马的疝病分析

# 实验环境

操作系统：win10

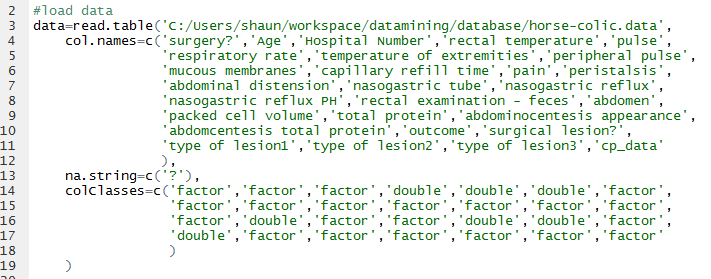
语言：R

# 实验过程

## 数据可视化和摘要

### 数据摘要

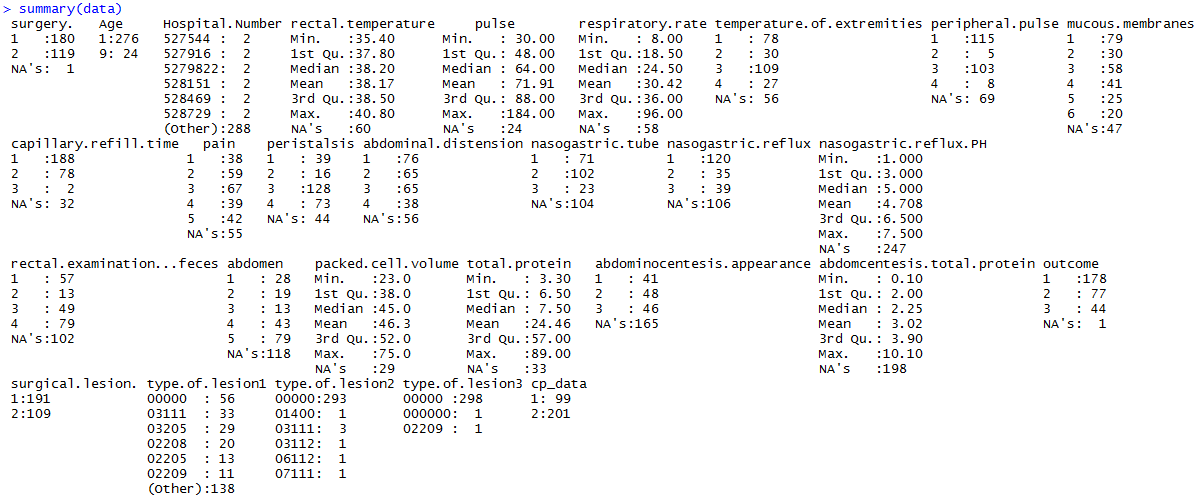
读取文件，同时设定列属性名，设置确实数据的字符串为‘？’，设置列的数据类型。



在R语言中使用summary函数可以方便的获得原始数据的摘要。对标称属性（数据类型为factor）给出每个可能取值的频数，对数值属性给出最大、最小、均值、中位数、四分位数及缺失值的个数。



摘要结果：



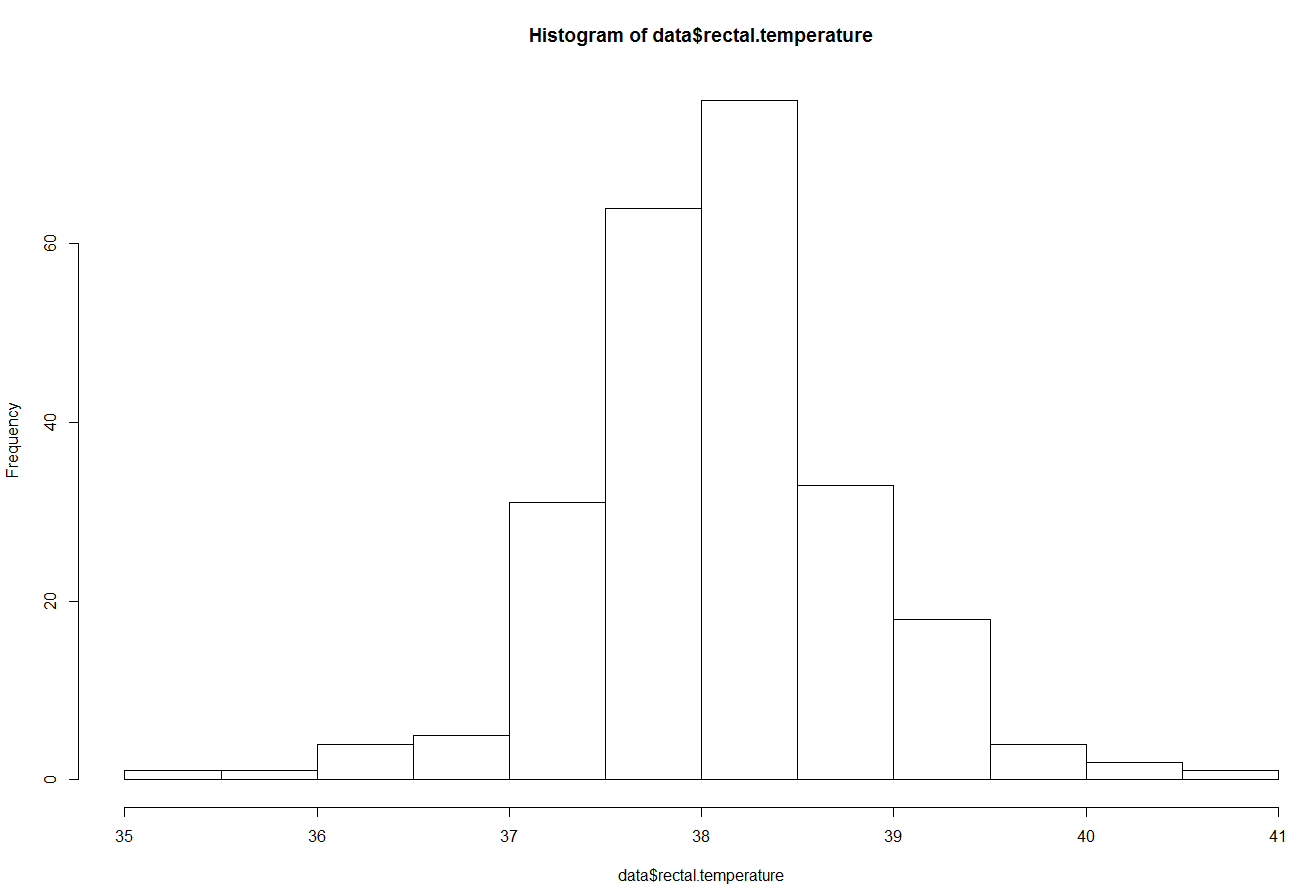
### 数据的可视化

针对数值属性，

* 绘制直方图，如mxPH，用qq图检验其分布是否为正态分布。
* 绘制盒图，对离群值进行识别

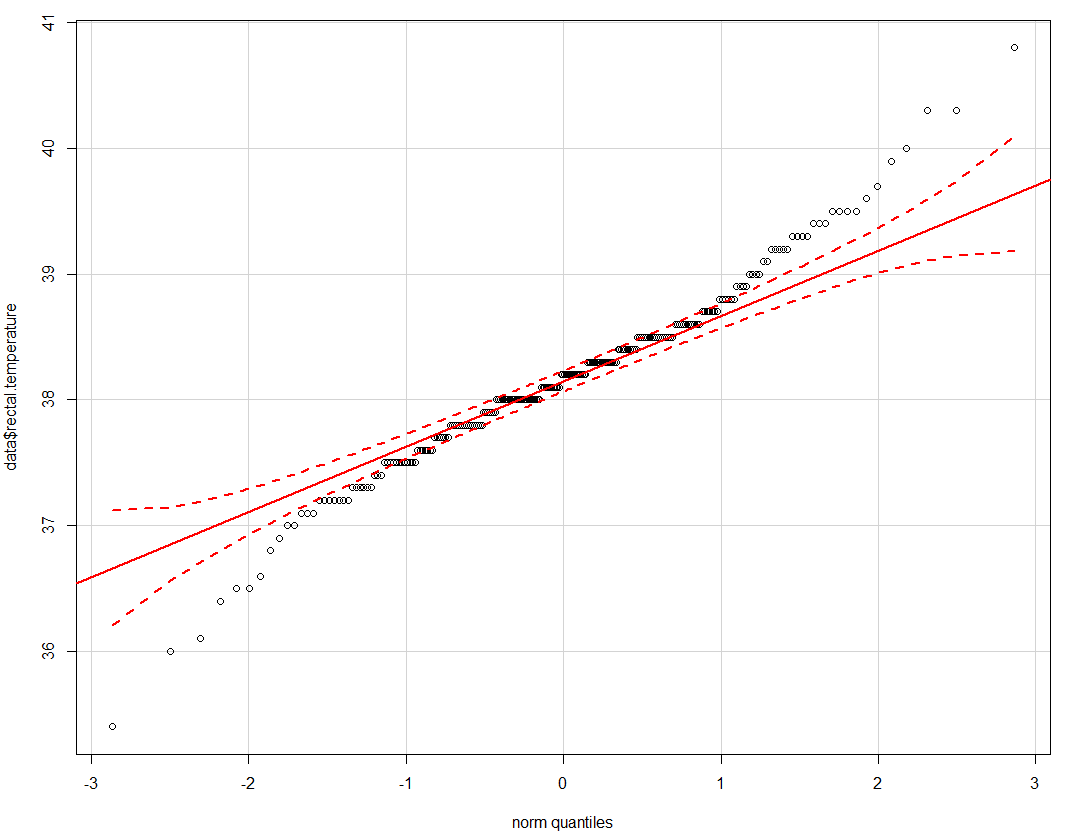
#### 以属性rectal temperature为例绘制直方图





#### 绘制qq图

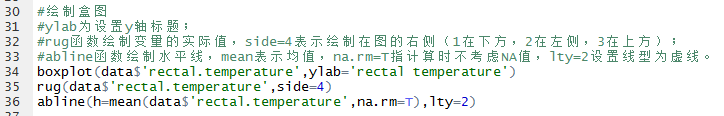




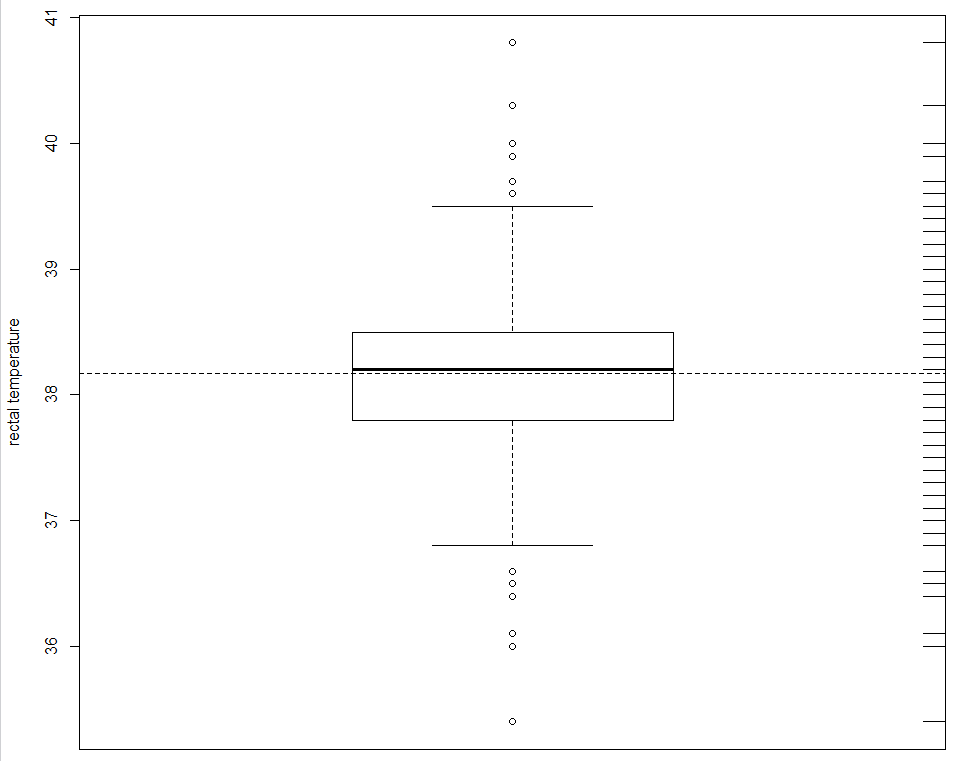
绘制出的直方图纵轴是其频数，横轴是分布区间。QQ图中，红色实线为QQ线，虚线为95%置信度的置信区间。结果表明直方图显示变量rectal temperature的分布近似正态分布。但是存在较多数据点处在95%置信区间之外。

#### 绘制盒图，对离群值进行识别

以rectal temperature为例绘制盒图



结果如下

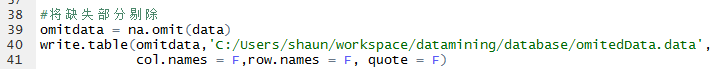


## 数据缺失的处理

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。

分别使用下列四种策略对缺失值进行处理:

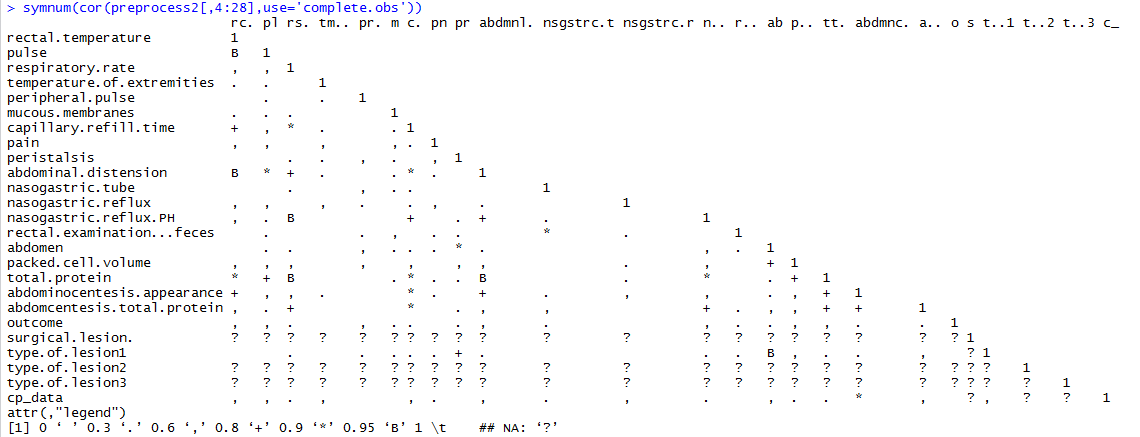
### 将缺失部分剔除



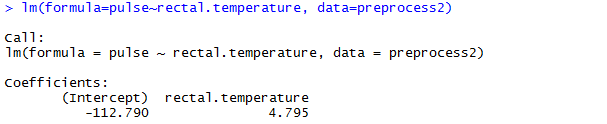
### 用最高频率值来填补缺失值

### 通过属性的相关关系来填补缺失值

利用cor()函数获得变量之间的相关值矩阵，首先重新读入数据为double格式。对读入数据的4至28列进行相关性计算，设定参数use=“complete.obs”后在计算相关值时忽略含有NA的记录。相关值在1周围表示相应的两个变量之间有强正线性相关关系，然后利用其它函数计算线性相关的变量之间的近似函数形式，使得可以通过一个变量的值计算得到另一个变量的值

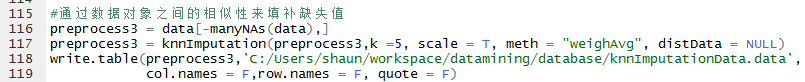
执行symnum函数后得到的相关值矩阵：

容易发现pulse和rectal.temperature属性表现出超过0.95的相关度。以pulse和rectal.temperature属性为例做相关分析，互相填补缺失数据，执行lm函数获得两个属性之间的线性模型：



之后根据两个属性的线性函数填补缺失值。

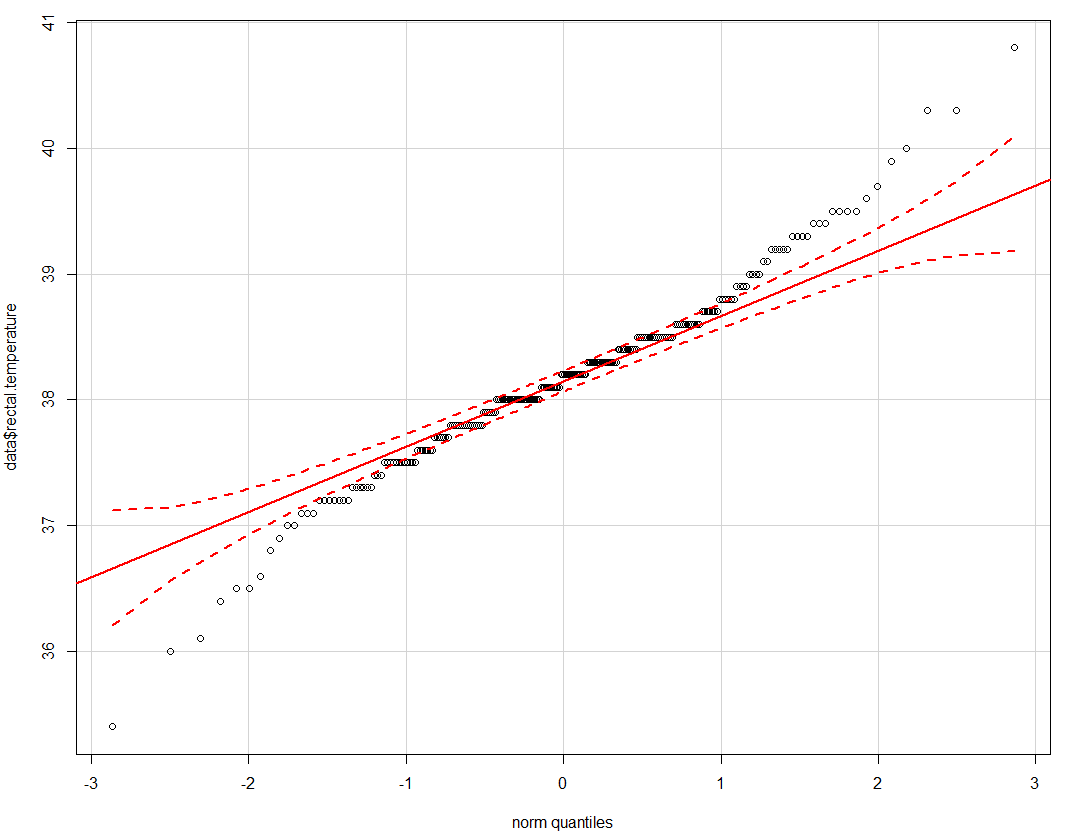
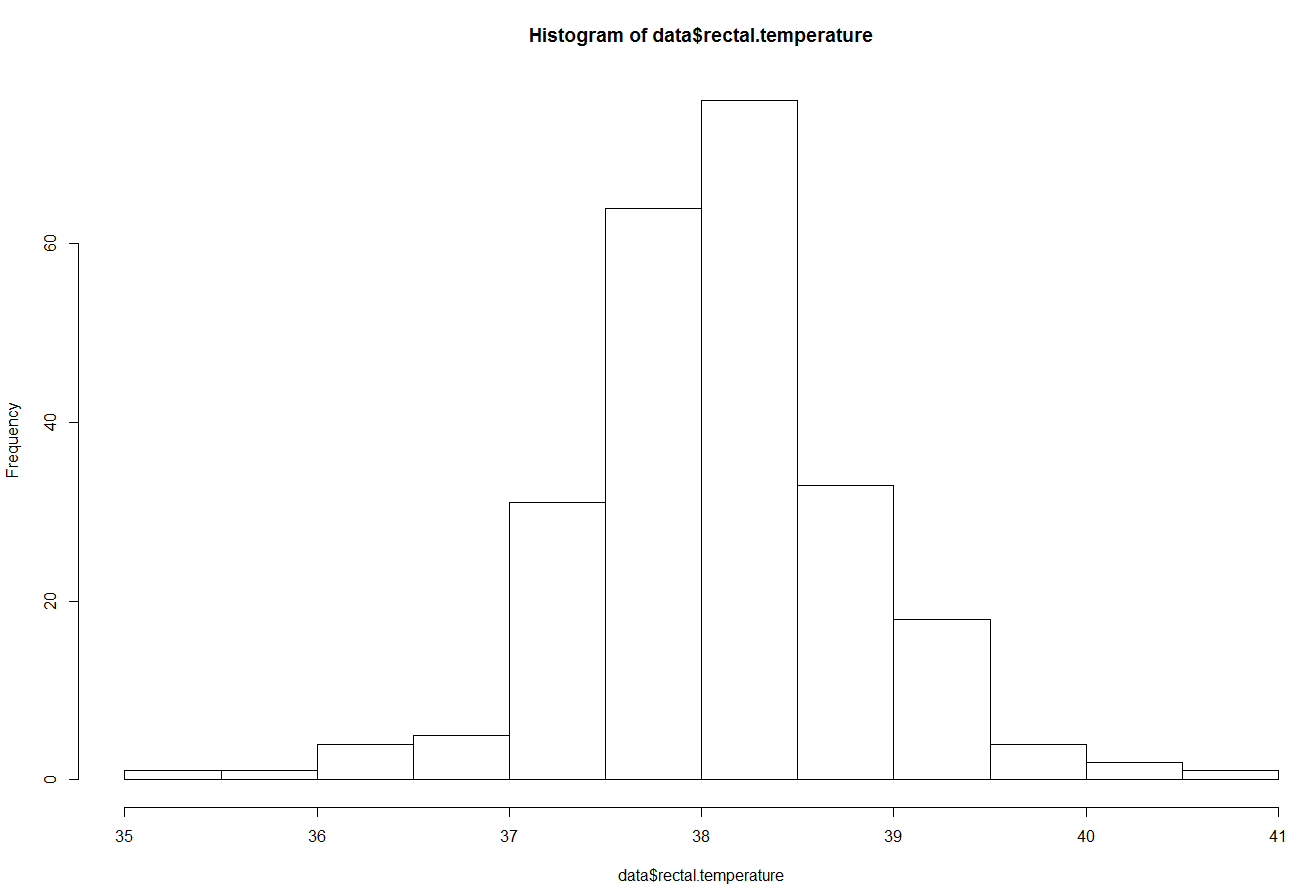
#### 通过数据对象之间的相似性来填补缺失值



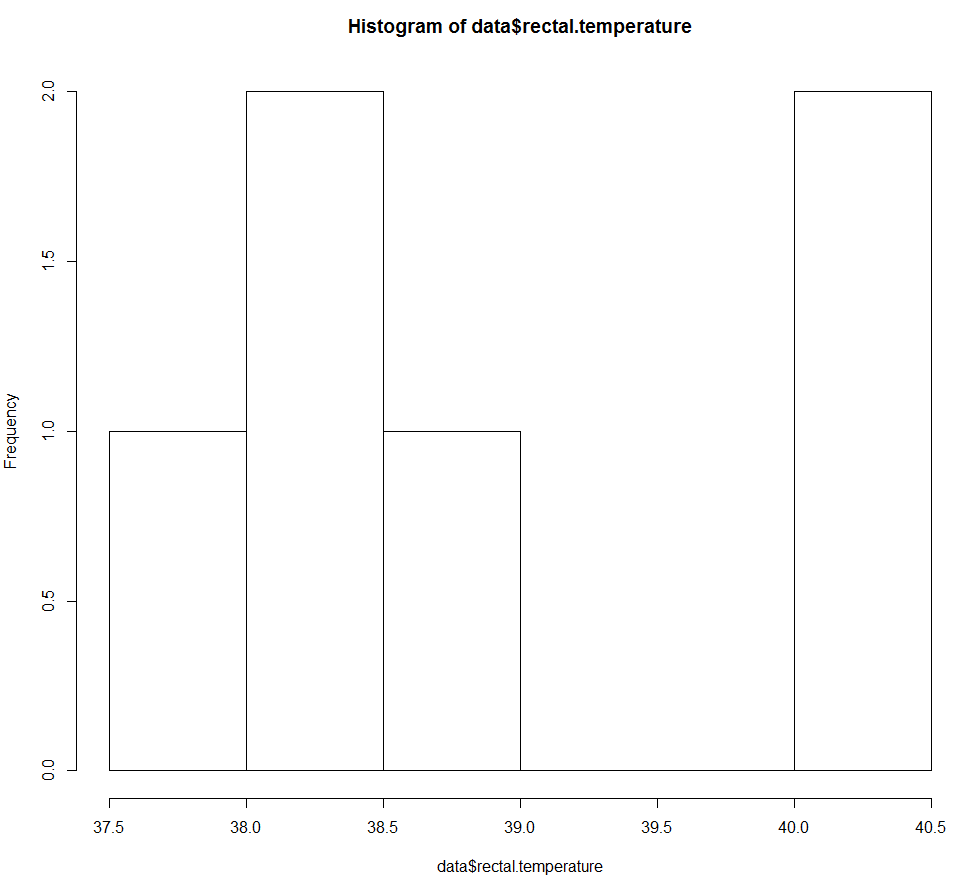
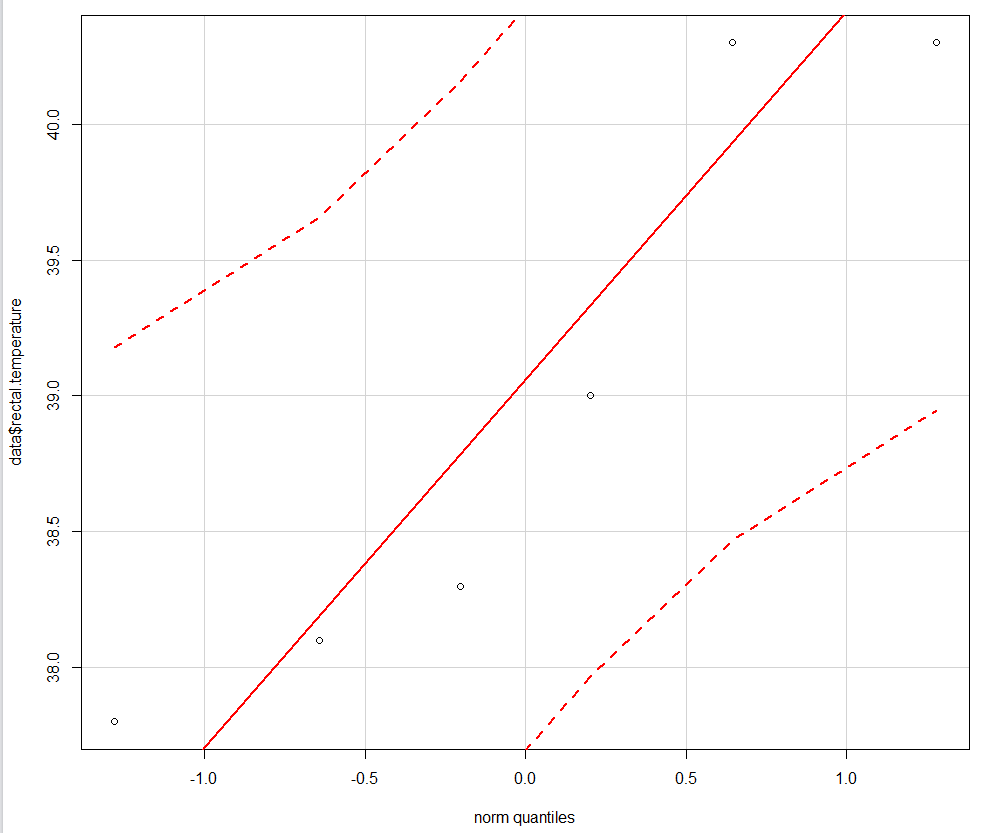
函数knnImputation()用欧氏距离来找到距离任何case最近的k个邻居。采用这些相似数据的加权均值填充缺失数据。权重的大小随着距待填补缺失值的case的距离变化：

w(d)=exp(-d)。

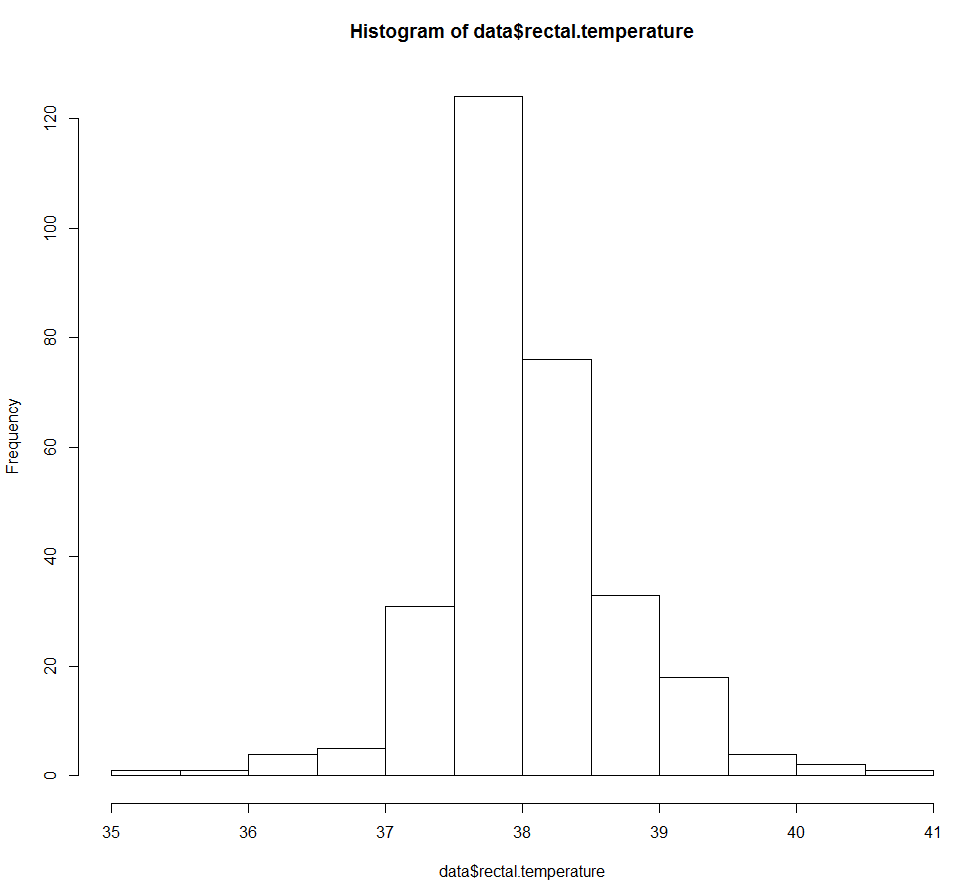
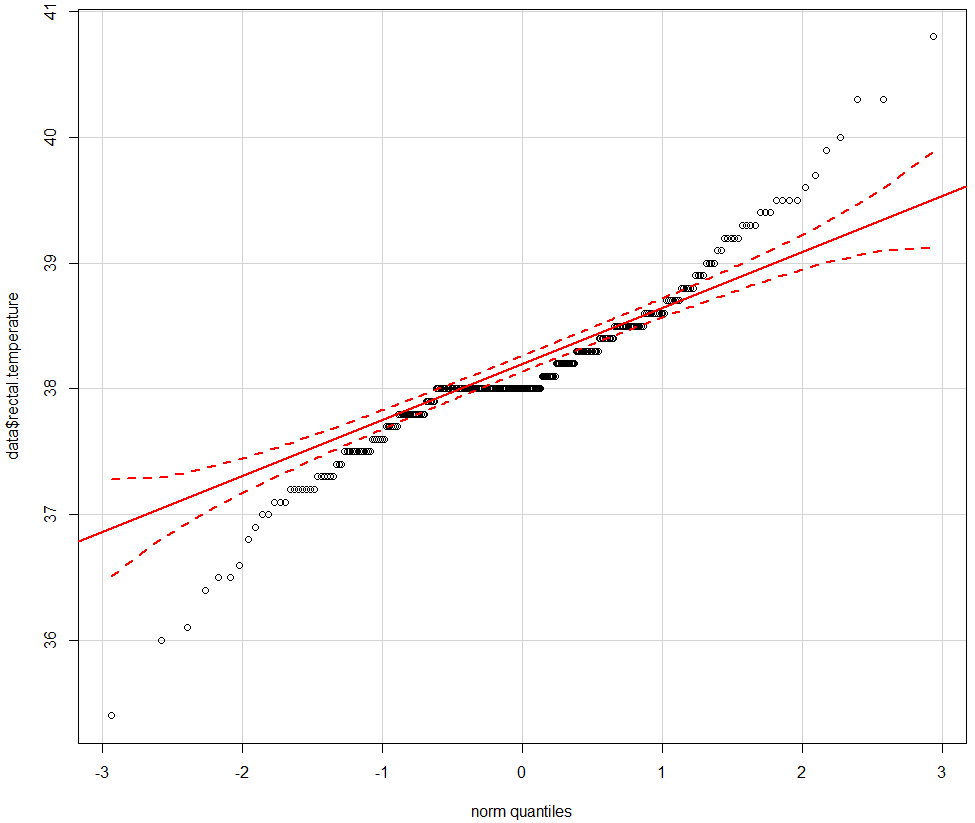
## 四个数据集上rectal temperature 属性的直方图和qq图与原始数据集对比



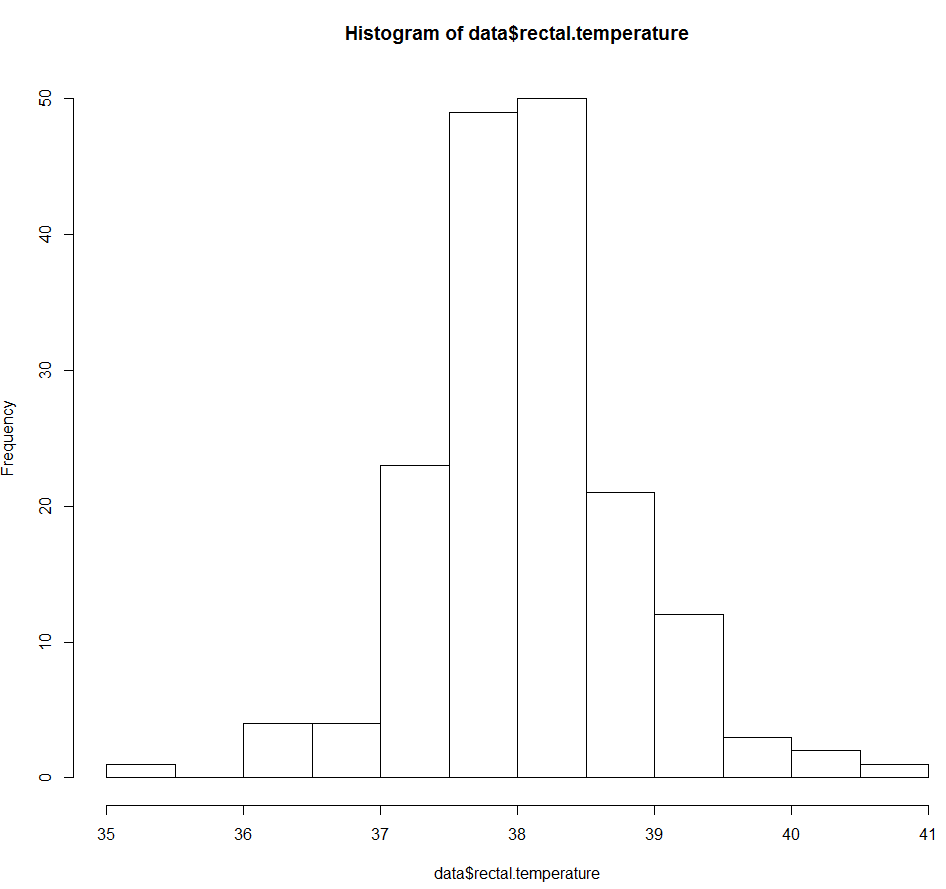
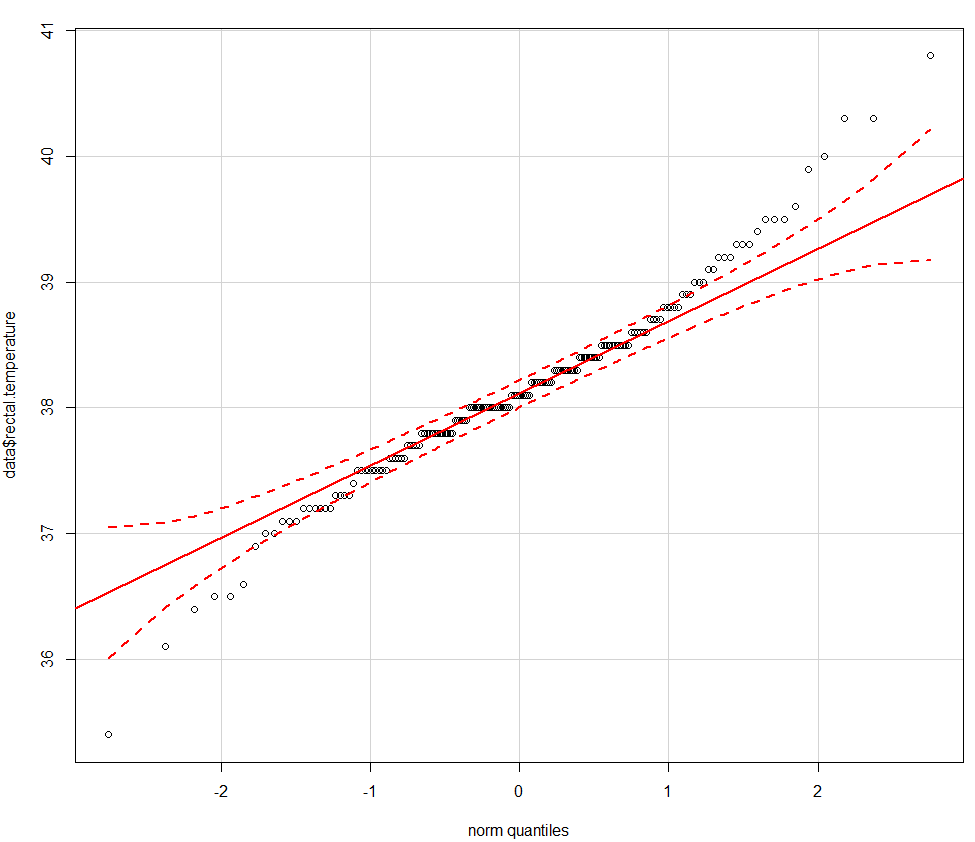
原始数据集

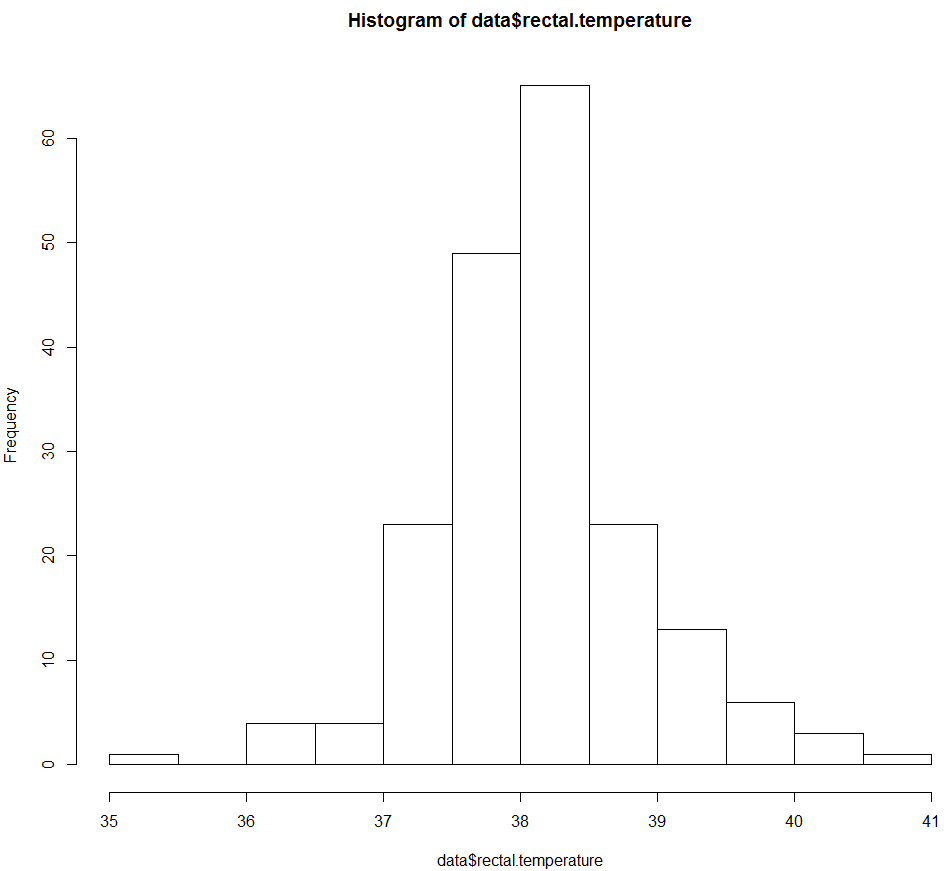
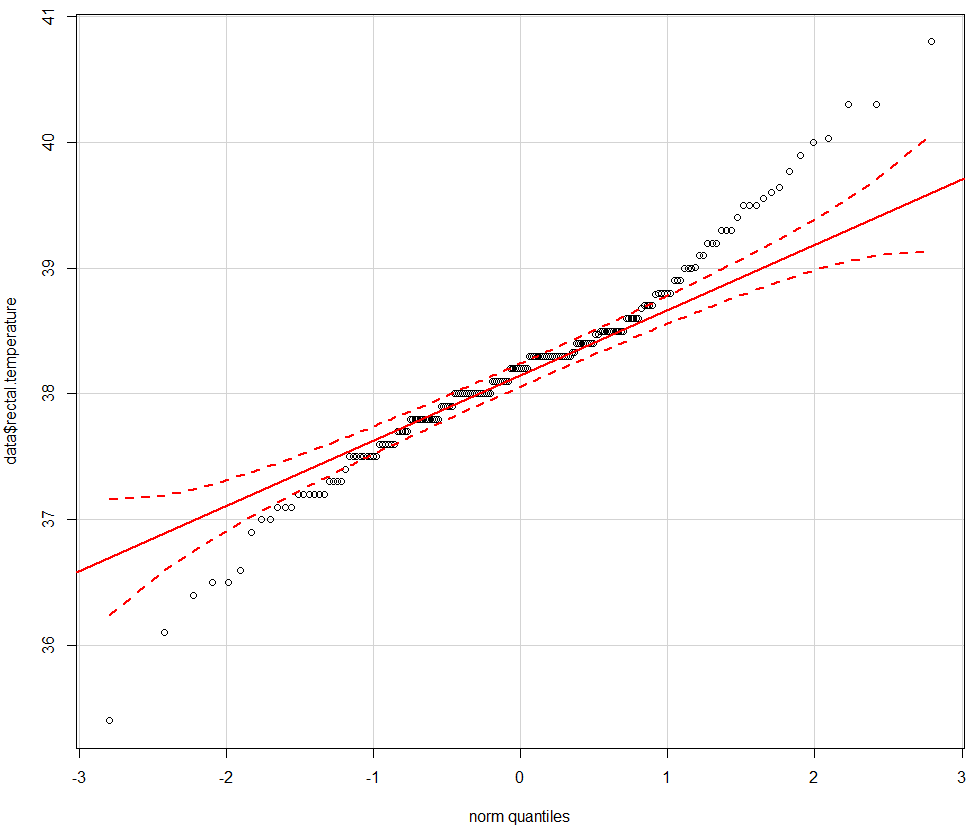
剔除缺失部分的数据集

用最高频率值来填补缺失值的数据集

通过属性的相关关系来填补缺失值的数据集

通过数据对象之间的相似性来填补缺失值的数据集