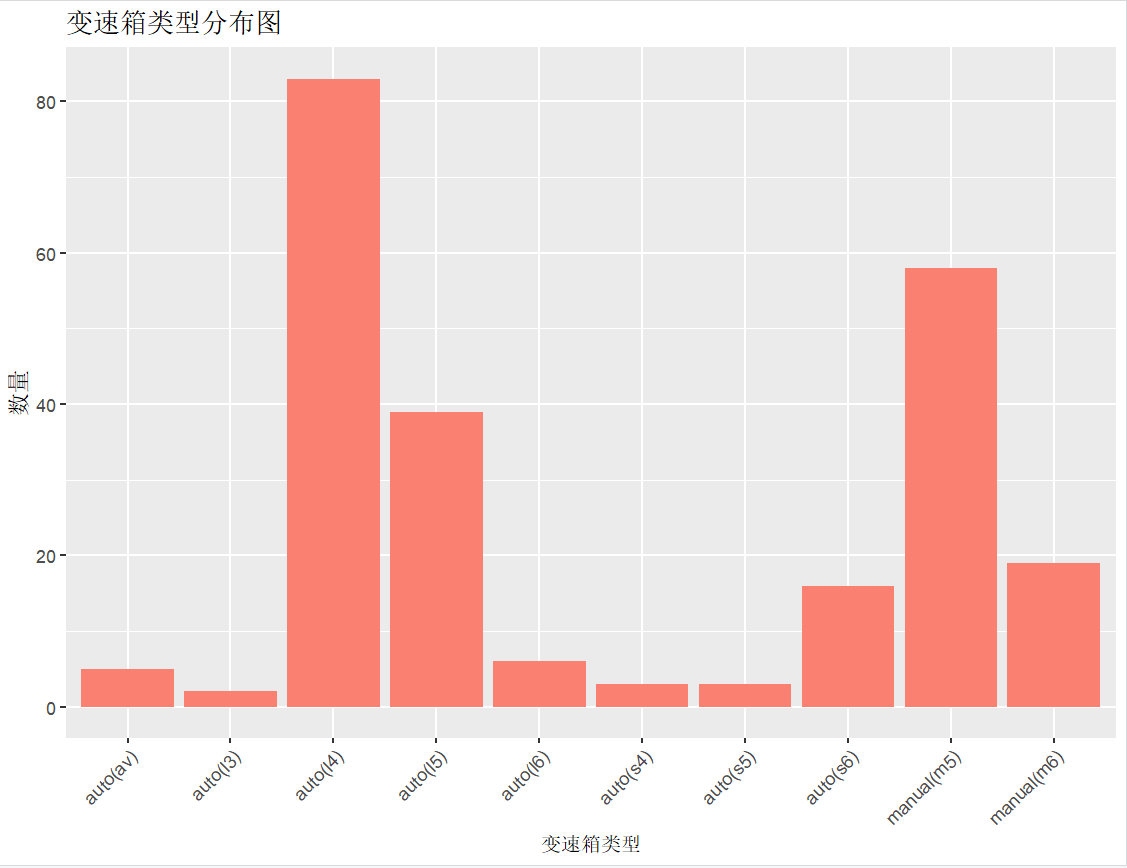


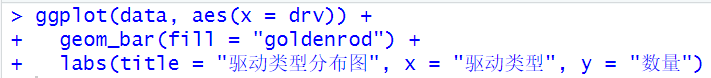


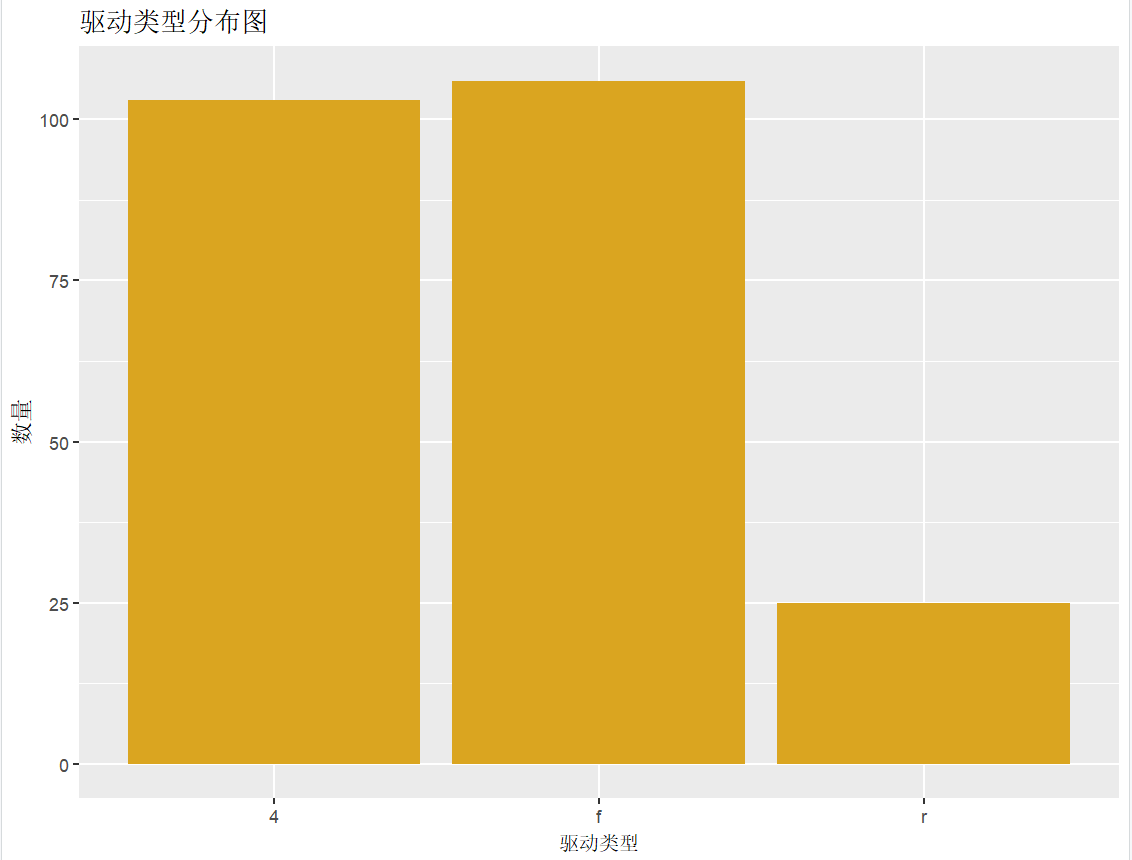


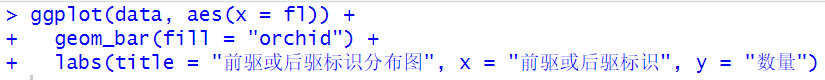


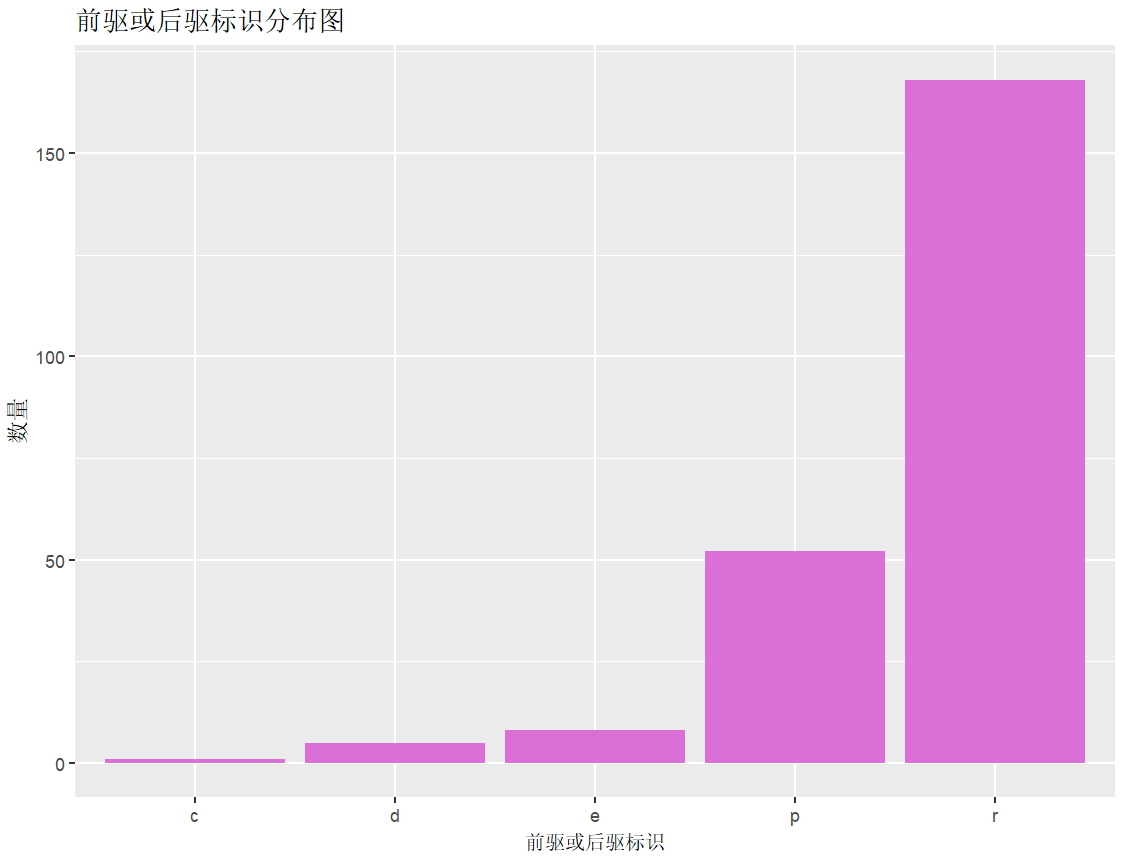


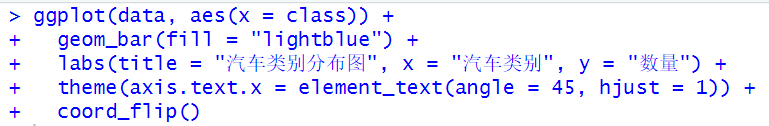


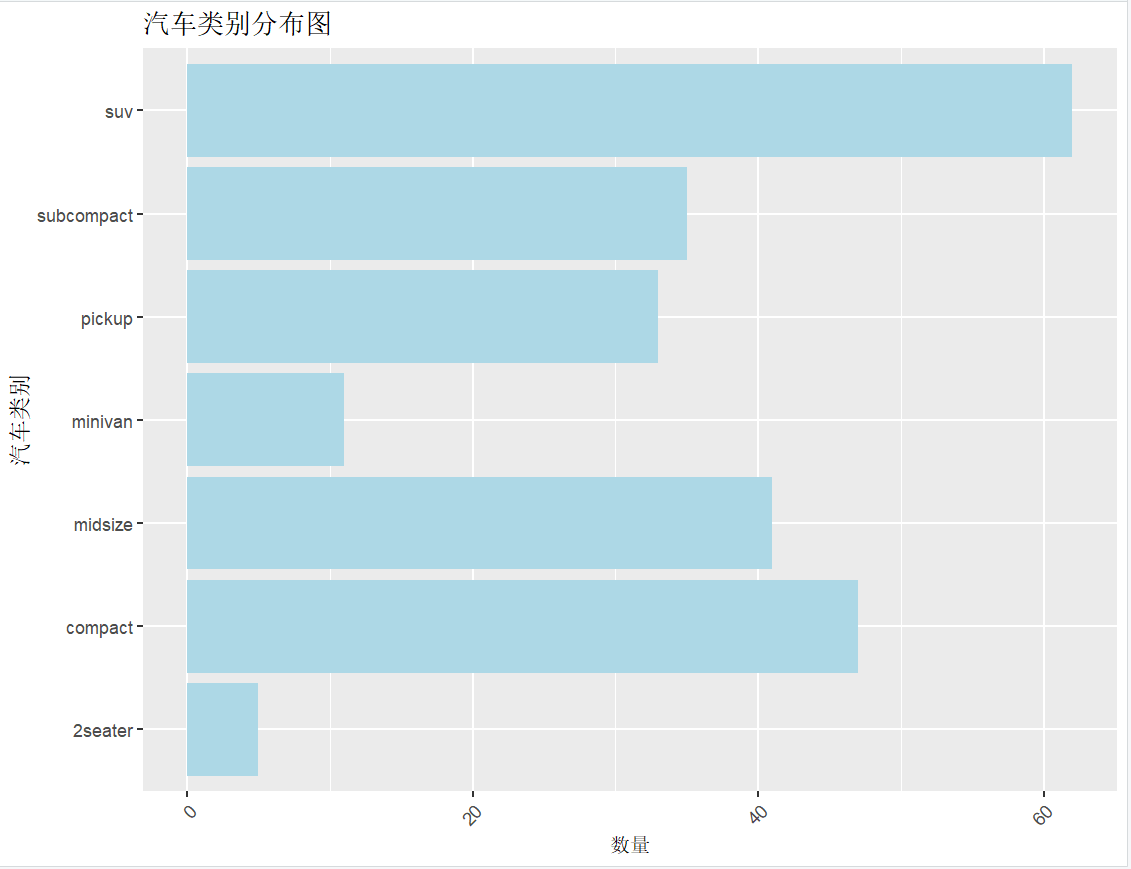


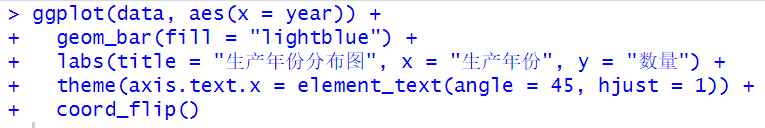


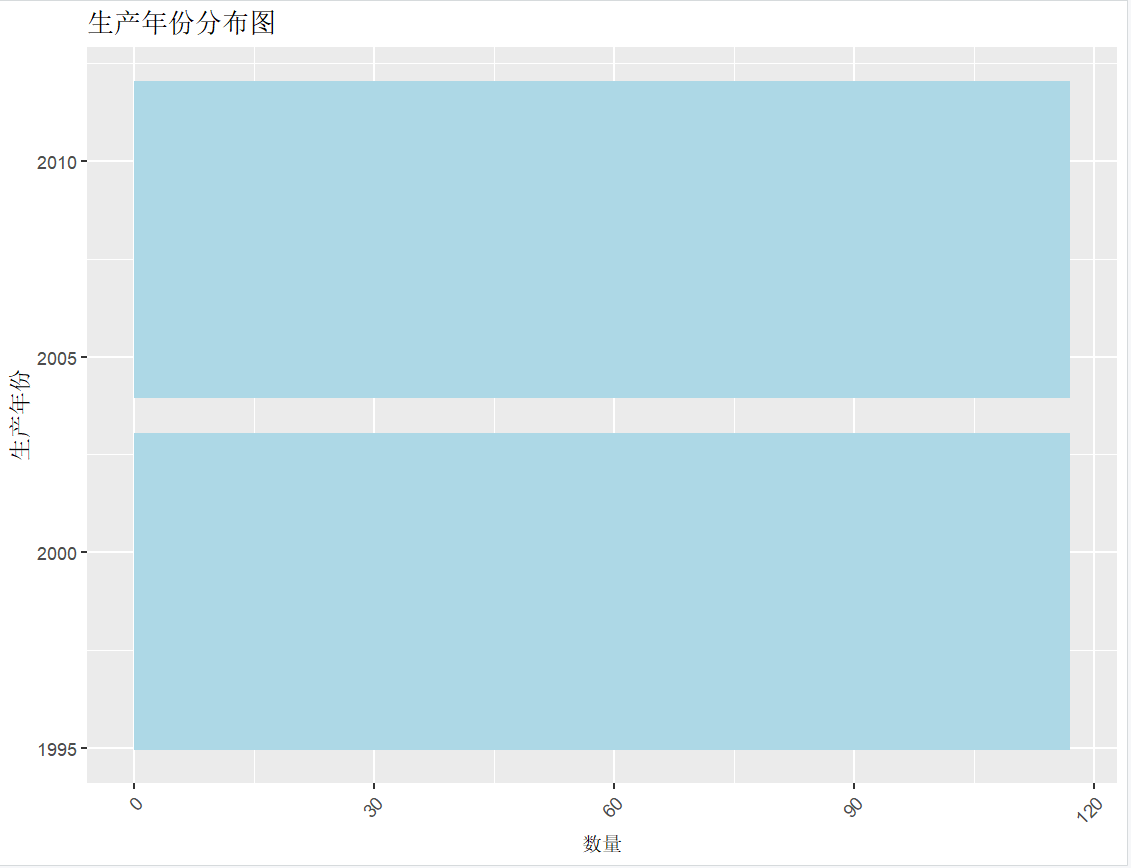


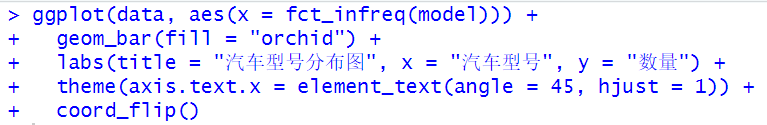


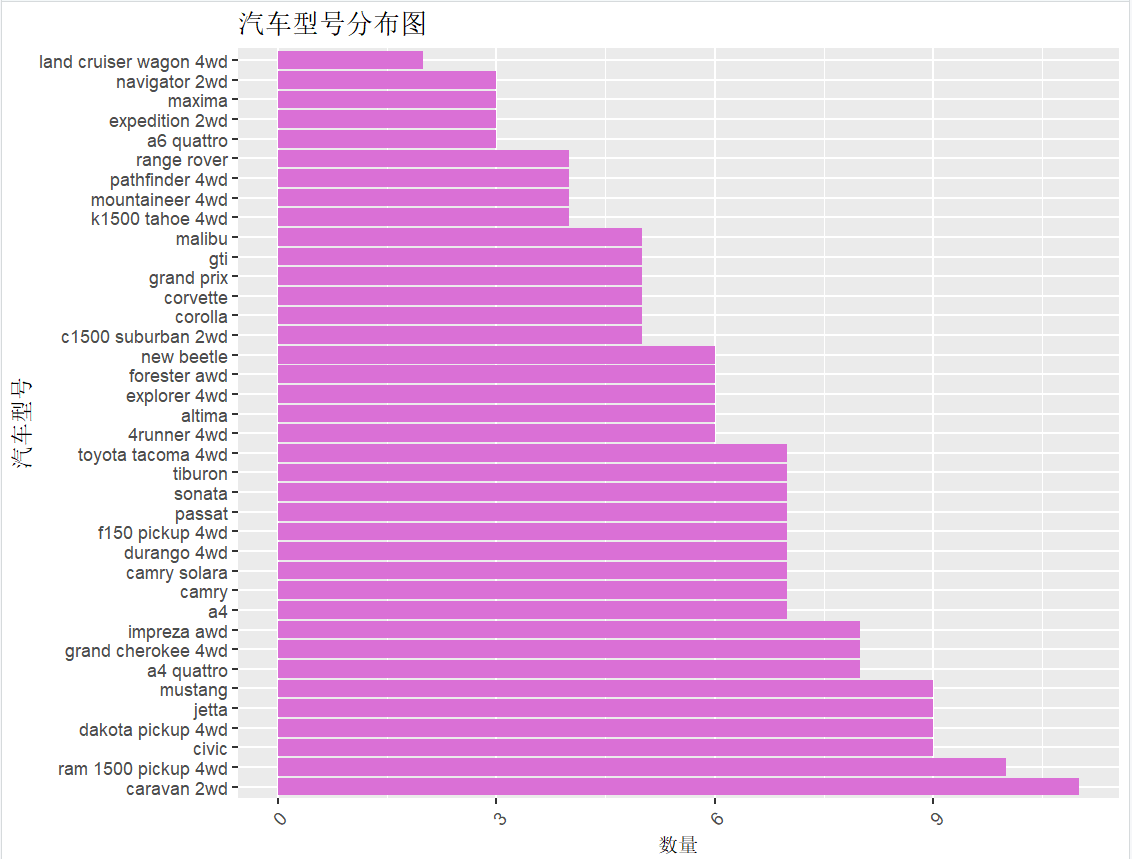






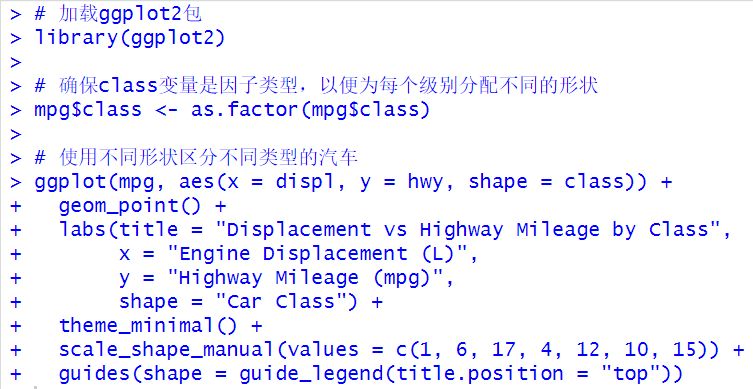


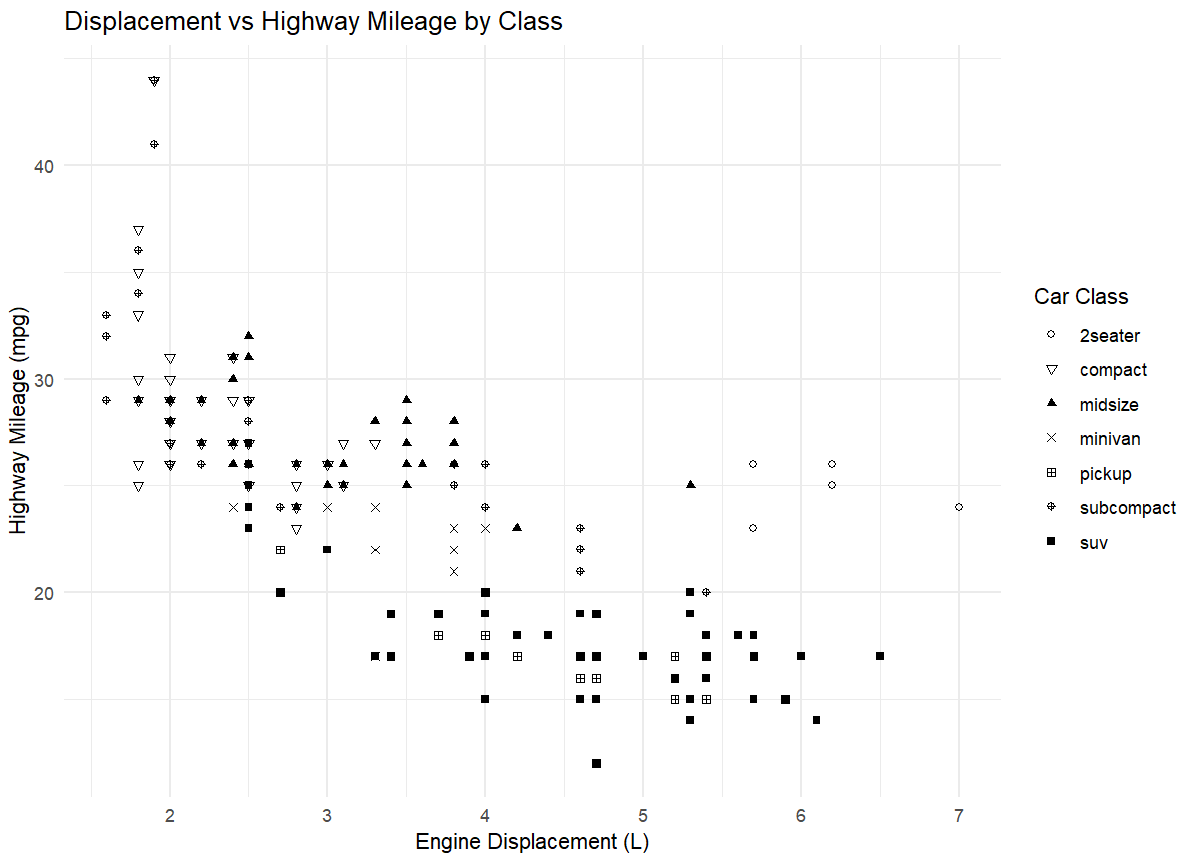




1. 利用普通散点图和条件散点图方式探索不同类型汽车（class）的排量（displ）和每加仑高速里程（hwy）的关系，并用简短的文字描述一下（注：普通散点图用不同颜色或者形状的点表示不同类型的汽车，不要忘记加图例和x、y轴的label）。

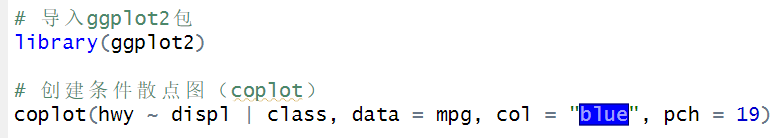
·普通散点图：

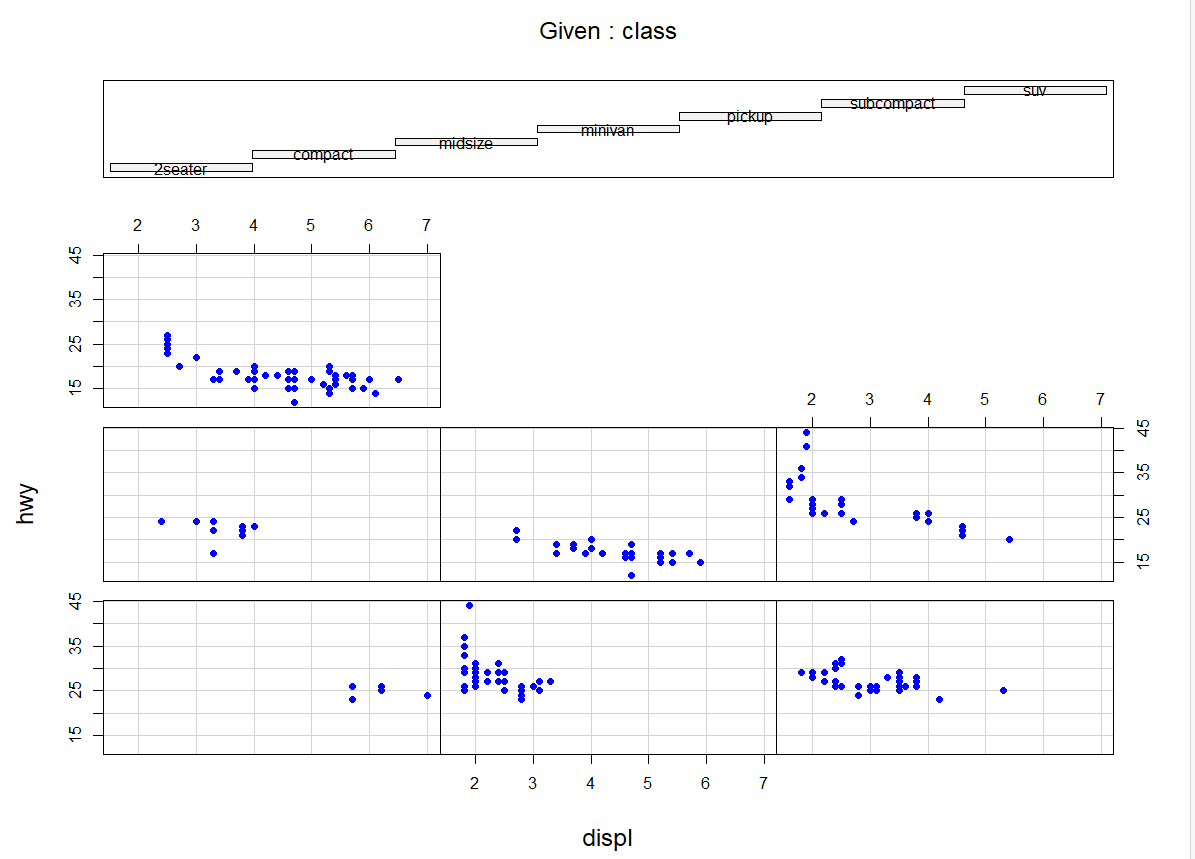




图中每个点代表一辆车，点的位置由其发动机排量和高速里程决定。不同的汽车类型（class）用不同形状的点来表示，使得观察者可以很容易地区分不同类型的汽车在这两个变量上的表现。图例解释了每种形状代表的汽车类别。

·条件散点图





这张图展示了不同类型汽车的发动机排量与高速公路里程之间的关系。每个汽车类型都有自己的子图，子图中的点表示个别车辆的数据。由于col参数被设置为"blue"，所以所有点都是蓝色的，形状为数字19指定的形状。观察者可以通过比较不同面板来分析不同汽车类型之间发动机排量与高速公路里程的关系是否存在显著差异。