马的疝病分析

**1. 问题描述**

疝病是描述马胃肠痛的术语，这种病不一定源自马的胃肠问题，其他问题也可能引发马疝病。所给数据集是医院检测的一些指标。

**2. 实验环境**

操作系统：Windows 10

平台：R(×64) 3.3.3

编程语言：R语言

**3. 数据说明**

数据来自UCI机器学习数据库。共368个样本，28个特征，详细说明如表1和表2所示。

**表1.样本数量说明**

|  |  |  |
| --- | --- | --- |
| 训练样本 | 测试样本 | 合计 |
| 300 | 68 | 368 |

**表2.特征说明**

| 特征名称 | 详细说明 |
| --- | --- |
| 手术（surgery） | 1=做过手术，2=没做过手术 |
| 年龄（Age） | 1=成年马，2=幼马（＜6个月） |
| 医院登记号（Hospital Number ） | 分配给马的病例号 |
| 直肠温度（rectal temperature） | 正常温度37.8 |
| 脉搏（pulse） | 反应心脏状况：成年马正常值30-40 |
| 呼吸率（ respiratory rate） | 正常值：8-10 |
| 四肢温度（temperature of extremities） | 1=Normal, 2=Warm, 3=Cool, 4=Cold |
| 外周脉搏（peripheral pulse） | 1=normal, 2=increased, 3=reduced, 4=absent |
| 黏膜（mucous membranes） | 1=normal pink, 2=bright pink, 3=pale pink, 4=pale cyanotic, 5=bright red, 6=dark cyanotic |
| 毛细血管再充盈时间 （capillary refill time ） | 1=小于3秒, 2=大于等于3秒 |
| 疼痛程度（pain） | 1=no pain, 2=depressed, 3=intermittent mild pain, 4=intermittent severe pain, 5=continuous severe pain |
| 肠蠕动（ peristalsis） | 1=hypermotile, 2=normal, 3=hypomotile,  4=absent |
| 腹胀（abdominal distension） | 1=none, 2=slight, 3=moderate, 4=severe |
| 胃管（nasogastric tube ） | 1=none, 2=slight, 3=significant |
| 胃反流（nasogastric reflux） | 1=none, 2=>1 liter, 3=<1 liter |
| 胃反流PH值(nasogastric reflux PH) | 范围：0-14 正常值：3-4 |
| 直肠检查(rectal examination) | 1=normal, 2=increased, 3=decreased, 4=absent |
| 腹部（abdomen） | 1=normal, 2=other, 3=firm feces in the large intestine, 4=distended small intestine, 5=distended large intestine |
| 红细胞体积（packed cell volume） | 线性值。正常值：30-50 |
| 总蛋白（total protein） | 正常值： 6-7.5 (gms/dL) |
| 腹腔穿刺术外表 （abdominocentesis appearance） | 1=clear, 2=cloudy, 3=serosanguinous |
| abdomcentesis总蛋白  (abdomcentesis total protein) | 线性值。单位：gms/dL |
| 结局（outcome） | 1=lived, 2=died, 3=was euthanized |
| 手术损伤（surgical lesion） | 1=Yes, 2=No |
| 损伤类型（type of lesion\_1） | 第一个数：损伤位置；第二个数：类型；第三个数：子类型；第四个数：特定码 |
| 损伤类型（type of lesion\_2） | 同上 |
| 损伤类型（type of lesion\_3） | 同上 |
| cp\_data | 1=Yes, 2=No |

由表2可以看出，直肠温度、脉搏、呼吸率、胃反流PH值、红细胞体积、总蛋白、abdomcentesis总蛋白共7个特征为数值属性，其余21个特征为标称属性。

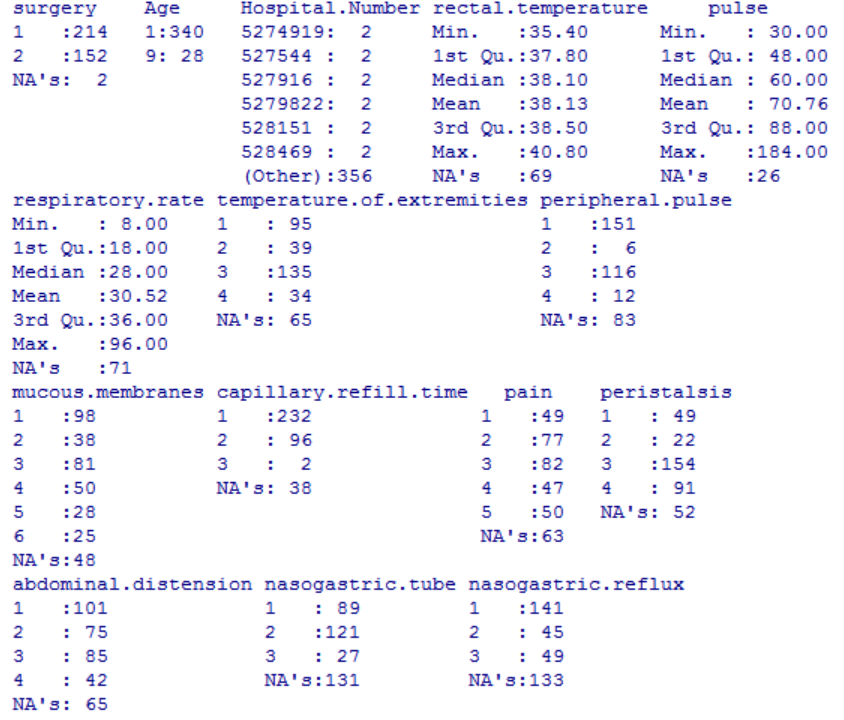
**4. 数据分析**

**4.1 数据可视化和摘要**

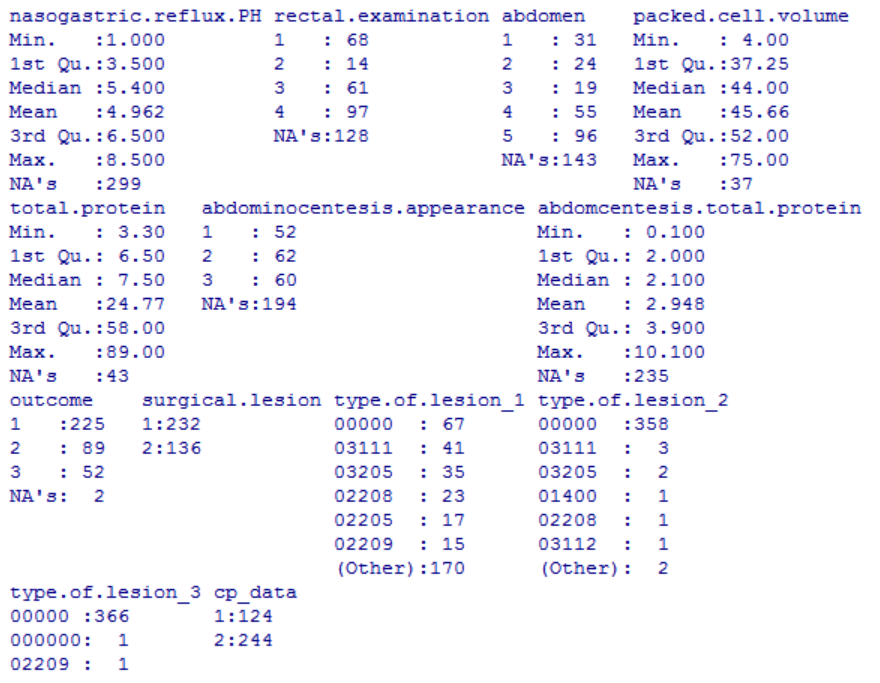
**数据摘要**

1. 对标称属性，给出每个可能取值的频数，
2. 数值属性，给出最大、最小、均值、中位数、四分位数及缺失值的个数。

数据摘要如图1，图2所示：

****

**图1.前15个特征的数据摘要**

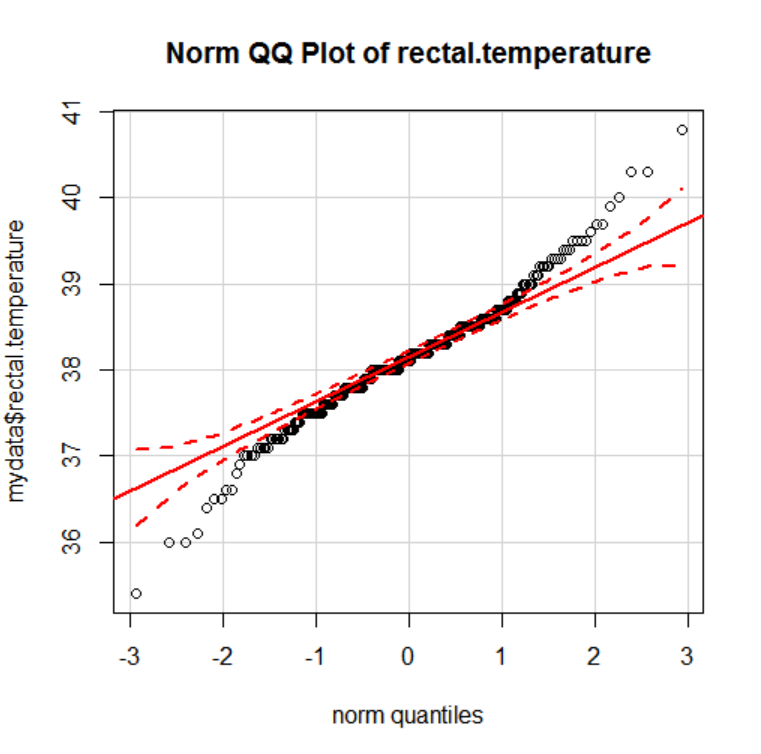
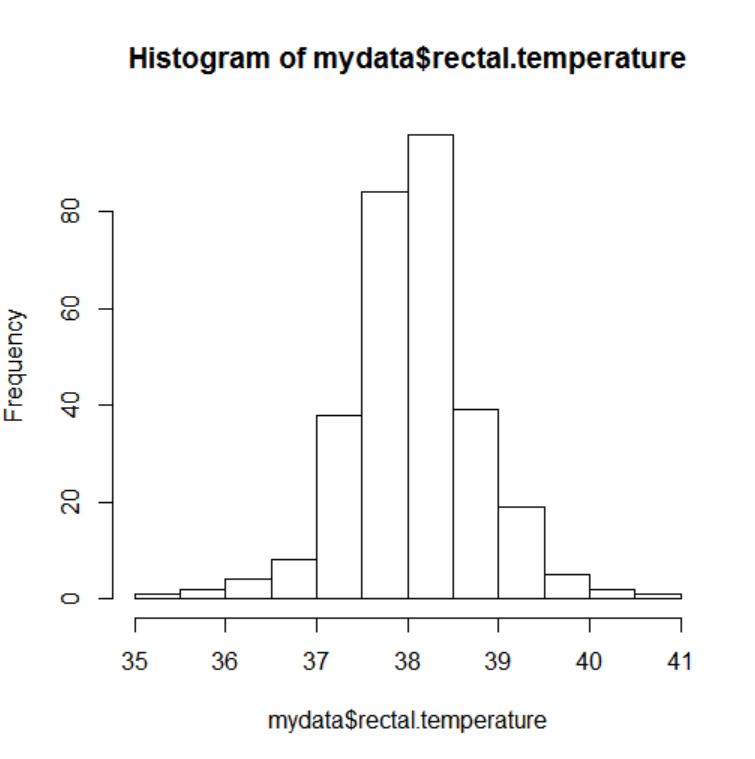
****

**图2.后13个特征的数据摘要**

**数据的可视化**

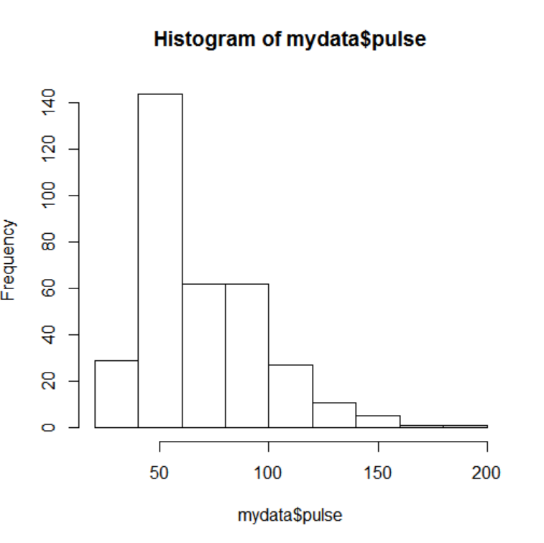
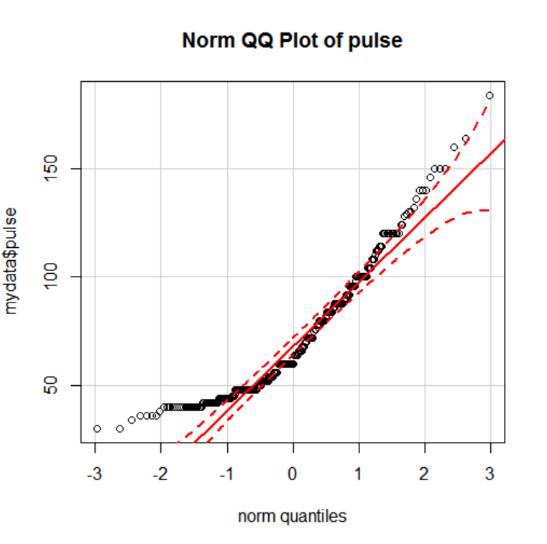
针对数值属性，

1. 绘制直方图，如rectal temperature，用qq图检验其分布是否为正态分布。
2. 直肠温度（rectal temperature）



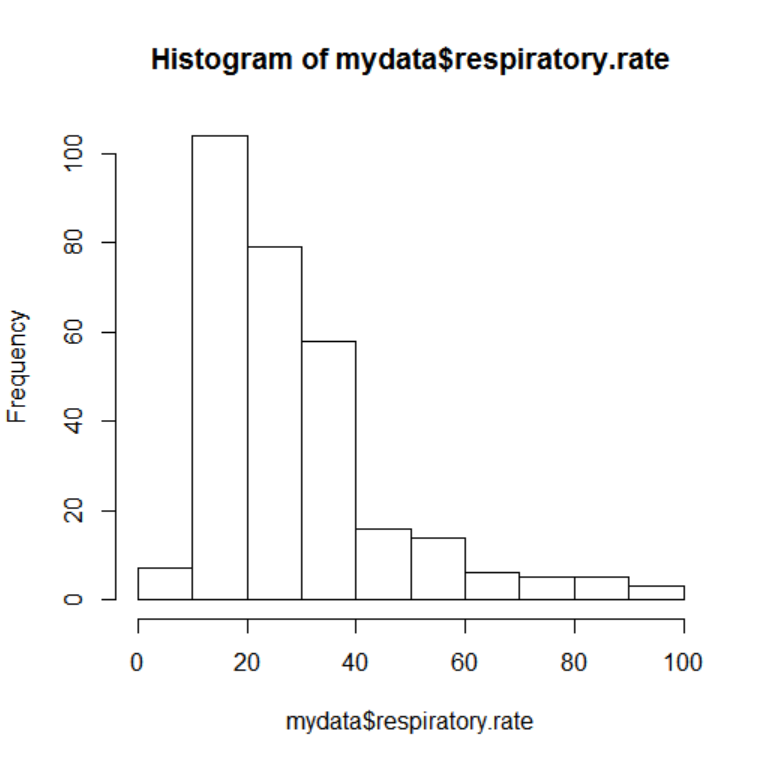
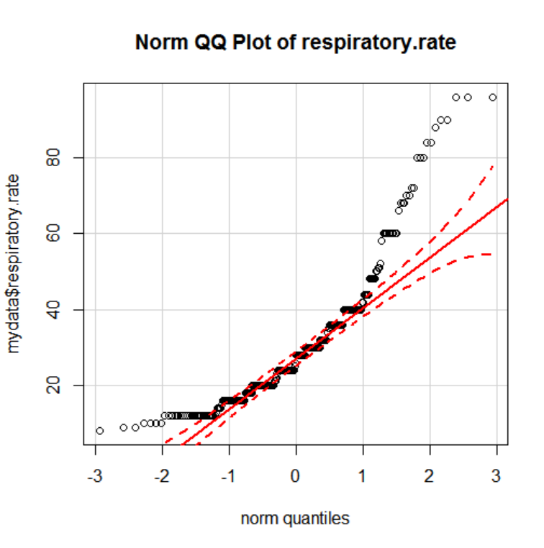
**图3. rectal temperature的直方图和qq图**

1. 脉搏（pulse）

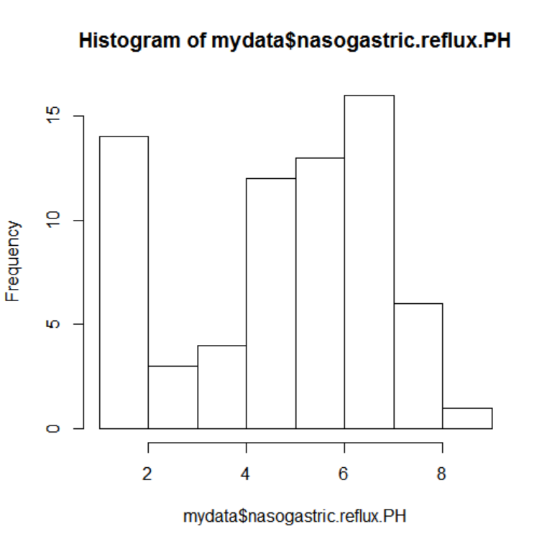
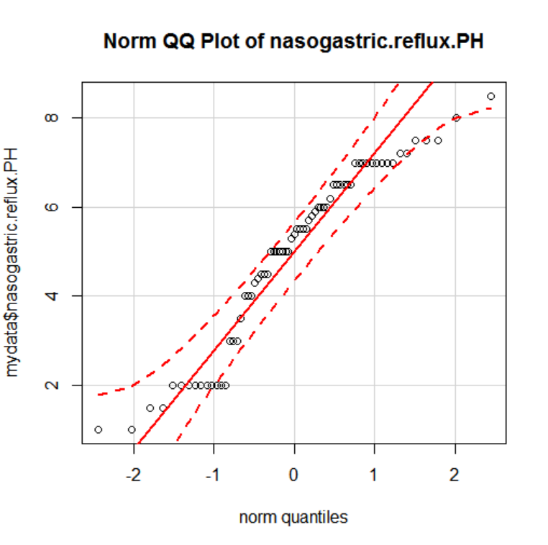
**图4. pulse的直方图和qq图**

1. 呼吸率（ respiratory rate）

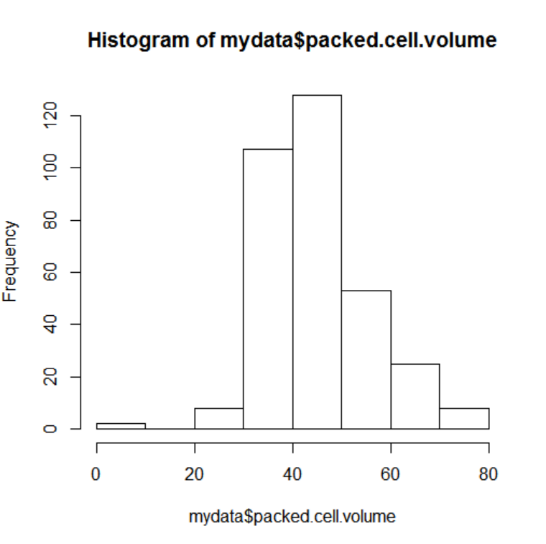
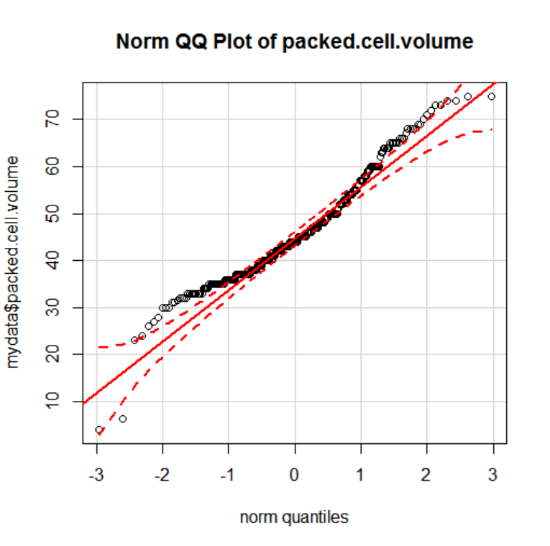
**图5.** **respiratory rate的直方图和qq图**

1. 胃反流PH值(nasogastric reflux PH)

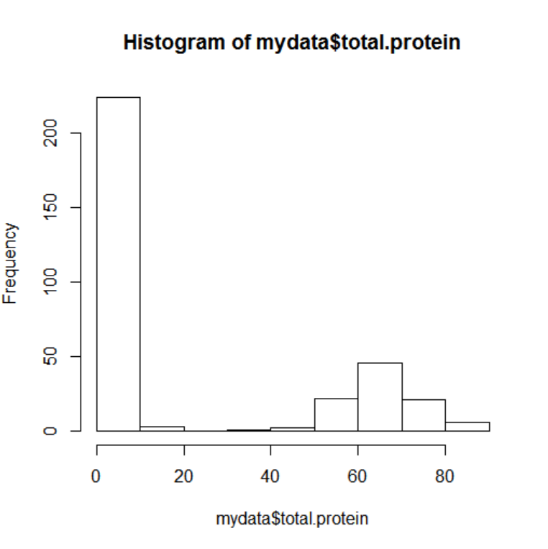
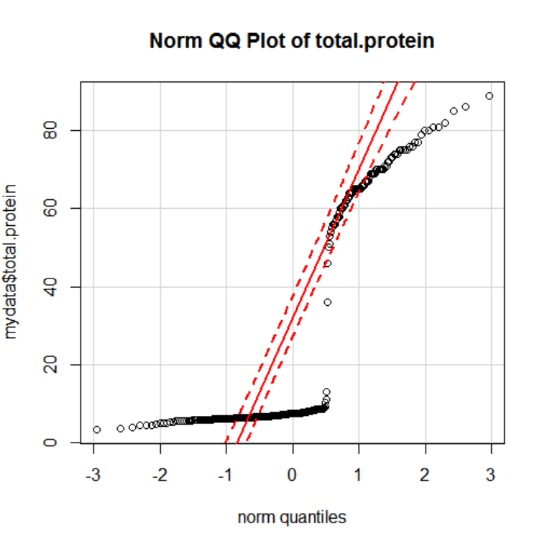
**图6.** **nasogastric reflux PH的直方图和qq图**

1. 红细胞体积（packed cell volume）

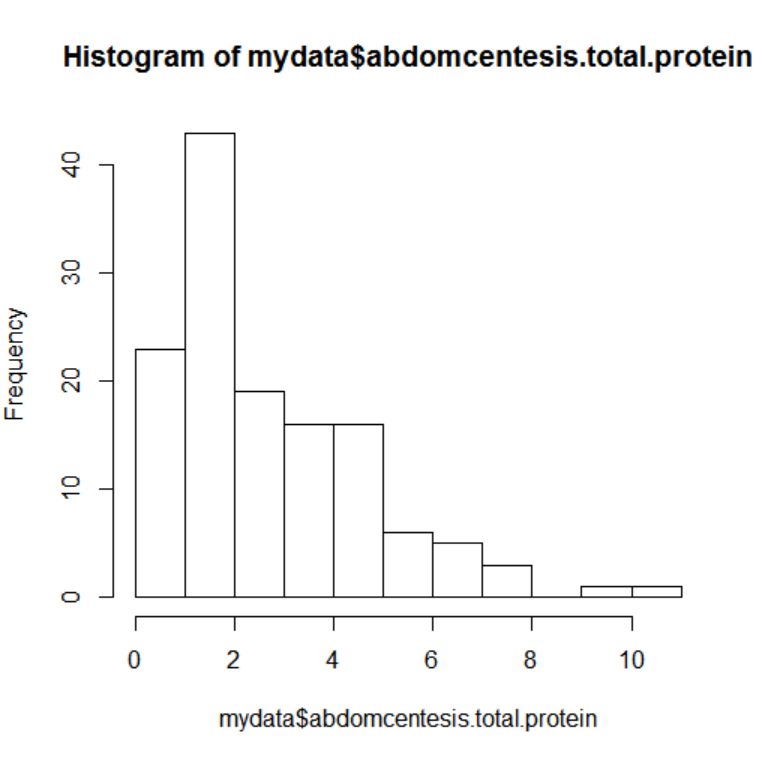
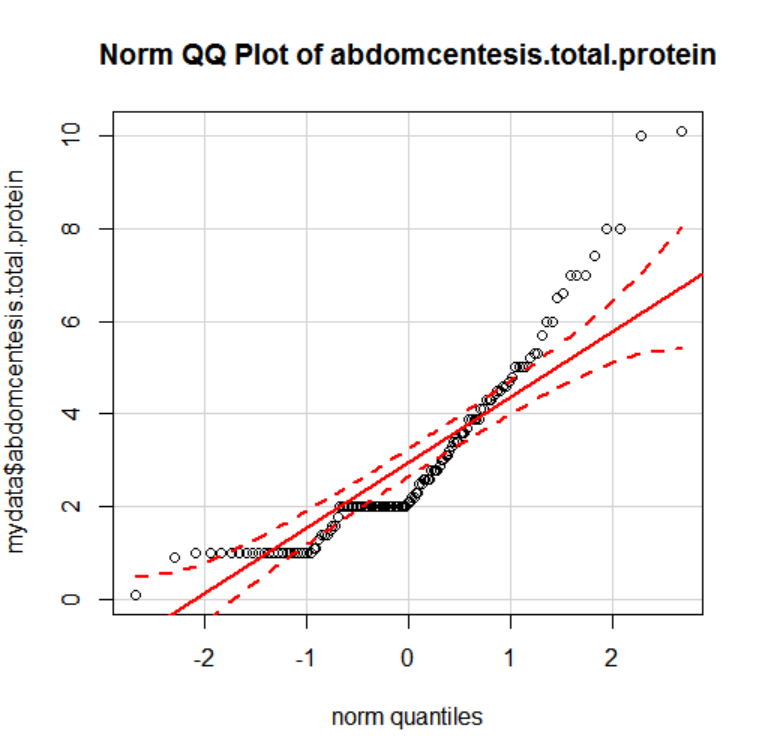
**图7.** **packed cell volume的直方图和qq图**

1. 总蛋白（total protein）

**图8.** **total protein的直方图和qq图**

1. abdomcentesis总蛋白(abdomcentesis total protein)

**图9.** **abdomcentesis total protein的直方图和qq图**

绘制出的直方图纵轴是其频数，横轴是其分布区间。qq图中，红色实线为其qq线，虚线为95%置信度的置信区间。

分析数据图表可以看出，属性均有明显数据点离开了95%的置信区间，所以不满足正态分布。

1. 绘制盒图，对离群值进行识别

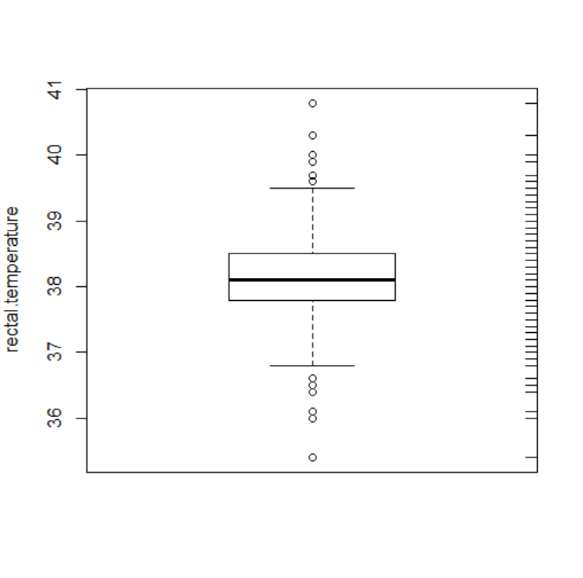
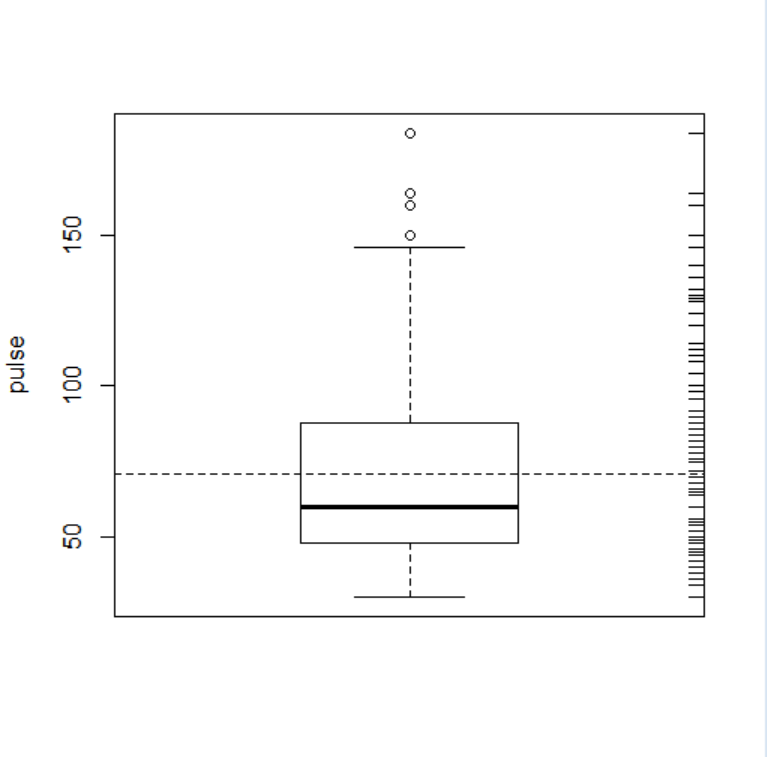
以直肠温度rectal.temperature为例，绘制其盒图的命令如下：

boxplot(mydata$rectal.temperature,ylab='rectal.temperature')

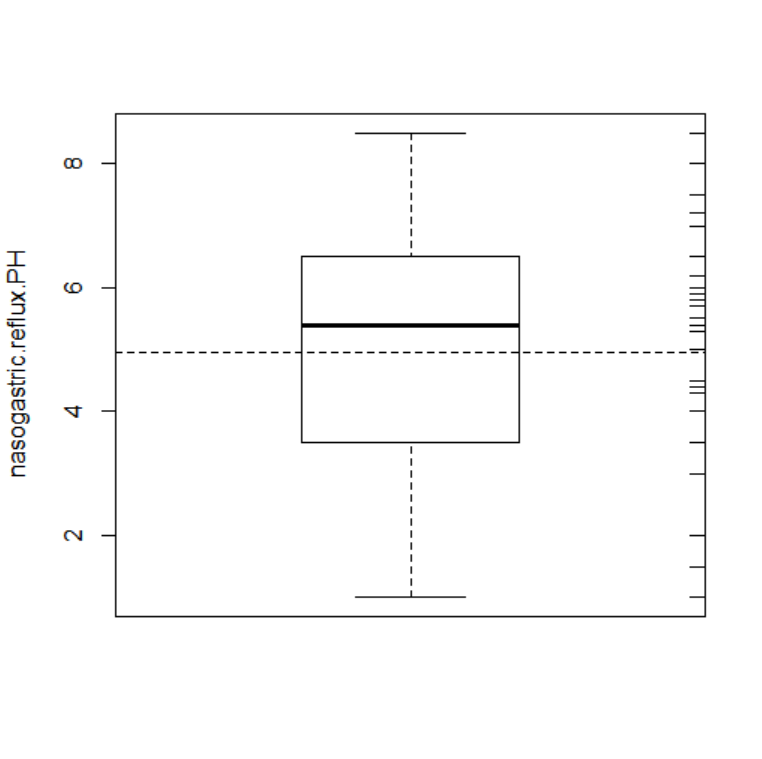
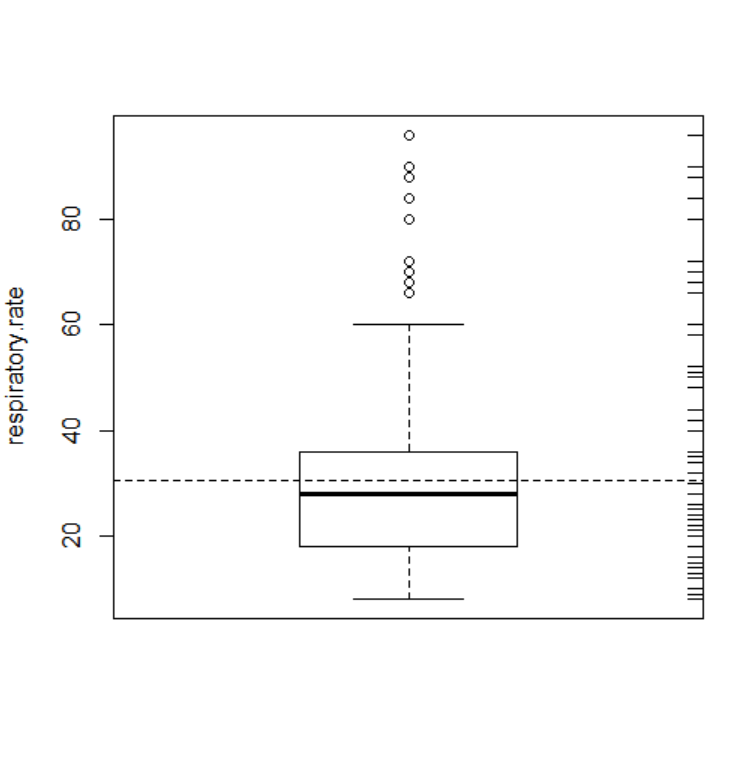
rug(mydata$rectal.temperature,side=4)

abline(h=mean(mydata$rectal.temperature,na.rm=T),lty=2)

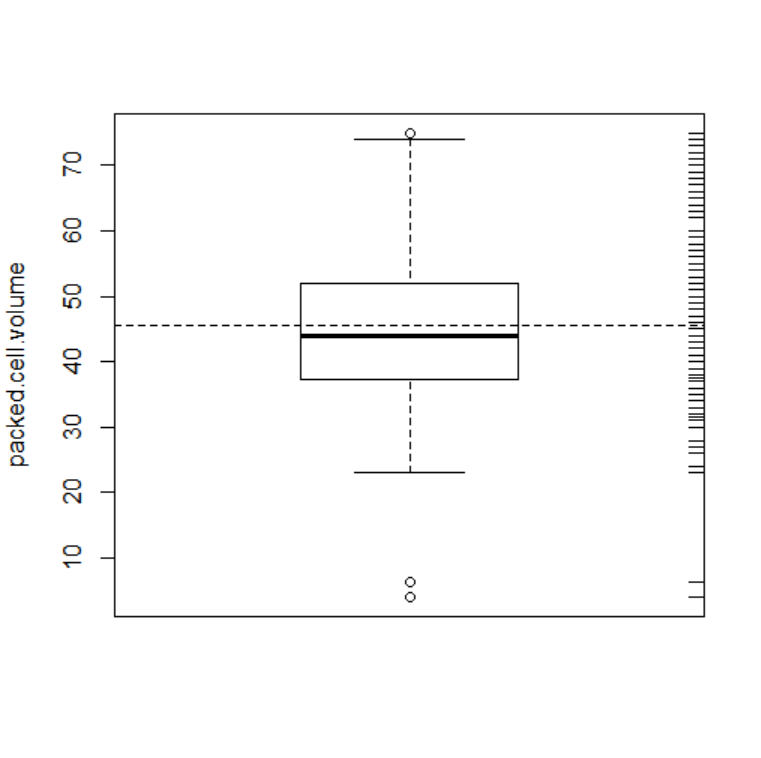
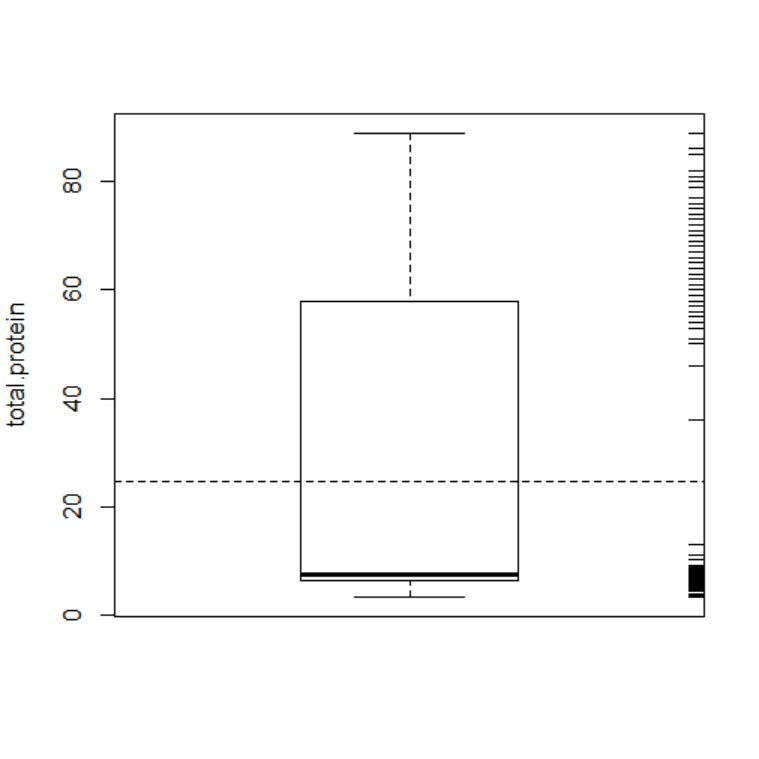
Rug函数绘制每个点在纵轴上的投影情况，abline绘制数据的均值，在图中以虚线的方式呈现。由各属性的盒图，可以分析出离群点的数量以及分布情况。每个属性的盒图如图10-16所示。

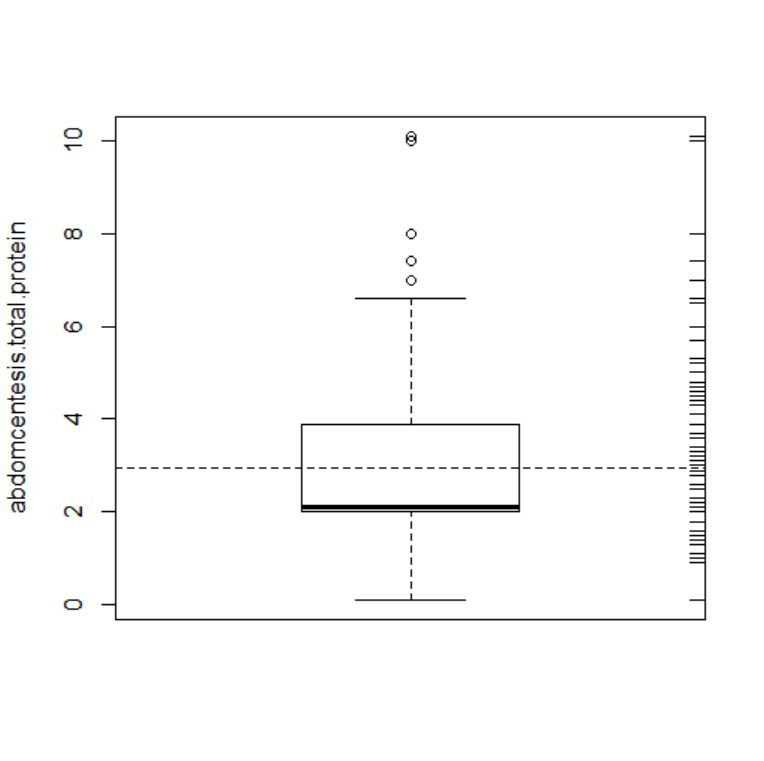
**图10.** rectal temperature**盒图 图11.** pulse**盒图**



**图12.** respiratory rate**盒图 图13.** nasogastric reflux PH**盒图**

**图14.** packed cell volume**盒图 图15.** total protein**盒图**



**图16.** abdomcentesis total protein**盒图**

**4.2 数据缺失的处理**

数据集中有30%的值是缺失的，因此需要先处理数据中的缺失值。

4.2.1 将缺失部分剔除

此方法最简单，直接丢弃所有含有缺失数据的数据行。剔除缺失数据与写入文件的命令如下：

#剔出缺失数据

omitdata = na.omit(mydata)

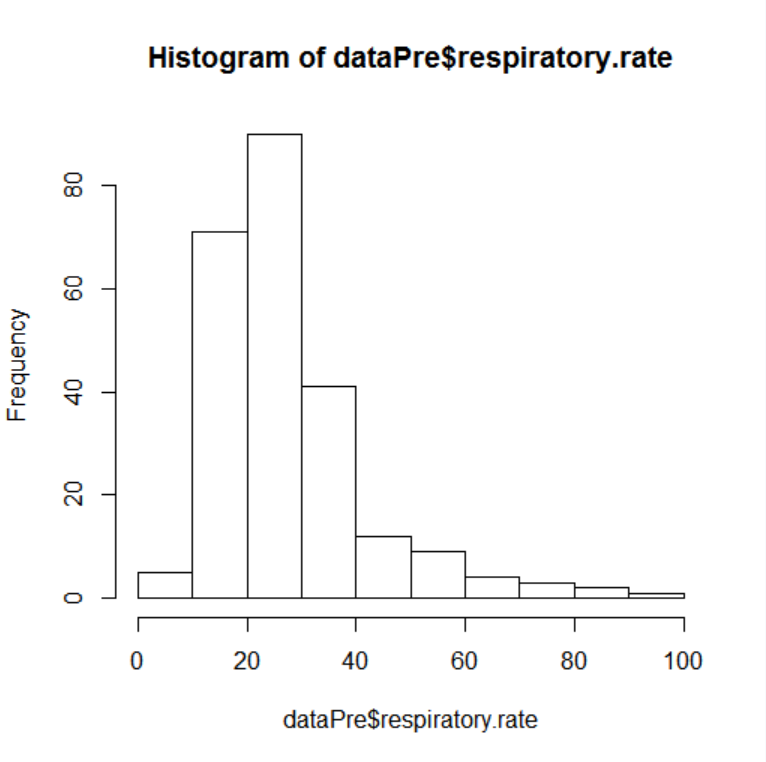
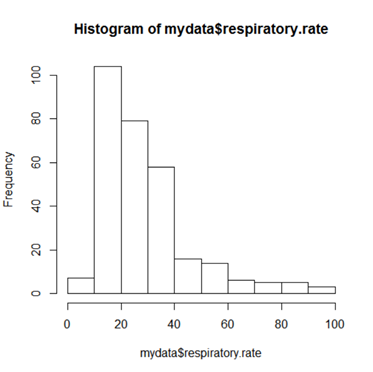
#写入文件

write.table(omitdata,'D:/DataMining/OmitedData.txt',col.names = F,row.names = F, quote = F)

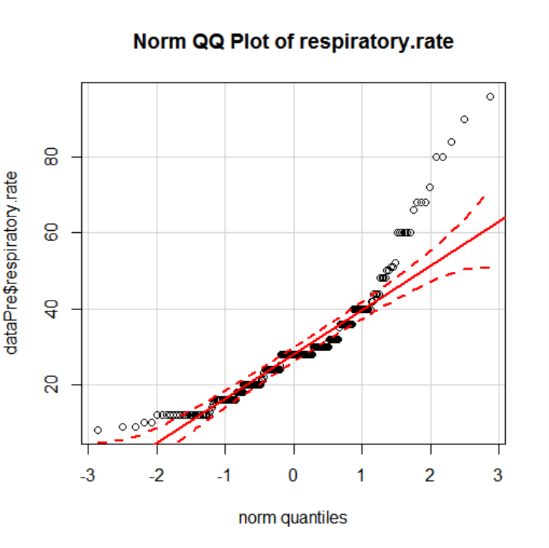
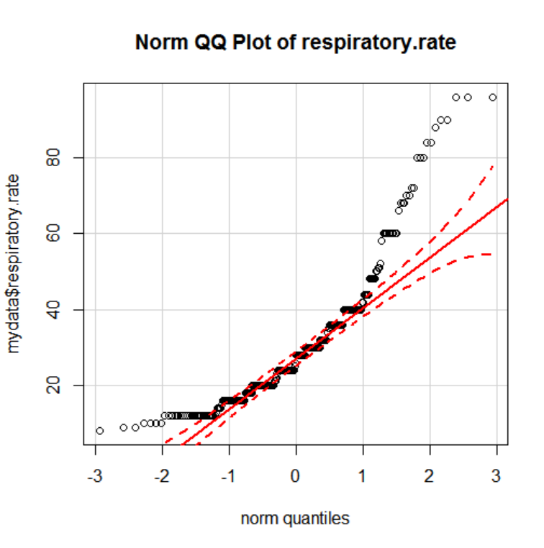
但只在含缺失值的记录在整个数据集中比例很小时适用。例如本实验数据集，缺失值比例较大，包含完整28个属性的数据样本较少，并不适用该方法处理缺失数据。

4.2.2 用最高频率值填补缺失值

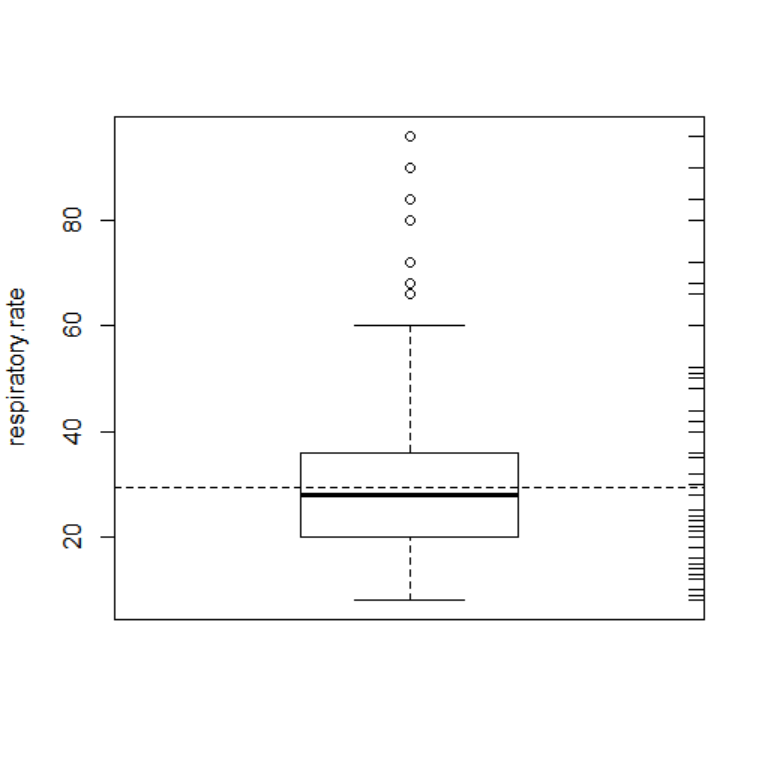
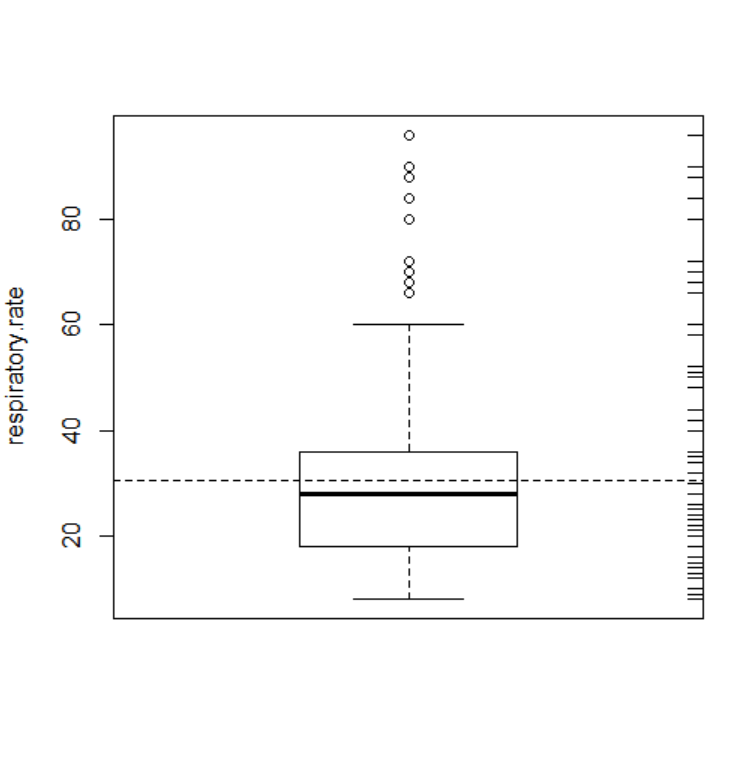
这种方法也比较简单，针对每个属性仅需计算其出现最高的频率值，然后填补到缺失值中即可。为简略对比预处理后的数据集变化情况，画图只选取一个数值属性——respiratory rate属性进行对比。处理前后数据的可视化对比如图17-19所示：



**图17.** **处理前后respiratory rate的直方图对比**



**图18.** **处理前后respiratory rate的qq图对比**

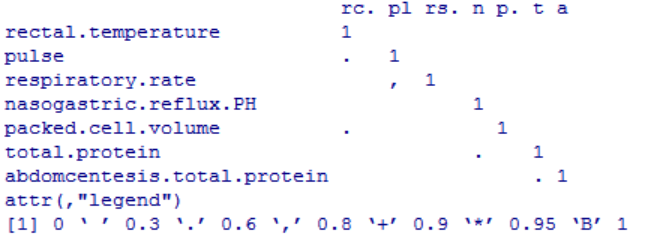


**图19.** **处理前后respiratory rate的盒图对比**

4.2.3 通过属性的相关关系填补缺失值

使用cor函数察看数值变量的相关性，结果如图20所示：

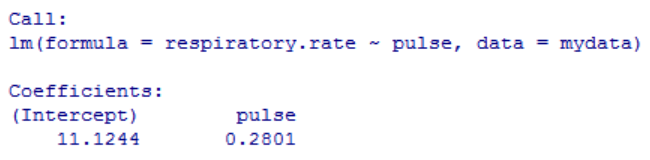
symnum(cor(mydata[,c(4,5,6,16,19,20,22)],use='complete.obs'))



**图20.** **属性相关度**

其中respiratory.rate和pulse相关度超过0.6，所以用这两个属性作相关分析，互相填补缺失数据。用如下代码获得其线性模型：

lm(formula=respiratory.rate~pulse, data=mydata)



**图21** **respiratory.rate与pulse的线性模型分析**

得到结果如图21所示，表示得到的线性模型为pulse=11.1244 + 0.2801\*respiratory.rate

preprocess3 = mydata[-manyNAs(mydata),]

fillpulse <- function(respiratory.rate){

if(is.na(respiratory.rate ))

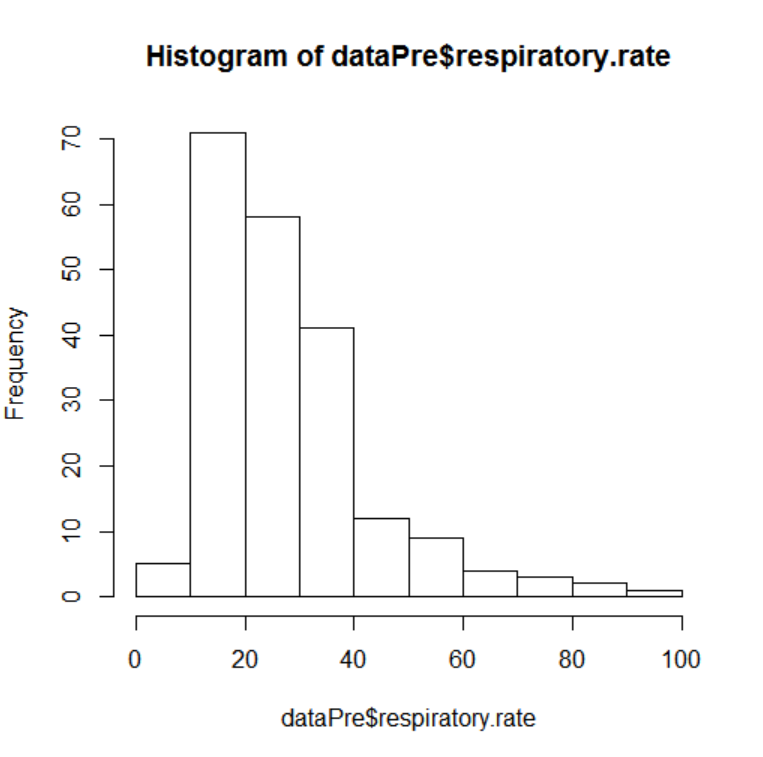
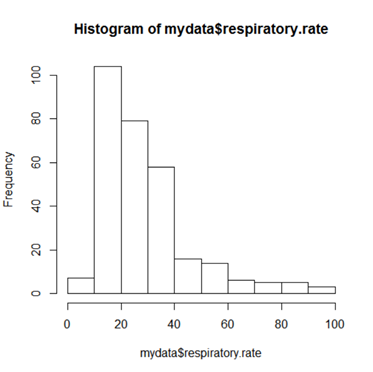
return(NA)

else return (11.1244 + 0.2801\*respiratory.rate)

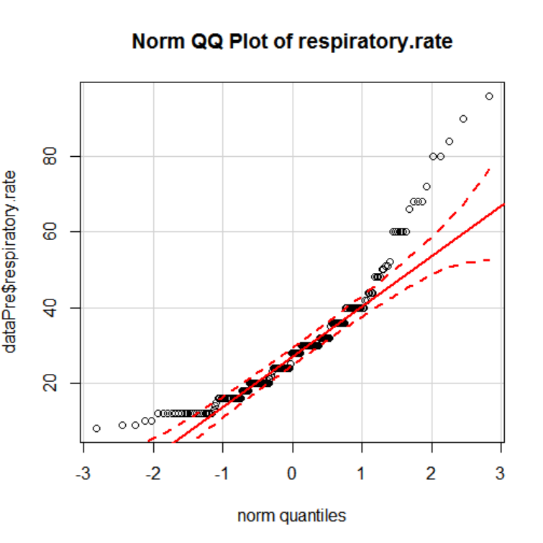
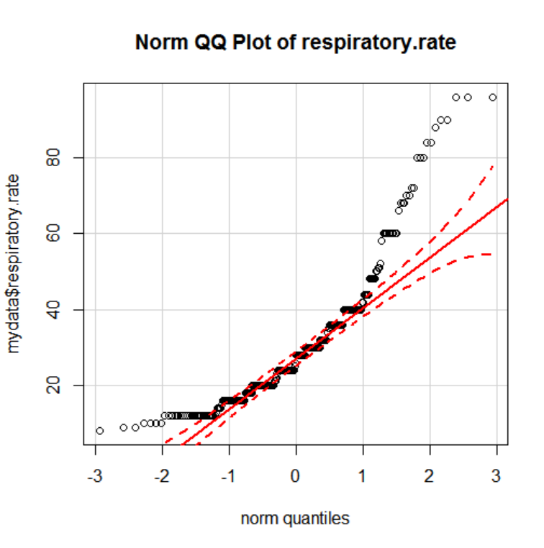
}

preprocess3[is.na(preprocess3$pulse),'pulse'] <- sapply(preprocess3[is.na(preprocess3$pulse),'respiratory.rate'],fillpulse)

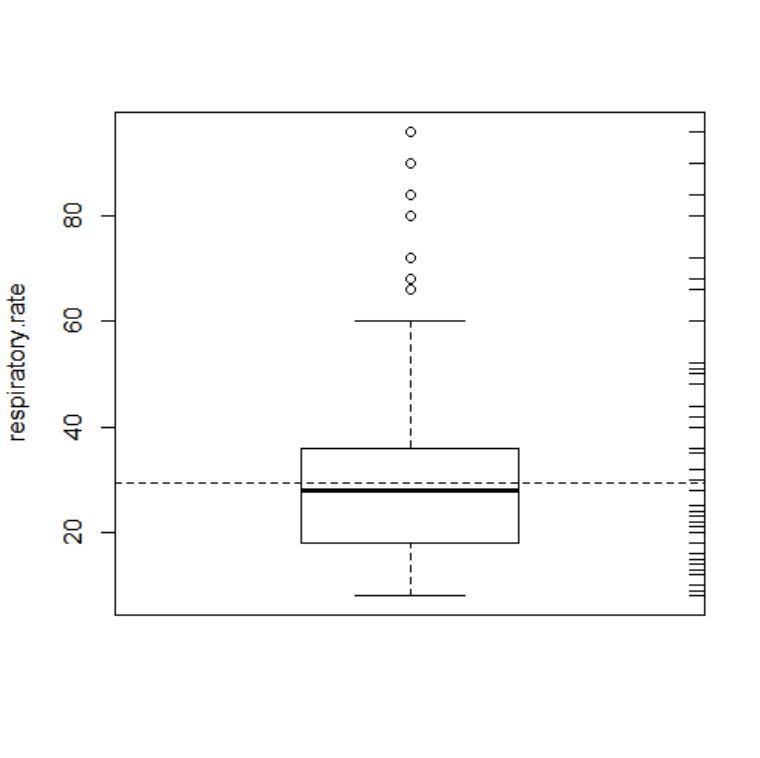
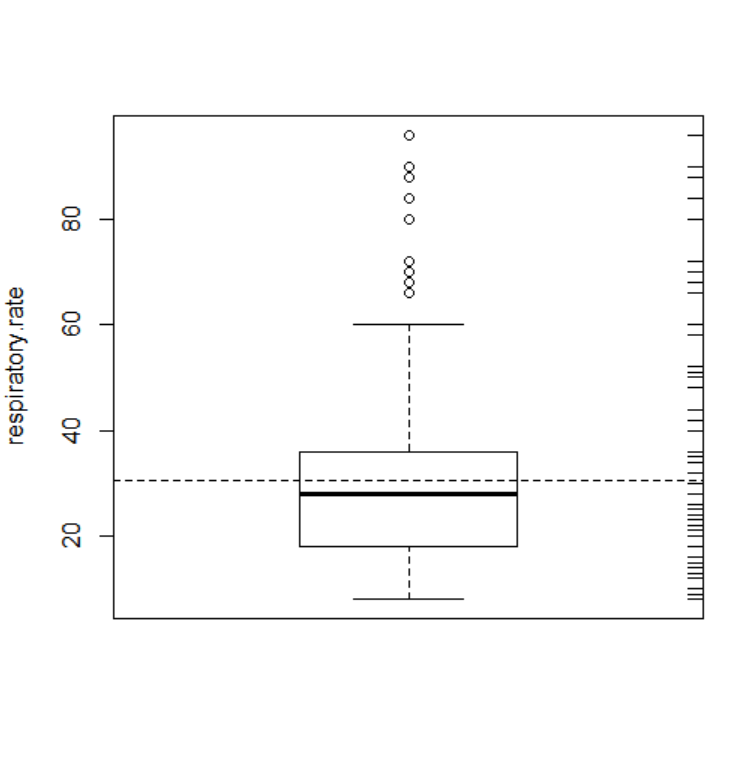
这样就使用线性模型填充了respiratory.rate和pulse的数据。处理前后数据的可视化对比如图22-24所示：



**图22.** **处理前后respiratory rate的直方图对比**



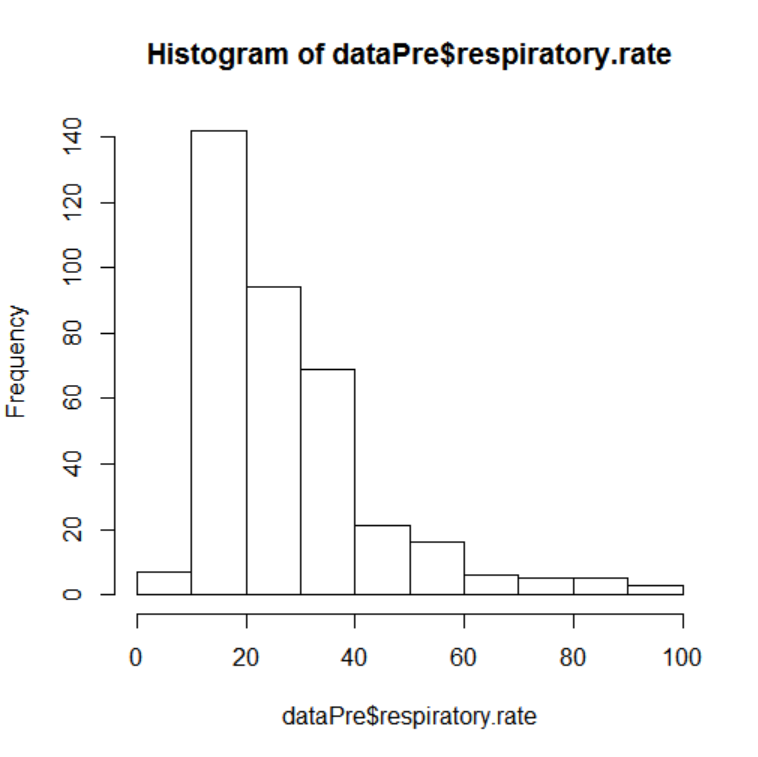
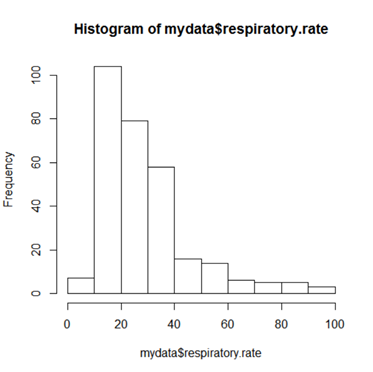
**图23.** **处理前后respiratory rate的盒图对比**



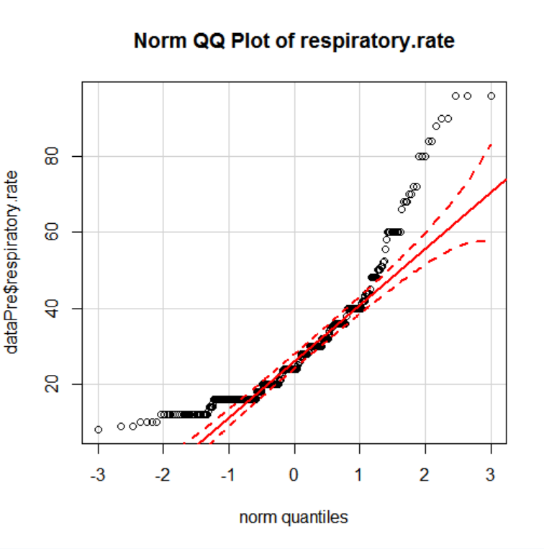
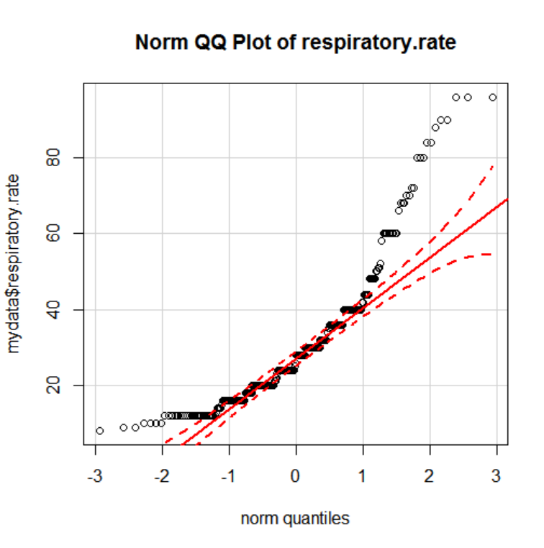
**图24.** **处理前后respiratory rate的盒图对比**

4.2.4 通过数据对象的相似性填补缺失值

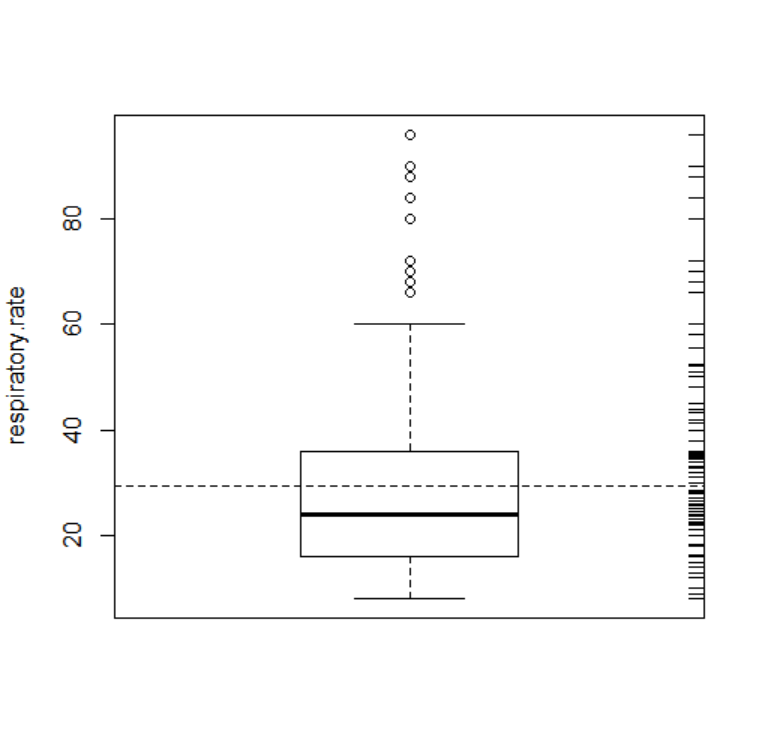
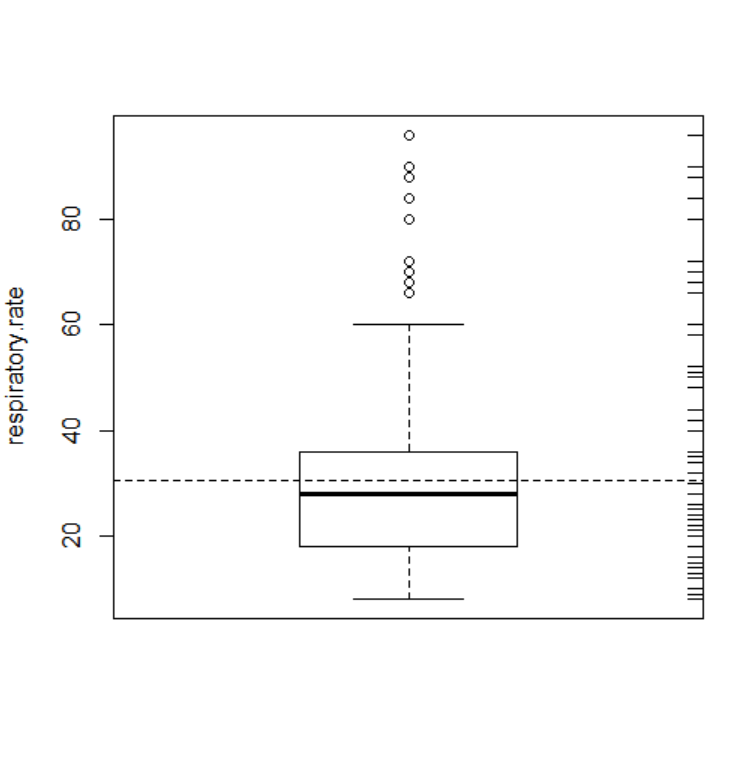
度量相似性时，最常用的指标是欧式距离。我们可通过使用这种度量来寻找与任何含有缺失值的案例最相似的10个样本，并用它们填补缺失值。处理前后数据的可视化对比如图25-27所示：



**图25.** **处理前后respiratory rate的直方图对比**



**图26.** **处理前后respiratory rate的盒图对比**



**图27.** **处理前后respiratory rate的盒图对比**