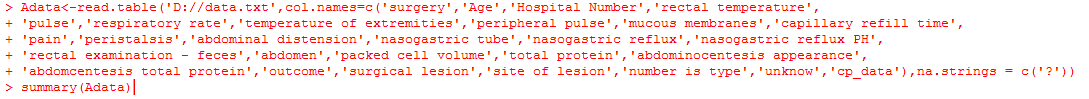
马的疝病分析

计算机学院 2120161028 庞荣

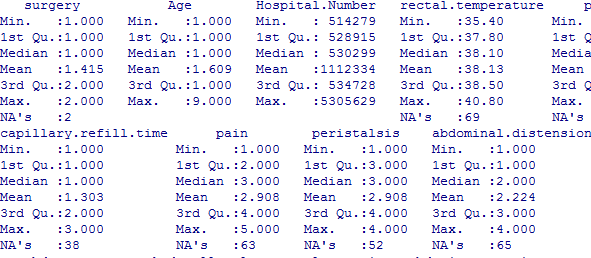
## 数据可视化和摘要

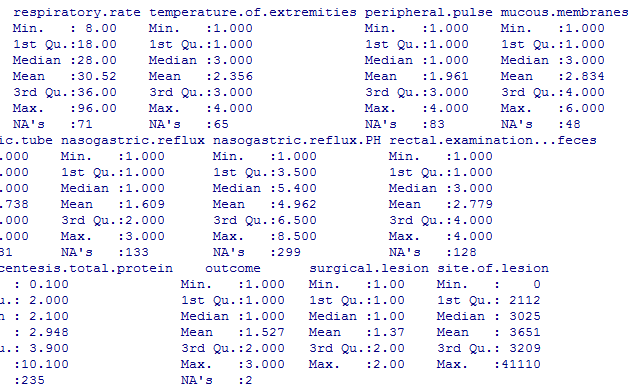
这次的数据挖掘实验主要是对马的疝病的分析数据做处理，疝病是描述马胃肠痛的术语，这种病不一定源自马的胃肠问题，其他问题也可能引发马疝病。所给数据集是医院检测的一些指标。本次实验利用R语言来对该数据集做分析

1. 数据摘要



部分结果展示为

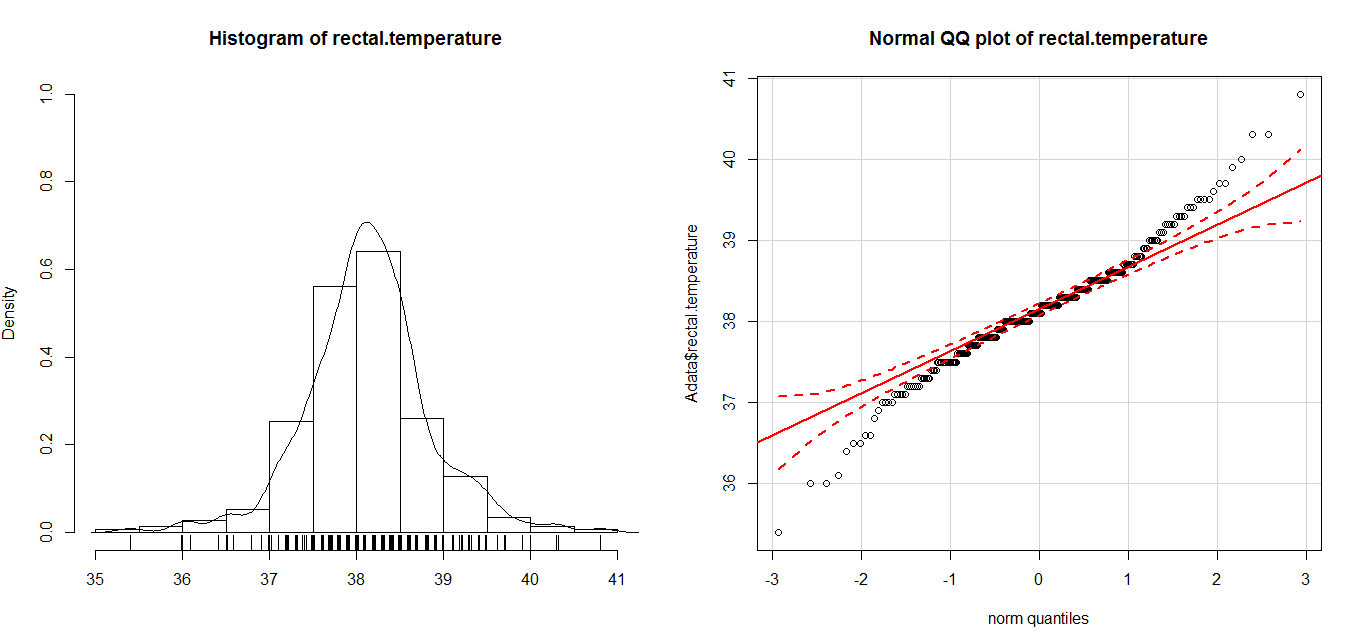




从图中可以看出针对于每一个属性所做出来的数据统计，其中涉及到的统计指标有四分之一位数，中位数，均值，四分之三位数，极值等一系列信息。这些数据可以初步地表现出数据的一种分布，让我们可以宏观地看出这些数据的分布。上面统计数据还将每个属性中缺失的信息显示出来，其中NAs就是缺失值的数量，这将会在下面的操作中做出处理

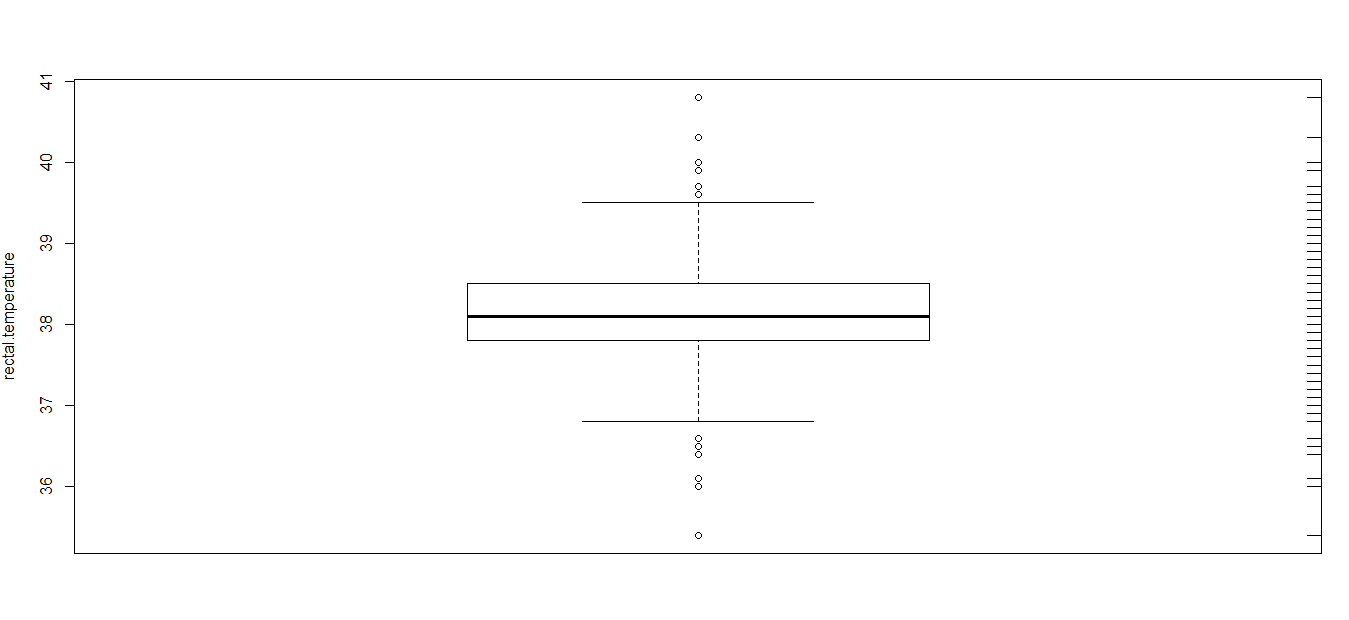
1. 数据可视化

**（1）绘制直方图，如rectal temperature，用qq图检验其分布是否为正态分布**



左图为变量rectal temperature的直方图，这是添加核密度曲线的直方图，曲线的走势可以看到该属性的基本分布，X轴上绘制的是变量的实际值，从这个分布可以很容易识别离群的点，例如我们可以发现有几个数值是显著低于其它所有值。离群点的筛选很重要，对数据挖掘的质量有很大的影响，其他的值大多数都集中在变量的均值附近（例如图中38°的位置左右）；右图为qq图，给出了变量值与正态分布的理论分位数的散点图，同时给出正态分布的95%的置信区间的带状图，由此可看出，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。

**（2）绘制盒图**



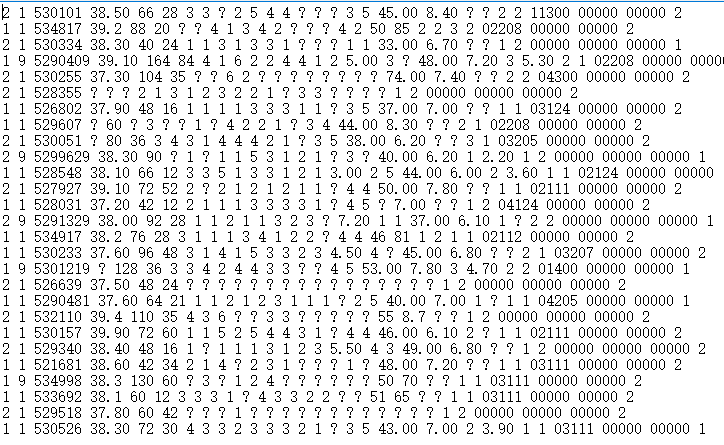
上图是对属性rectal temperature所制作的盒图，盒图上下的小圆圈的值是明显比大部分的值要大，这一类的数据表示的是离群的点，这类离群点会在数据挖掘之前去除掉。在盒子外面的两条水平线分别是最大值和最小值，盒内还有中位数和均值的标注。盒图可以比较清晰地看出一些数值上的分布，而且将一些标志性的值也展示了出来，进一步将该属性值的分布规律展示了出来。

## 二．数据缺失处理

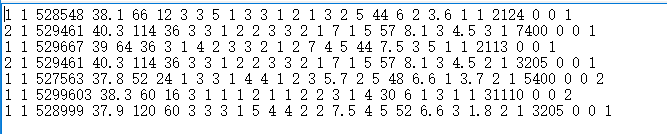
**1. 将缺失部分剔除**

也就是将存在遗漏信息属性值的对象（元组，记录）删除，从而得到一个完备的信息表。这种方法简单易行，在对象有多个属性缺失值、被删除的含缺失值的对象与信息表中的数据量相比非常小的情况下是非常有效的。

omiteddata = na.omit(Adata)



之前的数据



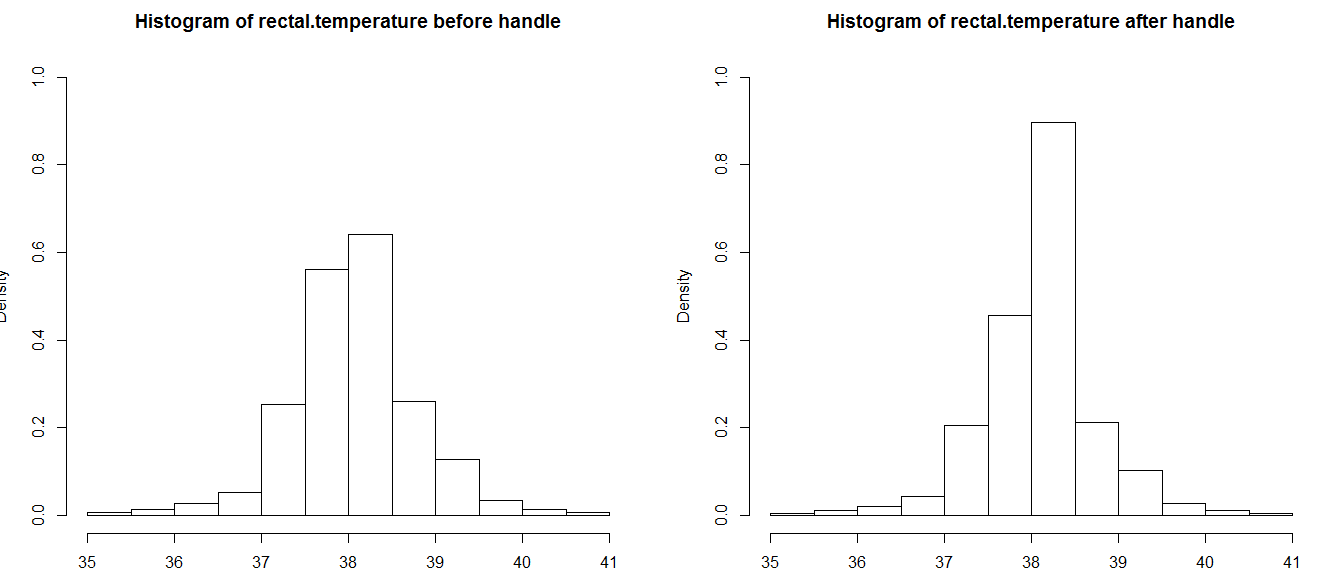
之后的数据

经过去除操作之后原先368条数据就只剩下了上图中的几条数据，这几条数据是完全没有缺失的，但是由于去除的过多已经无法进一步做数据的挖掘了。

**2. 用高频率值来填补缺失值**

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，因此中心趋势值是最自然的选择。有多个代表数据中心趋势的指标，例如平均值，中位数，众数等。最合适的选择由变量的分布决定。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。然而，对于偏态分布，或者离群值的变量来说，选择平均值就不好。偏态分布的大部分值都聚集在变量分布的一侧，因此平均值不能作为最常见的代表。另一方面，离群值（极值）的存在会扭曲平均值，这就导致了平均值不具有代表性的问题。因此，在对变量分布进行检查之前选择平均值作为中心趋势的代表是不明智的，例如某些R的绘制工具。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

Adata1 <- centralImputation(Adata)这是利用中位数来插值



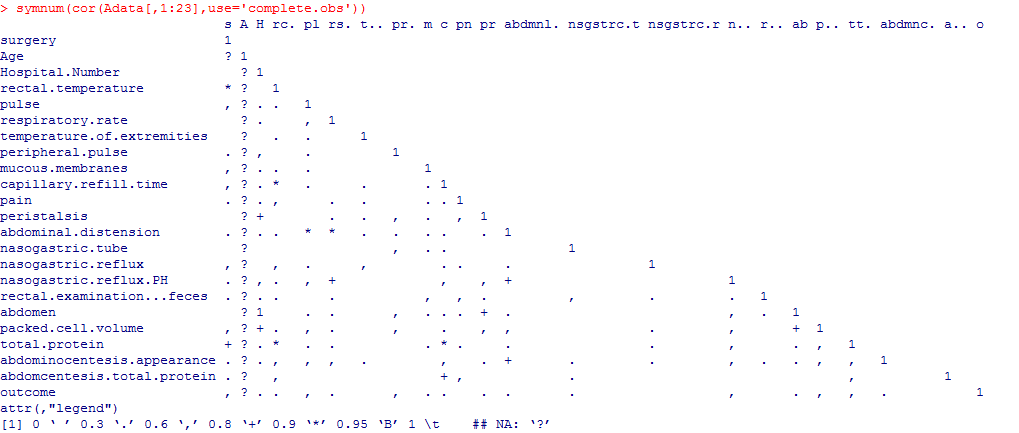
上图是处理前后的rectal.temperature分布的变换，左图是变换前，右图是变换之后的，企图中可以看出，数据更加集中于均值的位置，这可能是由于是利用中位数插值的结果。

**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。

首先探索变量之间的关系

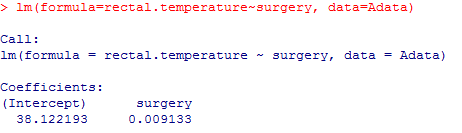
利用symnum(cor(Adata[,1:23],use='complete.obs'))



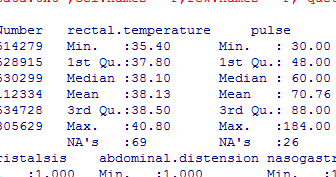
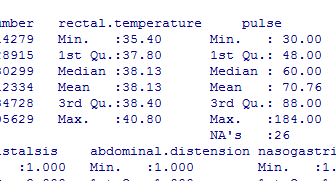
其中\*的位置表示两个变量相关关系在0.95以上的，本次实验中将探索变量surgery和rectal temperature之间的关系，然后利用surgery的值来插值rectal temperature的缺失值。

之后利用lm(formula=rectal.temperature~surgery, data=Adata)

来找变量surgery和rectal temperature之间的具体关系：

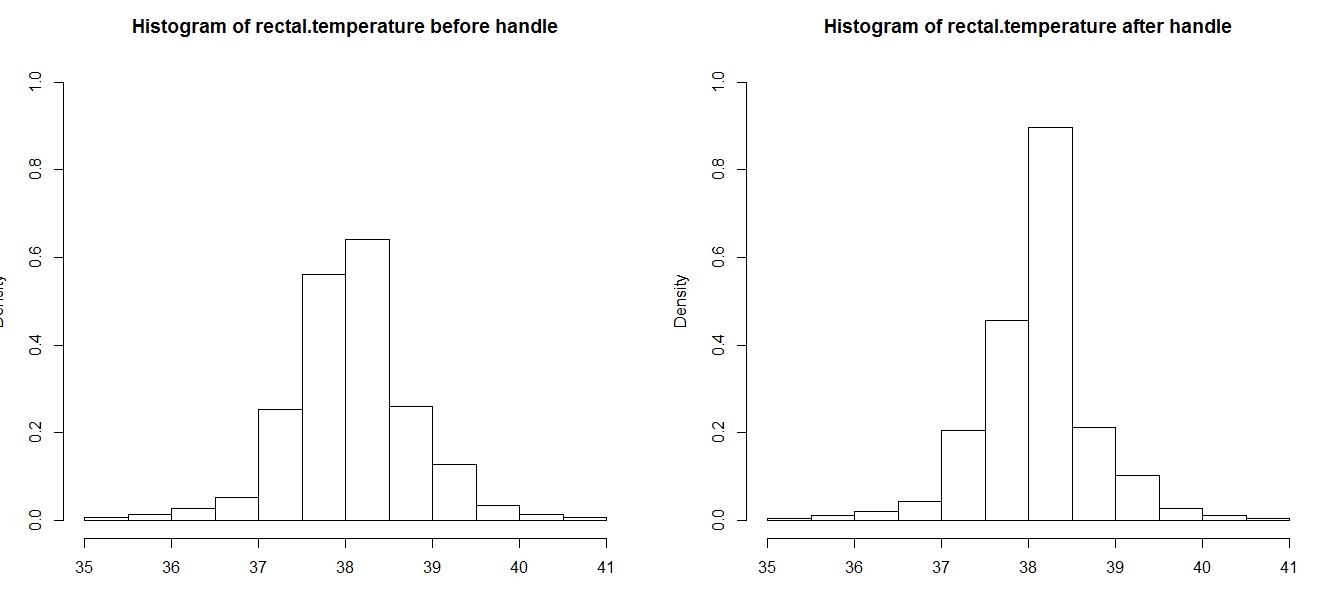


从中可以看出来rectal temperature = 38.122193 + 0.00913 \* surgery，然后利用这个公式插值rectal temperature。

处理之前 处理之后

首先可以看出来rectal temperature的缺失值已经处理了（只针对该属性）



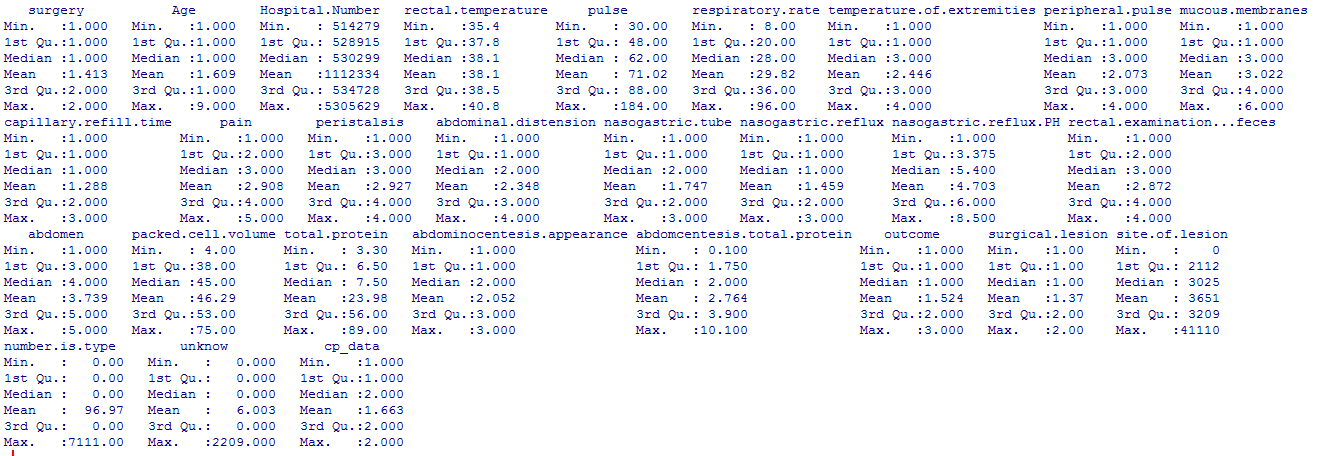
针对于rectal temperature的分布基本上和中位数插值是相同的。

**4. 通过探索案例之间的相似性来填补缺失值**

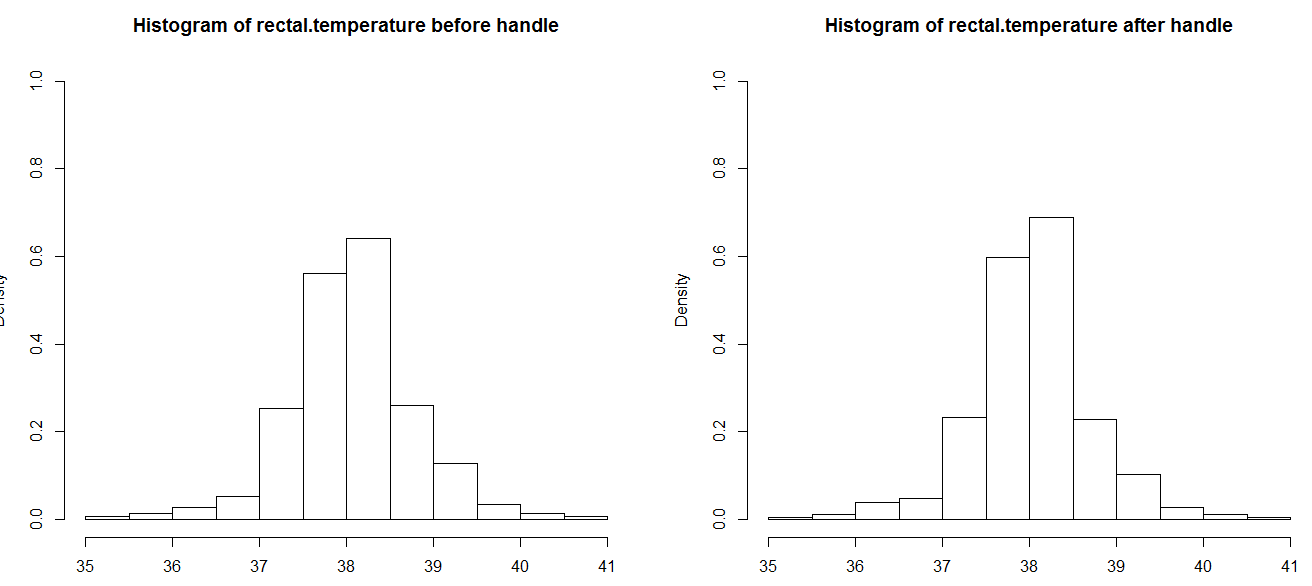
不同于探索数据集列之间的相关性，尝试使用行之间的相似性来填充缺失值。我们可以使用这种方法来填补剔除了那两个含有太多的NA值得样本外的其它缺失数据。

在本次的实验中是利用kNN的方法来探索不同案例之间的相似性

利用a <- kNN(Adata,k = 3)



这是操作之后的结果可以看出所有案例的缺失值都已经被处理了



这是针对rectal temperature的分布，这个处理的结果与2，3方法处理出来的相比更接近与原先数据的分布，所以这样的插值相对于比较科学。