北京理工大學

数据挖掘与 R 语言——

海藻数量频率预测及分析报告

学　　院：计算机学院

学　　号：2120151070

姓　　名：赵树阳

指 导 老 师：汤世平

完 成 日 期：2016.05.25

→关于 R 软件：

R 是一套完整的数据处理、计算和制图软件系统。其功能包括：数据存储和处理系统；数组运算工具（其向量、矩阵运算方面功能尤其强大）；完整连贯的统计分析工具；优秀的统计制图功能；简便而强大的编程语言：可操纵数据的输入与输出，可实现分支、循环、用户可自定义功能。R 在语义上是函数设计语言。它允许在“语言上计算”。这使得它可以吧表达式作为函数的输入参数，而这种做法对统计模型和绘图非常有用。R 是一个免费的自由软件，本问题中使用的是 3.0.1 版本。

→关于海藻问题的描述：

某些高浓度的有害藻类对河流生态环境的破坏是一个严重的问题。它们不仅破坏河流的生物，也破坏水质。能够监测并在早期对海藻的繁殖进行预测对提高河流质量是很有必要的。

针对这一问题的预测目标，在大约一年的时间内，在不同时间内收集了欧洲多条河流的水样。对于每个水样，测定了它们的不同化学性质以及 7 种有害藻类的存在频率。在水样收集过程中，也记录了一些其他特性，如收集的季节、河流大小和水流速度。

一．数据说明

本案例有两个数据集，第一个数据集有 200 个水样，这里我们命名为 algae，这个数据集的每条记录是同一条河流在该年的同一个季节的三个月内收集的水样的平均值。

每条记录由 11 个变量构成：
3 个是标称变量：

1>水样收集的季节 (春 夏 秋 冬)
2>收集样本的河流大小 (大 中 小)
3>河水速度 (高 中 低)

以及水样的化学参数的 8 个变量：

1>最大 pH 值(mxPH)
2>最小含氧量(mnO2)
3>平均氯化物含量(Cl)
4>平均硝酸盐含量(NO3)

5>平均氨含量(NH4)

6>平均正磷酸盐含量(oPO4)

7>平均磷酸盐含量(PO4)

8>平均叶绿素含量(Chla)

与这些参数相关的是 7 种不同有害藻类在相应水样中的频率数目：a1-a7。并未提供所观察藻类的名称的相关信息。

第二个数据集是 140 个不含 7 种藻类的频率数目的测试集。名为 algae.sols，本案例的主要目标是预测 140 个水样中 7 种海藻的频率。

即：两个数据集：训练样本（200 个）+测试样本（140 个）

3 个标称变量+8 个水样化学参数+7 种有害藻类的频率

二．数据加载到 R

利用 R 软件载入 DMwR 添加包，里面有我们需要的名为 algae 的数据框。这个数据框里含有前面提到的 200 个观测值：

```
> library(DMwR)
> head(algae)
```

	season	size	speed	mxPH	mnO2	Cl	NO3	NH4	oPO4	PO4	Chla	a1	a2	a3	a4	a5	a6	a7
1	winter	small	medium	8.00	9.8	60.800	6.238	578.000	105.000	170.000	50.0	0.0	0.0	0.0	0.0	34.2	8.3	0.0
2	spring	small	medium	8.35	8.0	57.750	1.288	370.000	428.750	558.750	1.3	1.4	7.6	4.8	1.9	6.7	0.0	2.1
3	autumn	small	medium	8.10	11.4	40.020	5.330	346.667	125.667	187.057	15.6	3.3	53.6	1.9	0.0	0.0	0.0	9.7
4	spring	small	medium	8.07	4.8	77.364	2.302	98.182	61.182	138.700	1.4	3.1	41.0	18.9	0.0	1.4	0.0	1.4
5	autumn	small	medium	8.06	9.0	55.350	10.416	233.700	58.222	97.580	10.5	9.2	2.9	7.5	0.0	7.5	4.1	1.0
6	winter	small	high	8.25	13.1	65.750	9.248	430.000	18.250	56.667	28.4	15.1	14.6	1.4	0.0	22.5	12.6	2.9

函数 head()将显示数据框的前 6 行。数据框的每一行代表一个观测值。

三．数据可视化和摘要

鉴于开始我们对该领域一无所知，首先我们要了解一些数据的统计特性，为后面的数据处理与建模提供更多的信息。

获取数据统计特性的一个方法是获取数据的如下描述性统计摘要。命令为：

```
>summary(algae)
```

```
   season       size       speed          mxPH
autumn:40   large :45   high  :84   Min.   :5.600
spring:53   medium:84   low   :33   1st Qu.:7.700
summer:45   small :71   medium:83   Median :8.060
winter:62                           Mean   :8.012
                                    3rd Qu.:8.400
                                    Max.   :9.700
                                    NA's   :1
     mnO2              Cl               NO3
Min.   : 1.500   Min.   :  0.222   Min.   : 0.050
1st Qu.: 7.725   1st Qu.: 10.981   1st Qu.: 1.296
Median : 9.800   Median : 32.730   Median : 2.675
Mean   : 9.118   Mean   : 43.636   Mean   : 3.282
3rd Qu.:10.800   3rd Qu.: 57.824   3rd Qu.: 4.446
Max.   :13.400   Max.   :391.500   Max.   :45.650
NA's   :2        NA's   :10        NA's   :2
     NH4               oPO4             PO4
Min.   :    5.00   Min.   :  1.00   Min.   :  1.00
1st Qu.:   38.33   1st Qu.: 15.70   1st Qu.: 41.38
Median :  103.17   Median : 40.15   Median :103.29
Mean   :  501.30   Mean   : 73.59   Mean   :137.88
3rd Qu.:  226.95   3rd Qu.: 99.33   3rd Qu.:213.75
Max.   :24064.00   Max.   :564.60   Max.   :771.60
NA's   :2          NA's   :2        NA's   :2
     Chla              a1               a2
Min.   :  0.200   Min.   : 0.00   Min.   : 0.000
1st Qu.:  2.000   1st Qu.: 1.50   1st Qu.: 0.000
Median :  5.475   Median : 6.95   Median : 3.000
Mean   : 13.971   Mean   :16.92   Mean   : 7.458
3rd Qu.: 18.308   3rd Qu.:24.80   3rd Qu.:11.375
Max.   :110.456   Max.   :89.80   Max.   :72.600
NA's   :12
      a3               a4               a5
Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
Median : 1.550   Median : 0.000   Median : 1.900
Mean   : 4.309   Mean   : 1.992   Mean   : 5.064
3rd Qu.: 4.925   3rd Qu.: 2.400   3rd Qu.: 7.500
Max.   :42.800   Max.   :44.600   Max.   :44.400
      a6               a7
Min.   : 0.000   Min.   : 0.000
1st Qu.: 0.000   1st Qu.: 0.000
Median : 0.000   Median : 1.000
Mean   : 5.964   Mean   : 2.495
3rd Qu.: 6.925   3rd Qu.: 2.400
Max.   :77.600   Max.   :31.600
```

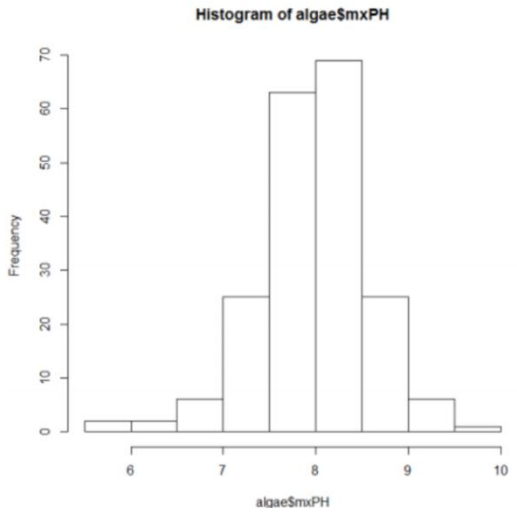
这个简单函数立即给出了数据的的统计特征概括。对于名义变量，他给出了每个取值的变量的频数。例如，从结果中可知冬季采集的水样比其他季节更多，河流 size 为最大的有 45 个，河流的流速 low 比较少。对于数值型变量，R 为我们提供了均值，中位数，四分位数，极值等一系列统计信息。这些统计信息提供了变量值分布的初步信息，在变量有缺失值的情况下，字符串 NA’s 后面的数值即为缺失值的个数，通过观察这些值，我们可以了解数据分布的偏度和分散情况。这些信息大多数都可以通过图形来表达出来。

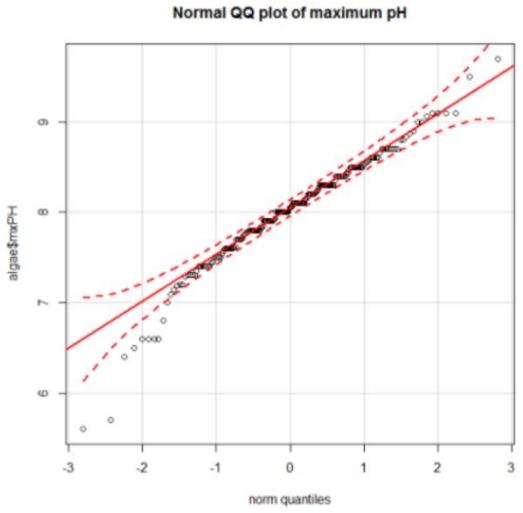
1. 利用如下命令绘制得到变量 mxPH 等各个属性的直方图：

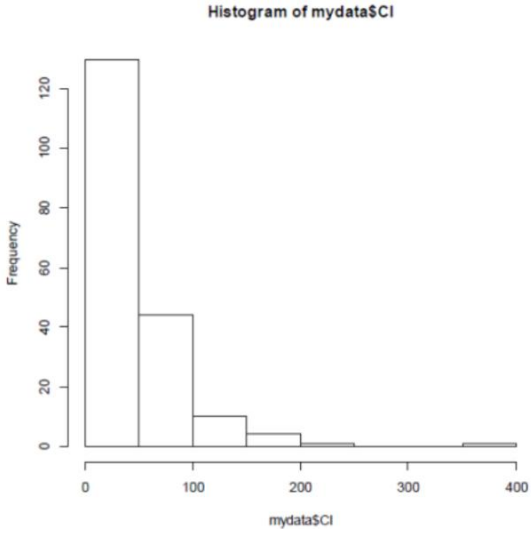
```
> hist(algae$mxPH,probability = T)
> hist(alage$mxPH)
```

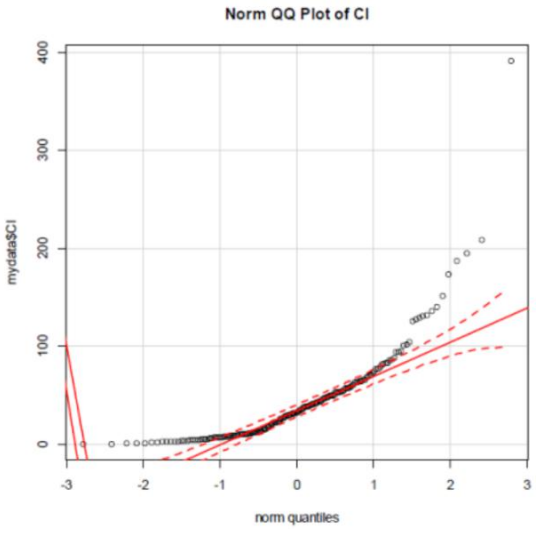
2. 利用如下命令绘制得到变量 mxPH 等各个属性的 Q-Q 图：

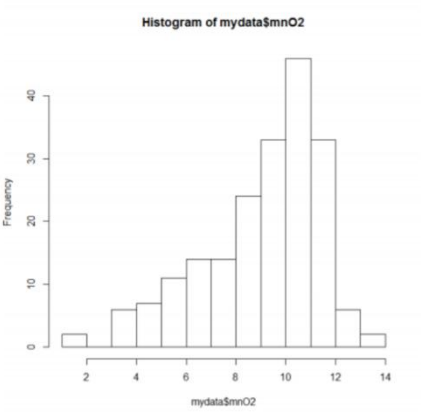
```
> library(car)
> qq.plot(algae$mxPH,main="Normal QQ plot of maximum pH")
```

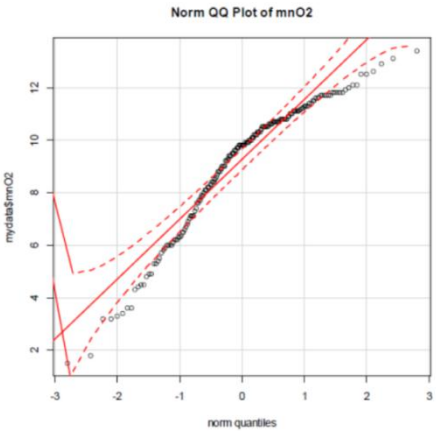


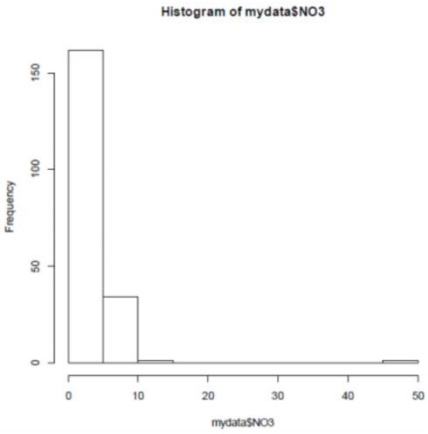


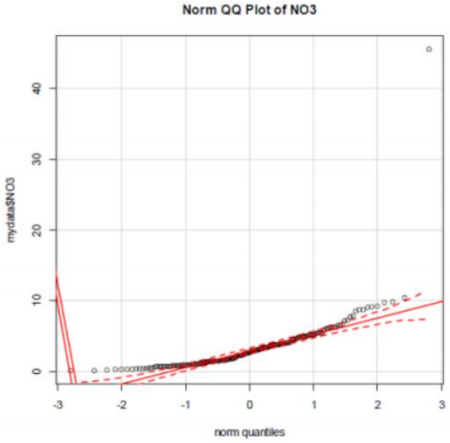


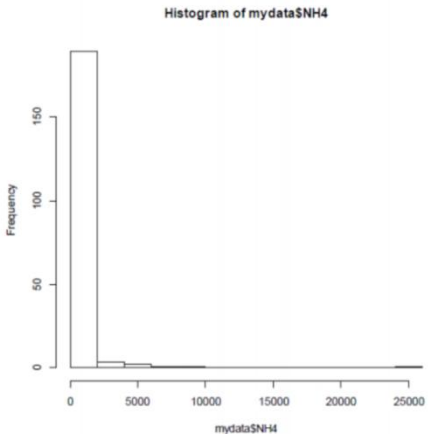


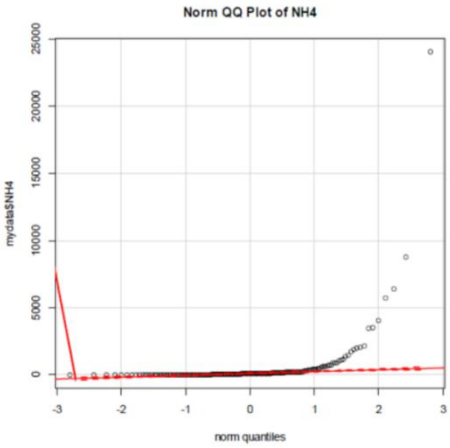

Histogram of mydata$mnO2
Frequency
mydata$mnO2

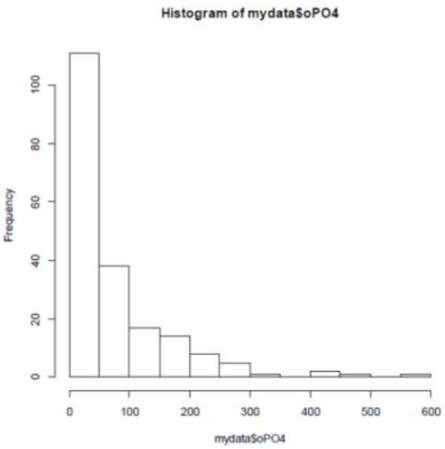

Norm QQ Plot of mnO2
mydata$mnO2
norm quantiles

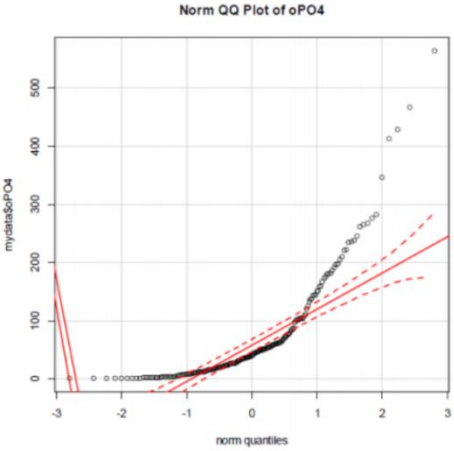

Histogram of mydata$NO3
Frequency
mydata$NO3

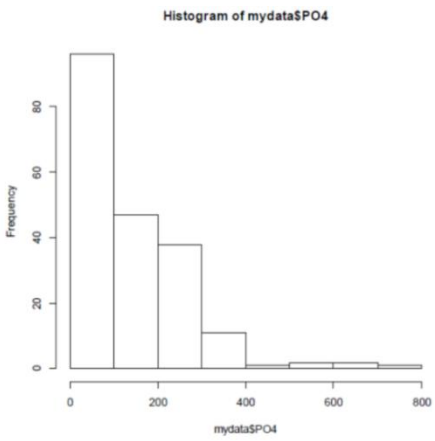

Norm QQ Plot of NO3
mydata$NO3
norm quantiles

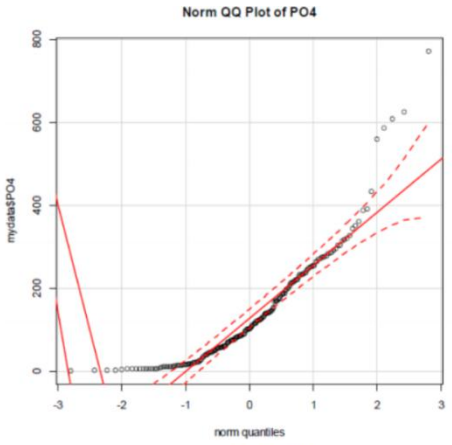

Histogram of mydata$NH4
Frequency
mydata$NH4

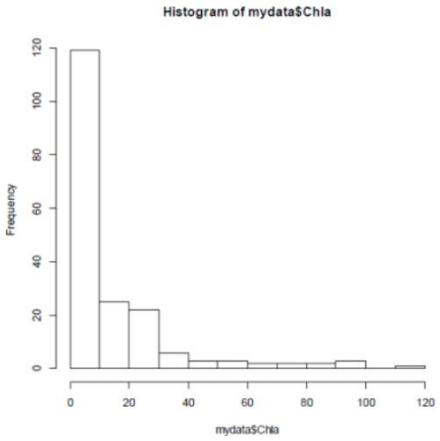

Norm QQ Plot of NH4
mydata$NH4
norm quantiles

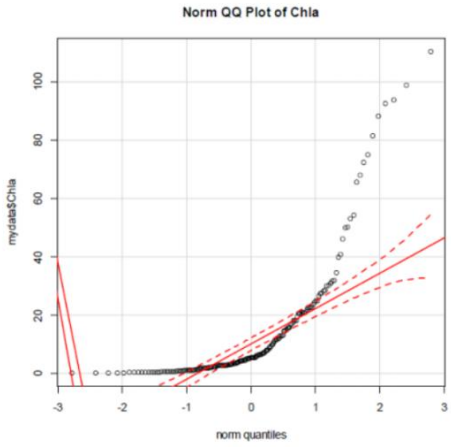

Histogram of mydata$oPO4
Frequency
mydata$oPO4


Norm QQ Plot of oPO4
mydata$oPO4
norm quantiles





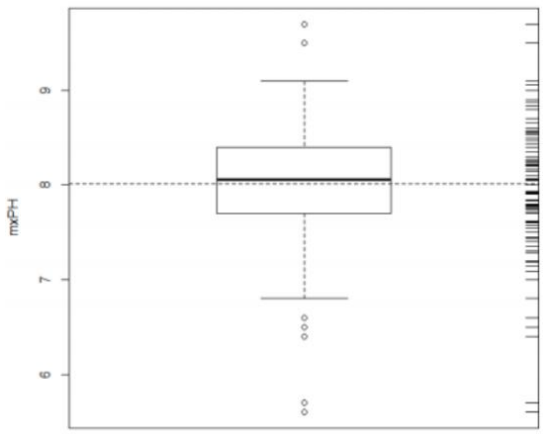


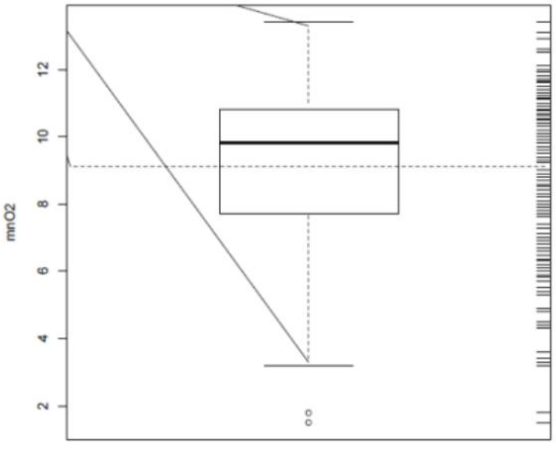


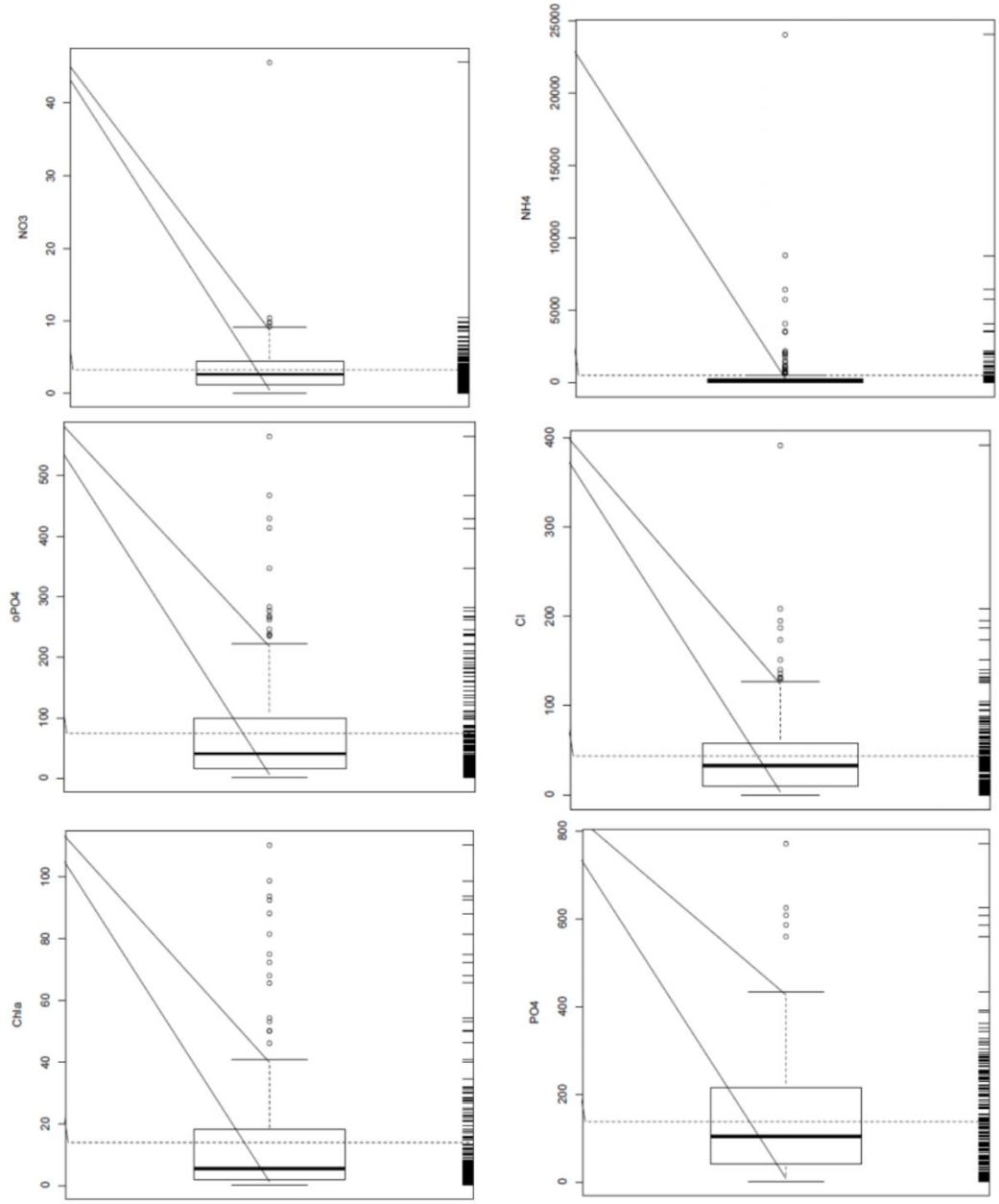
右图是用函数 qq.plot()得到的 Q-Q 图，它绘制变量值和正态分布的理论分位数（黑色实线）的散点图。同时，它给出正态分布的 95%置信区间的带状图（虚线）。从右图可知，变量有几个小的值明显在 95%置信区间之外，它们不服从正态分布。

3. 利用如下命令绘制得到变量 oPO4 等各个数值属性的盒图：

```
> boxplot(algae$oPO4,ylab="Orthophosphate")
> rug(jitter(algae$oPO4),side = 2)
> abline(h=mean(algae$oPO4,na.rm = T),lty=2)
```







ylab 为设置 y 轴标题；

rug 函数绘制变量的实际值，side=4 表示绘制在图的右侧（1 下 2 左 3 上）；

abline 函数绘制水平线，mean 表均值，na.rm=T 指计算时不考虑 NA 值，lty=2 设直线型为虚线。

盒图上方小横线上面的小圆圈表示与其他值比较特别大的值，通常认为是离群值，这意味着箱线图给出大量的信息，它不仅给出变量的中心趋势，也给出了变量的发散情况和离群值。 上图中与 X 轴平行的直线，是变量的均值所在的位置，将均值线与中位数线进行比较，就可以知道变量的多个离群值使得作为变量中心的均值产生了扭曲。

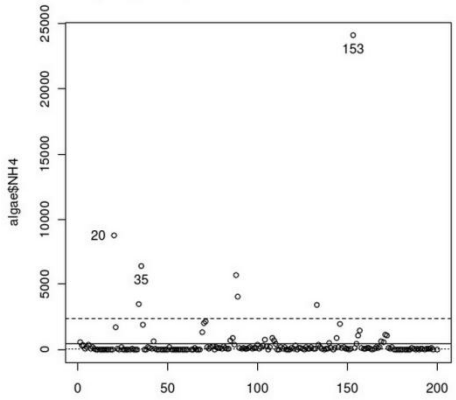
总的来说，较小的离群点多于较大的离群点导致了平均值较中位线稍小，离群点总体较少。

4. 利用如下命令进行离群点标示

当有离群点时，需要确定哪些有离群值的观测，可以使用图形法。如绘制 NH4 的值，将会有一个特别大的值，使用以下方法识别特大值相应的水样：

```
> plot(algae$NH4,xlab = "")
> abline(h=mean(algae$NH4,na.rm = T),lty=1)
> abline(h=mean(algae$NH4,na.rm = T)+sd(algae$NH4,na.rm = T),lty=2)
> abline(h=median(algae$NH4,na.rm = T),lty=3) > identify(algae$NH4)
```

上面函数第一条绘制变量的所有值，调用函数 abline 绘制三条有用的直线，第一条为均值，第二条为均值加一个标准差，第三条为中位数，对于离群值的识别尽管这三条不是必须的，但它们能提供有用的信息。



从上图中知道，水样 153 的 NH4 值是一个极大异常值，20，34,35,88,89,153 等水样中 NH4 值都异常高，可以把这些水样选择出来，看看他们与别的水样中海藻有什么种类有什么差异。

5. 利用如下命令进行 7 种藻类与河流大小的条件盒图绘制

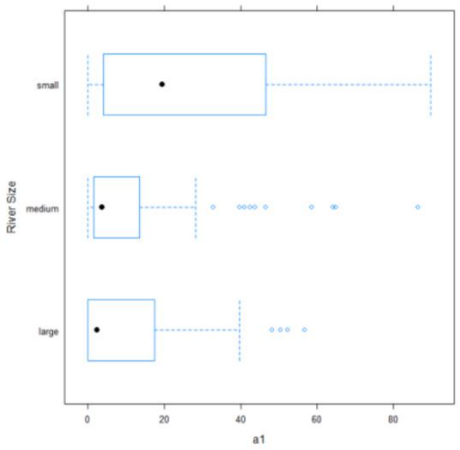
假设需要研究海藻变量 a1 的值得分布，然而，这里需要分布如何依赖于其他变量，就需要新的变量，新的工具。

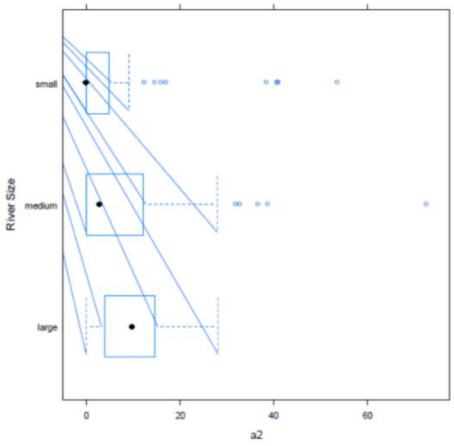
条件绘图是依赖于某个特定因子的图形表示，因子是一个为一个取值为有限集合的名义变量。例如，对于 size 的不同取值，可以绘制变量 a1 的一组盒图。每个盒图是对应于变量 size 的某个特定值的水样子集。通过这些盒线图可以研究名义变量 size 如何影响变量 a1 值得分布。

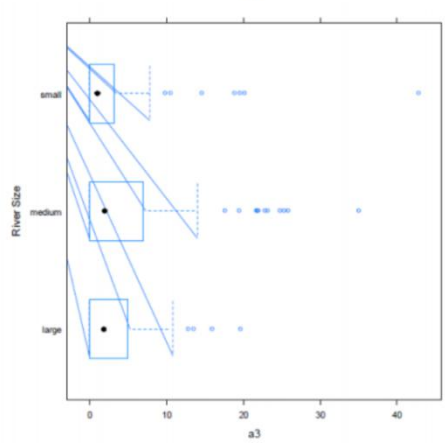
```
> library(lattice)
> bwplot(size~a1,data=algae,ylab = "River Size",xlab = "Algal A1")
```

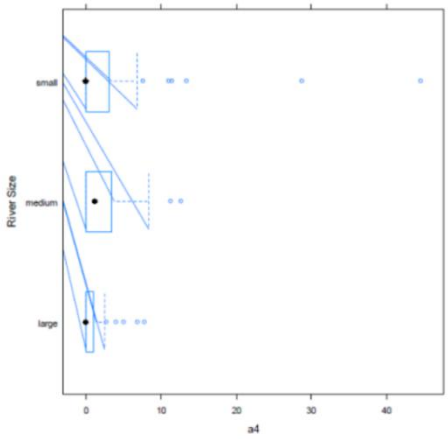
上面的第一条指令载入 lattice 包。第二条指令绘制这些图 lattice 版本的箱图，这条指令可以读做：对变量 size 的每个值绘制 a1。其他参数的意义显而易见。

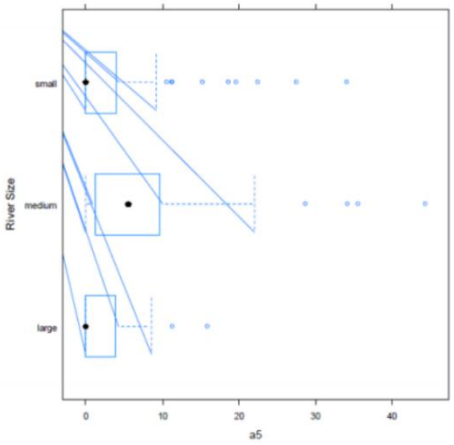
由 a1 的盒图可知，在规模较小的河流中，海藻 a1 的频率较高。

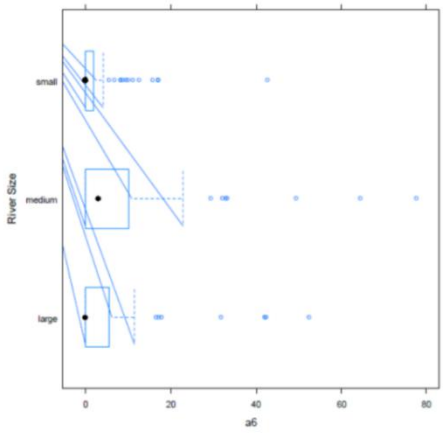

River Size
small
medium
large
0
20
40
60
80
a1

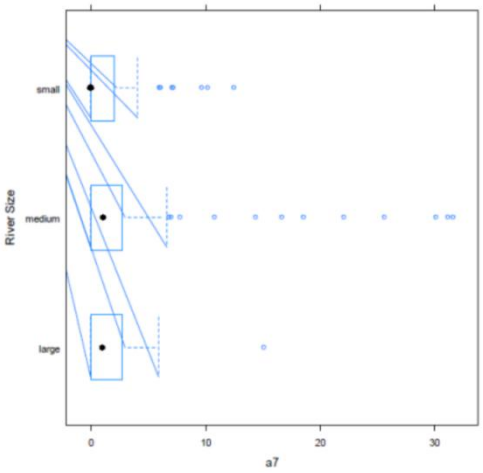

River Size
small
medium
large
0
20
40
60
a2


River Size
small
medium
large
0
10
20
30
40
a3


River Size
small
medium
large
0
10
20
30
40
a4


River Size
small
medium
large
0
10
20
30
40
a5

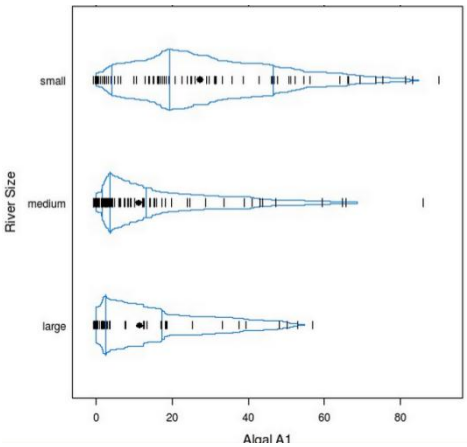

River Size
small
medium
large
0
20
40
60
80
a6


River Size
small
medium
large
0
10
20
30
a7

6. 利用如下命令进行条件分位盒图的绘制

这种盒图的另外一个类型是分位盒图，它可以给出绘制变量的更多信息，R 添加包 Hmisc 可以绘制分位盒图，下面绘制上面例子中 a1 变量的条件分位箱图：

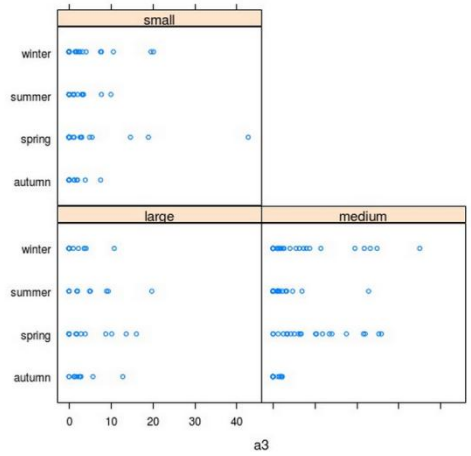
```
> library(Hmisc)
> bwplot(size~a1,data=algae,panel = panel.bpplot,probs=seq(.01,.49,by=.01),datade
nsity=T,ylab="River Size",xlab = "Algal A1")
```



上图中的点代表代表不同大小的河流中海藻频数的均值，而图中的竖线分别代表变量的第一份位数，中位数和第三分位数。图中的小竖线代表数据的真实取值，这些值分布信息由分位数图来体现。分位数箱图提供的信息要多于传统的箱图。例如我们可以确认上面的观测结论，小型河流有更高频率的海藻，但我们也观察到小型河流的海藻频率的分布比其他类型的河流的海藻频率的分布分散。

7. 利用如下命令绘制更复杂的条件盒图与连续因子离散化

```
> stripplot(season ~ a3|size,data-algae[!Is.na(algae$size),])
```

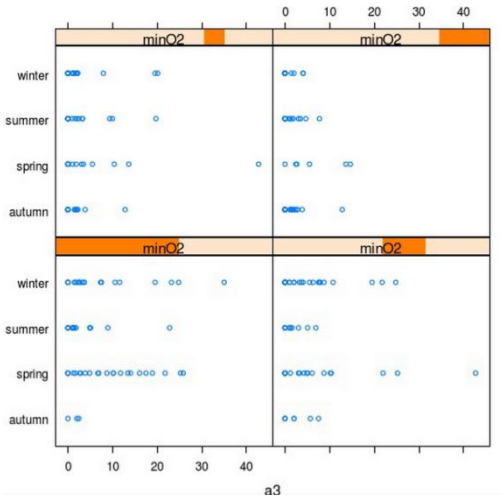


这种类型的条件绘图不局限于名义变量，也不局限于单个因子。只需把连续变量离散化，也同样可以进行条件绘图，下面给出两因子条件绘图的例子。考虑变量 a3 在给定变量 season 和变量 mnO2 下的条件绘图，变量 mnO2 是一个连

续变量，代码如下：

```
> m1<-equal.count(na.omit(algae$mnO2),number=4,overlap=1/5)
> stripplot(season~a3|m1,data = algae[!is.na(algae$mnO2),])
```

代码第一行是调用函数 equal.Count()对连续变量 mnO2 离散化，把该变量转化为因子类型。参数 number 设置需要的区间的个数。每个区间的观测值的个数相等。第二个调用函数 stripplot()，该函数是 lattice 包中的一个绘图函数，他根据另一个变量 season 把变量的实际值绘制到不同的图形中。然后对变量 mnO2 的每个不同的区间绘制不同的图形。这些区间按照从左到右，从上到下的顺序来排列。左下方对应的是较小的 mnO2 值。



四．数据缺失的处理

在许多水样中，一些变量含有缺失值。这种情形在实际问题中非常普遍，这会导致一些不能处理缺失值的分析方法无法应用。处理含有缺失值的数据是常用的几种策略：

将含有缺失值的部分剔除 ；
用最高频率值来填补缺失值；
根据属性之间的相关关系填补缺失值 ；
根据数据对象之间的相似性填补缺失值 。

1. 直接剔除缺失部分

剔除含有缺失数据的记录非常容易实现，尤其是当这些记录所占的比例在可用数据集中非常小的时候，这个选择就比较合理。

(1)先检查观测值，或者至少得到这些观测值的个数：

```
> algae[!complete.cases(algae),]
```

```
> nrow(algae[!complete.cases(algae),])
...
...
[1] 16
```
16 个含缺失的样本

(2)剔除所有含缺失值的记录：

```
> algae <- na.omit(algae)
```

(3)找到海藻数据集中每行数据的缺失值个数：

```
> apply(algae,1,function(x) sum(is.na(x)))
```

函数 apply（）是元函数，它可以在某些条件下对对象应用其他函数。第二个参数 “1” 标识第一个参数 algae 中的对象的第一个维度，即行数据。第三个参数是临时函数，功能是计算对象 x 中 NA 的数量。在 R 中逻辑值 TRUE 等于数值 1，逻辑值 FALSE 等于数值 0，这意味着当加一个布尔值向量时，得到向量中取值为 TRUE 的元素的个数。

(4)根据以上代码，可以编写一个程序找到 algae 中含有给定数目缺失值的行。然后应用如下 manyNA()函数找出缺失值个数大于列数 20%的行：

```
> data(algae)
> manyNAs(algae,0.2)
[1] 62 199
```

在第二个参数中可以设置一个精确的列数作为界限，这里默认值为 0.2。因此，用下面的代码就无须知道含有缺失值较多的行的具体数量。

```
> algae <- algae[-manyNAs(algae),]
```

以上操作可以把样本中 62，199 号水样数据剔除。

2. 用最高频率值填补缺失值

填补缺失数据最简单和快捷的方法是使用一些代表中心趋势的值，代表中心趋势的值反映了变量分布的最常见的值，因此最高频率值是最自然的选择。对于接近正态的分布来说，所有的观测值都较好地聚集在平均值周围，平均值数就是最佳选择。对偏态分布或者有离群值的分布而言，中位数是更好的代表数据中心趋势的指标。

```
> algae[48,'mxPH'] <- mean(algae$mxPH,na.rm=T)
```

这里，函数 mean（）计算数值向量的平均值。变量 Chla 的分布偏向于较低的数值，并且它有几个极端值，这些都使得平均值不能代表大多数的变量值。因此，我们使用中位数来填补这一类的缺失值：

```
> algae[is.na(algae$Chla),'Chla'] <- median(algae$Chla,na.rm=T)
```

对于名义变量，则采用众数。用以下命令完成填补所有缺失值：

```
> data(algae)
> algae<-alage[-mmahyNAs(algae),]
> algae<-centralImputation(algae)
```

由于缺失值的存在会导致某些方法不能使用，所以使用上面的方法填补缺失值通常也认为不是很好的方法。虽然上述的方法速度快，特别适用于大数据集，但是它可能导致较大的数据偏差，影响后期的数据分析工作。然而，使用无偏差方法来寻找最佳数据调补值复杂，对于大型数据挖掘问题可能并不适用。

3. 通过变量的相关关系来调补缺失值

另一种获取缺失值较少偏差估计值的方法是探寻变量之间的相关关系。先用以下命令来获得变量之间的相关矩阵：

```
> data(algae)
> symnum(cor(algae[,4:18],use="complete.obs"))
```

其中函数 cor（）的功能是产生变量之间的相关值矩阵。（忽略前 3 个名义变量）设定参数 use= “complete.obs” 时，R 在计算相关值时忽略含有 NA 的记录。函数 cor（）的输出结果并不是很清晰，用 symnum 来改善结果的输出形式。

```
     mP mO Cl NO NH o P Ch a1 a2 a3 a4 a5 a6 a7
mxPH 1
mnO2    1
Cl        1
NO3         1
NH4         , 1
oPO4    .  .       1
PO4     .  .       * 1
Chla .                 1
a1        .       . .   1
a2   .                .    1
a3                          1
a4      .          . .           1
a5                                  1
a6           .  .                    .  1
a7                                        1
attr(,"legend")
[1] 0 ‘ ’ 0.3 ‘.’ 0.6 ‘,’ 0.8 ‘+’ 0.9 ‘*’ 0.9 ‘B’ 1
```

结果显示，有两个相关性较大的值：NH4 和 NO3 之间，PO4 和 oPO4 之间。前者的相关性不是特别明显（0.6~0.8），考虑到样本 62 和样本 199 含有过多的缺失值，若剔除它们，样本中 NH4 和 NO3 就不存在缺失值了。后者相关值很高（大于 0.9），可用变量的相关性填补缺失值。

为了达到这个目标，我们需要找到 PO4 和 oPO4 这两个变量之间的线性关系，命令如下：

```
> data(algae)
> algae<-algae[-manyNAs(algae),]
> lm(PO4~oPO4,data = algae)
Call:
lm(formula = PO4 ~ oPO4, data = algae)
Coefficients:
(Intercept)        oPO4        42.90        1.29
```

可以得到线性模型： PO4=1.29oPO4+42.90 ；

在剔除样本 62 和样本 199 后，还剩下一个样本 28 在 PO4 上有缺失值，可以简单地使用上面的线性关系计算缺失值的填补值：

```
> algae[28,'PO4'] <- 42.897 + 1.293 * algae[28,'oPO4']
> algae[28,"PO4"]
[1] 48
```

通过变量之间的相关关系求出水样 28，PO4 的缺失值填补为 48。

4. 通过数据对象之间的相似性填补缺失值

假如两个水样是相似的，其中某些变量有缺失值，那么该缺失值很可能与另一个水样的值是相似的。有许多度量相似性的指标，最常用的是欧式距离。这个距离可以非正式的定义为任何两个案例的观察值之差的平方和，我们可以通过使用这种度量来寻找与任何含有缺失值的案例最相似的 10 个水样，并用它们填补缺失值。

我们考虑两种应用这些值的方法。第一种方法简单地计算这 10 个最相近的案例的中位数并用这个中位数来填充这些缺失值。第二种方法是采用这些相似数据的加权均值。这里用高斯核函数从距离获得权重。

$$w(d)=e^{-d} \qquad d(x,y)=\sqrt{\sum_{i=1}^{p}\delta_i(x_i,y_i)}$$

$$\delta_i(x_i, y_i) = \begin{cases} 1 & \text{名义变量且 } x_i \neq y_i \\ 0 & \text{名义变量且 } x_i = y_i \\ (x_i - y_i)^2 & \text{数值变量} \end{cases}$$

命令如下：

```
> algae<-knnImputation(algae,k=10,meth = "median")
```

总之使用这些简单地操作，数据中不再含有缺失值 NA，未使用 R 的其他函数进行分析做好充分的准备。

当决定使用哪种方法来填充缺失值时，大多数时候根据你所分析的领域知识来确定，根据个案之间的相似性来填补缺失看起来合理，但这种方法也存在其他问题，例如可能存在不相关的变量扭曲相似性，甚至造成大型数据集的计算特别复杂的问题。另外，对于这些大数据集问题，可以通过随机抽取样本的方法来计算他们之间的相关性。

五．海藻问题分总结

海藻问题是数据挖掘所处理的诸多问题中的一类重要问题。在这个问题中，主要任务是建立预测模型，并预测在给定预测变量的取值时相应的目标变量的值。预测模型也可能会说明哪一个预测变量对目标变量有较大的影响，即模型可能提供影响目标变量因素的一个综合描述。

通过此次分析报告，我深入了解并学习了数据可视化分析的具体过程，也熟悉了 R 软件，获益匪浅。