**数据挖掘**

**--作业一实验报告**

学 院：计算机学院

专 业：委培生

年 级：2016 级

姓 名: 李昱燃

学 号：2620160009

提交日期：2017/4/16

**一、实验目的**

### 根据现有马的疝病相关的数据集，获得数据的统计摘要和可视化结果。

### 用不同的策略对数据中的缺失值进行处理，并可视化地对比新旧数据集。

**二、运行环境**

* Windows 32
* R i386 3.3.3

**三、实验步骤及内容**

1. 准备工作

所需数据集是医院检测的有关马疝病的一些指标。共368个样本，27个特征。下载地址：

<http://archive.ics.uci.edu/ml/datasets/Horse+Colic>

下载后得到的horse-colic-data.txt和horse-colic-test.txt分别为包含300个实例的训练集和包含68个实例的测试集。

下文实验步骤的描述将以训练集为例，测试集类似。

1. 数据可视化和摘要

（1）数据摘要

安装RMySQL和DMwR两个库

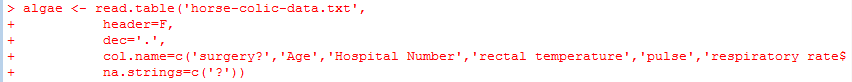


更改存储路径

为了方便，把horse-colic-data.txt和horse-colic-test.txt两个数据集文件存储到当前运行R的目录下，可以再R命令行下用命令getwd()来获取该目录。

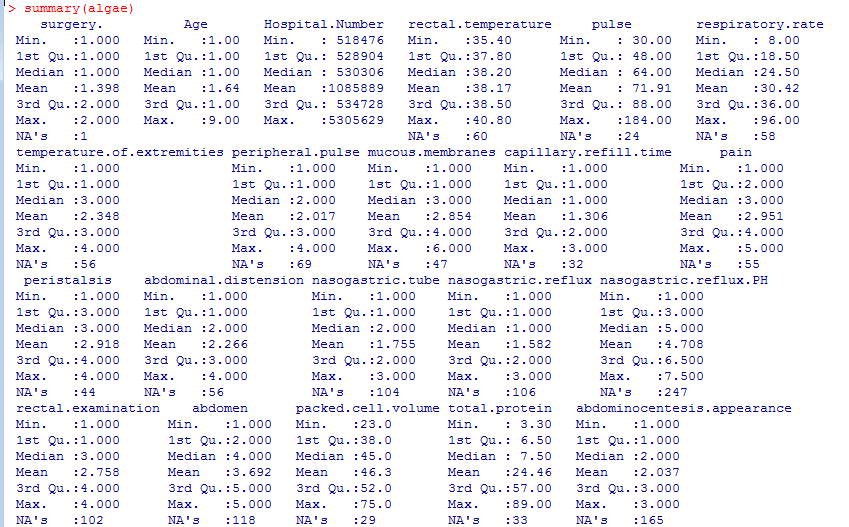
加载数据集

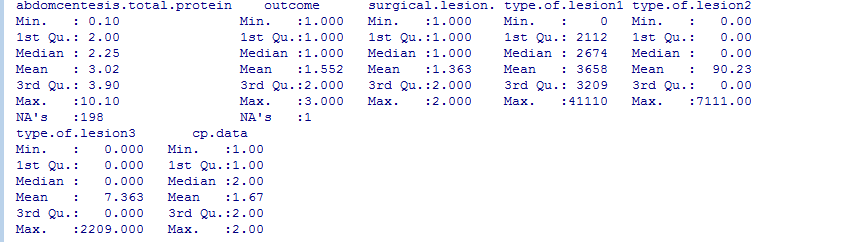
library



给各属性列加上标头，将“？”解释为未知值。

用summary()函数获取数据的描述性统计摘要，结果如下图：

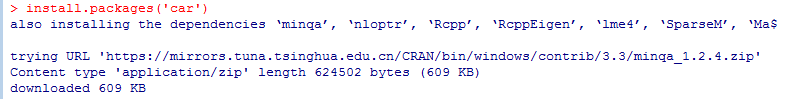




内容包括最大（Max）、最小（Min）、均值（Mean）、中位数（Median）、四分位数（1st Qu.和3rd Qu.）及缺失值的个数（NA’s）。

1. 数据可视化

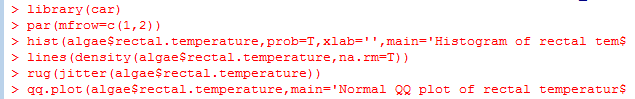
安装添加包car



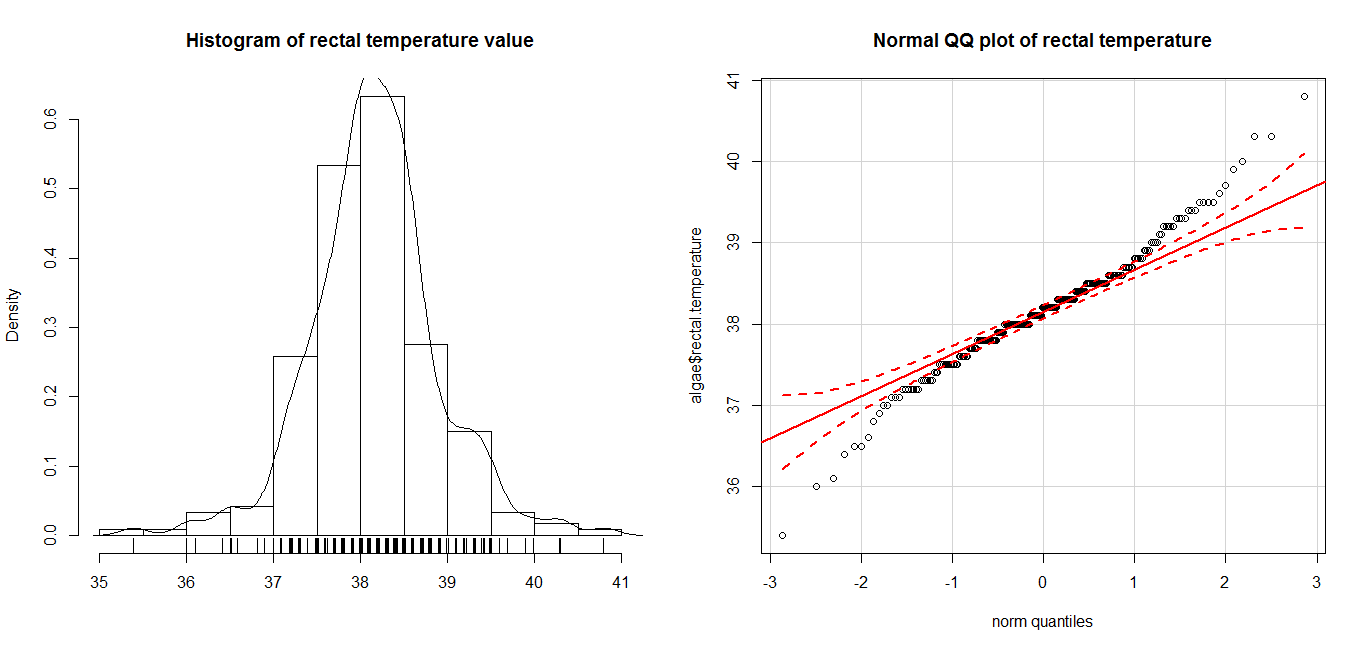
绘制直方图和qq图

首先我们需要选出数值属性，共有7个，分别为：rectal temperature、pulse、respiratory rate、nasogastric reflux PH、packed cell volume、total protein和abdomcentesis total protein，对应训练集中的第4、5、6、16、19、20和22列。

绘制**rectal temperature**属性的直方图和qq图，语句如下：



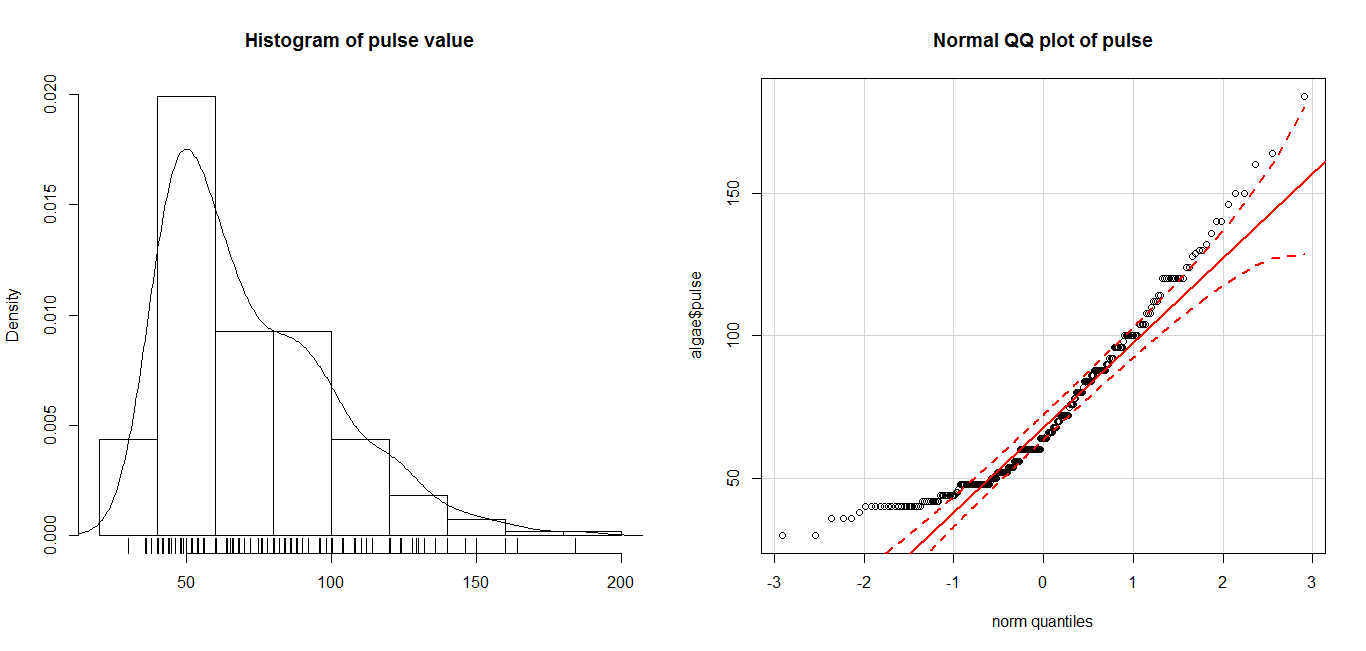
输出结果：



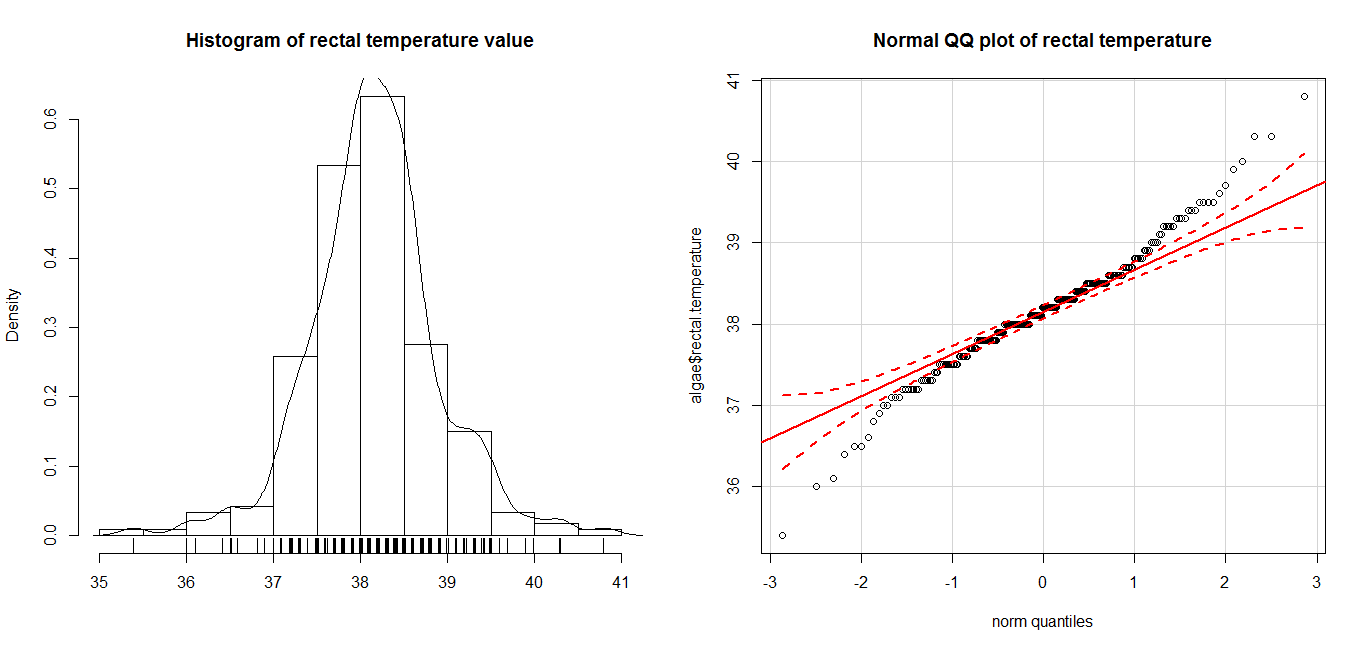
左图为直方图，可以看到，该属性的分布基本为正态分布，用右图的qq图得以检验。

同样的，得到其余6个数值属性的直方图和qq图。

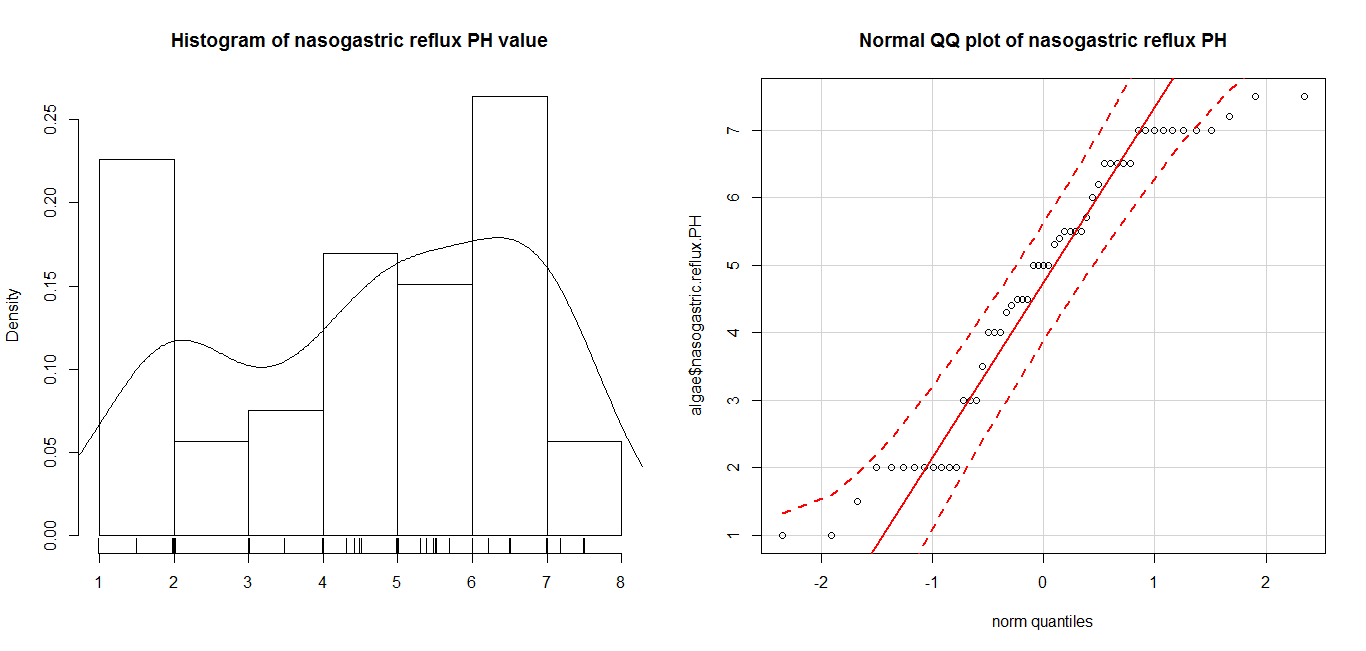
**pluse**属性：



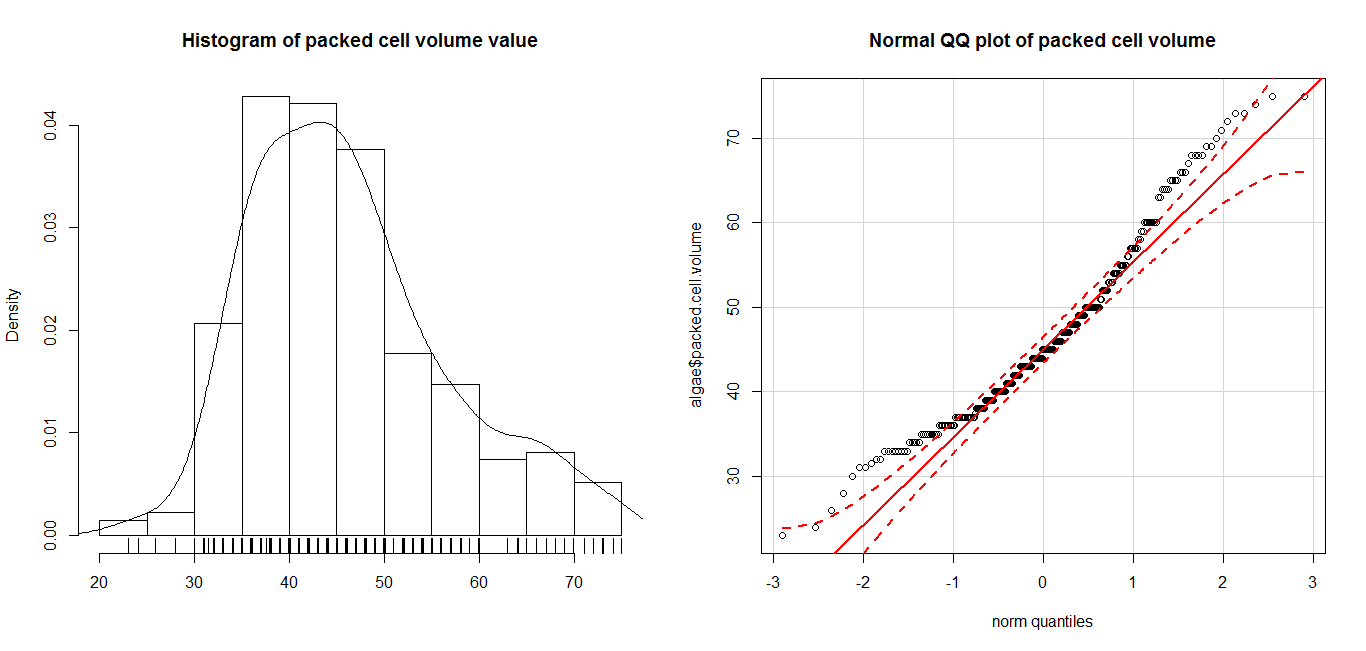
**respiratory rate**属性：



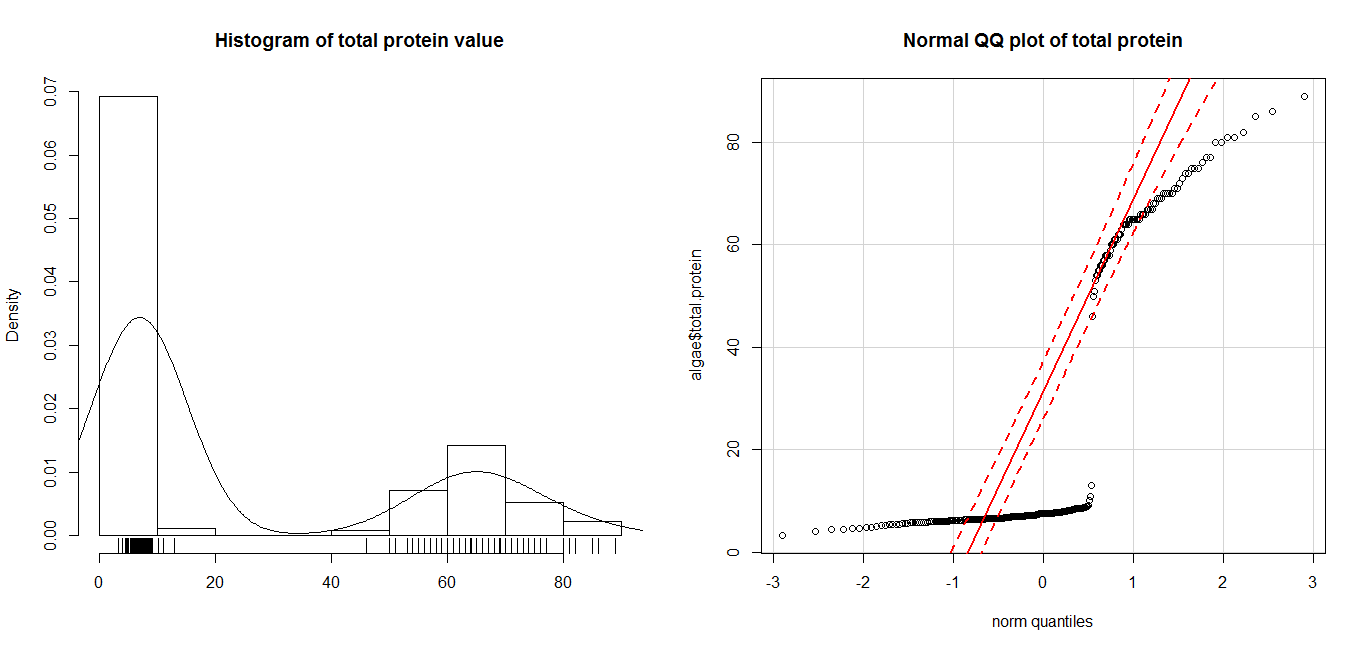
**nasogastric reflux PH**属性：



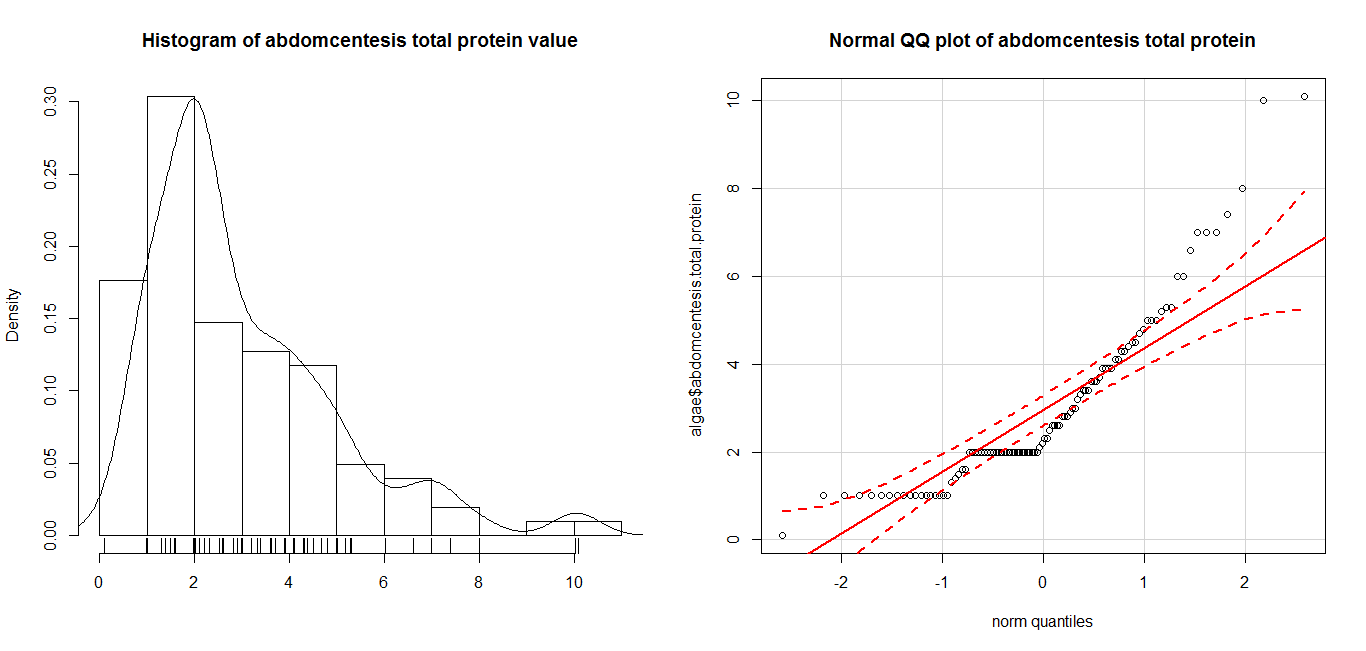
**packed cell volume**属性：



**total protein**属性：



**abdomcentesis total protein**属性：

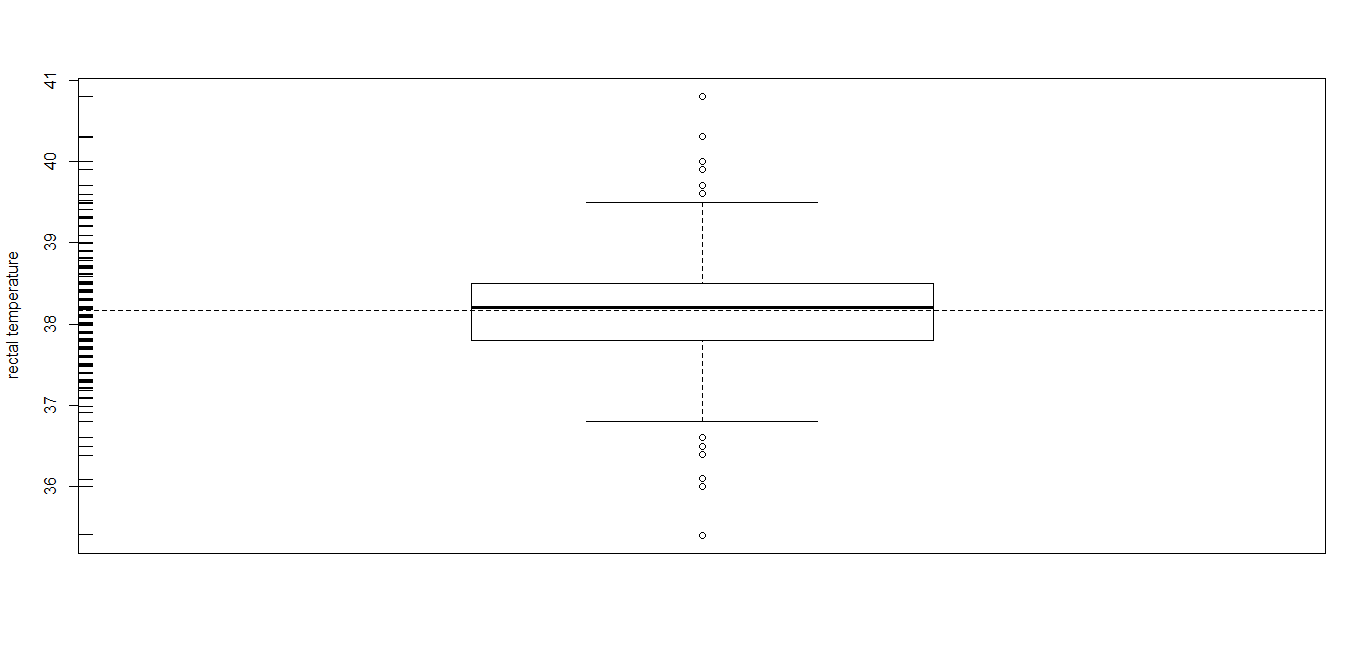


绘制盒图

绘制rectal temperature属性的盒图，语句如下：

rectal temperature盒图语句

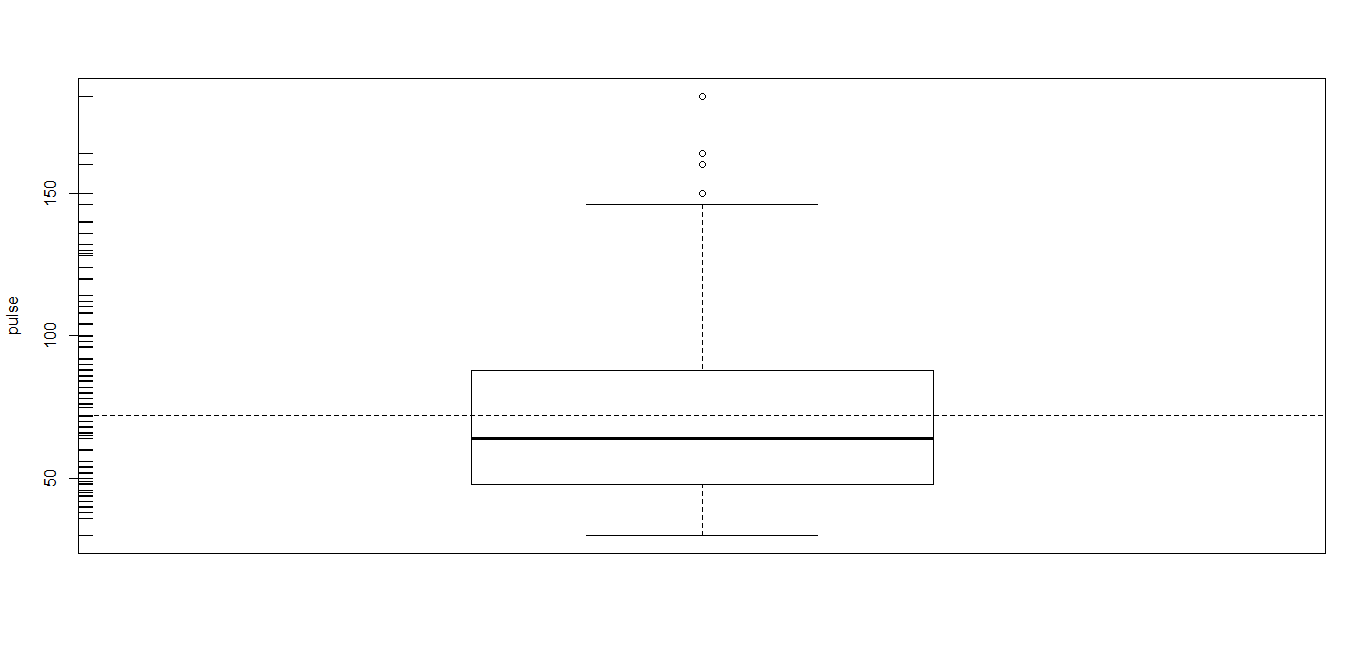
输出结果：



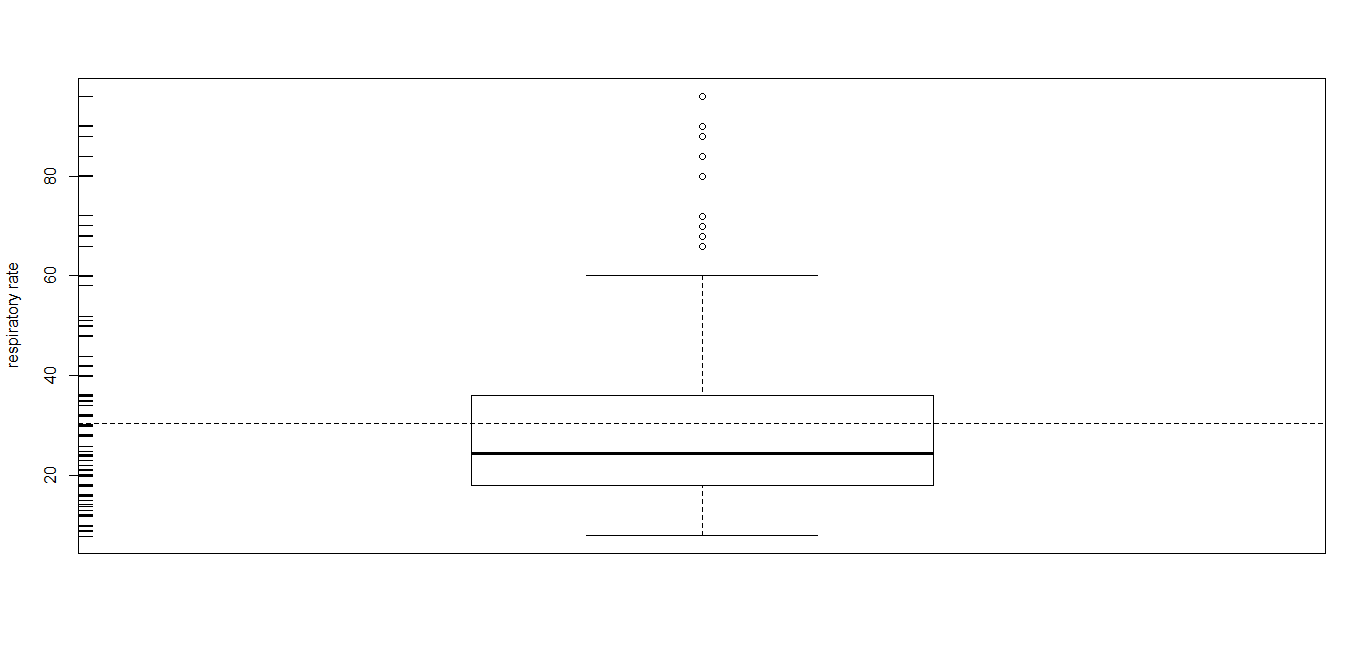
从盒图中我们可以清楚地看到该属性数据的最大值、最小值、均值、中位数、四分位数等信息。

同样的，得到其余6个数值属性的盒图。

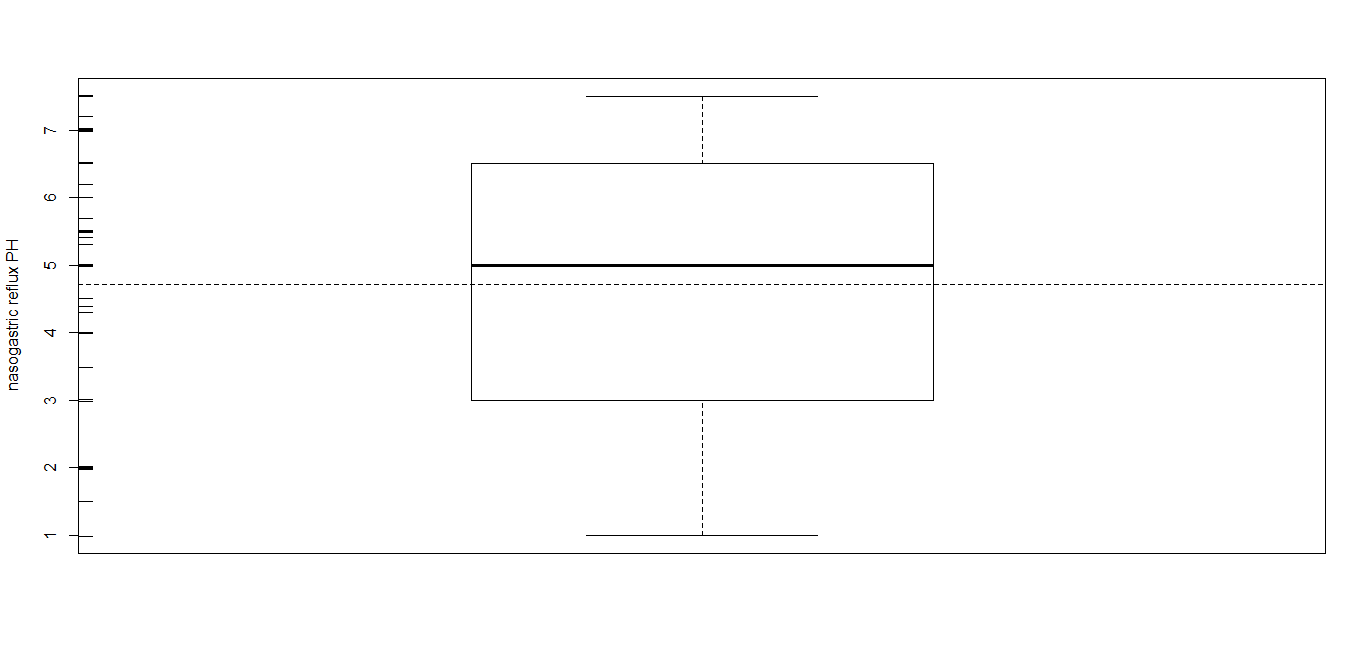
**pluse**属性：



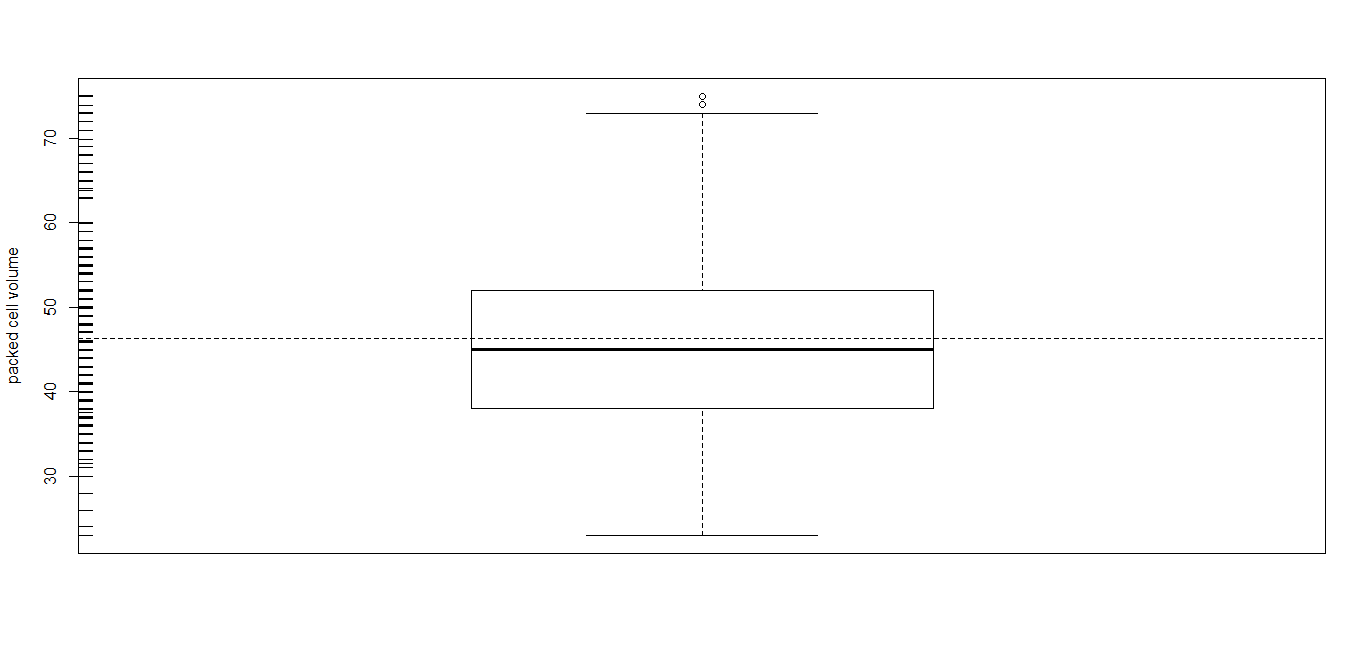
**respiratory rate**属性：



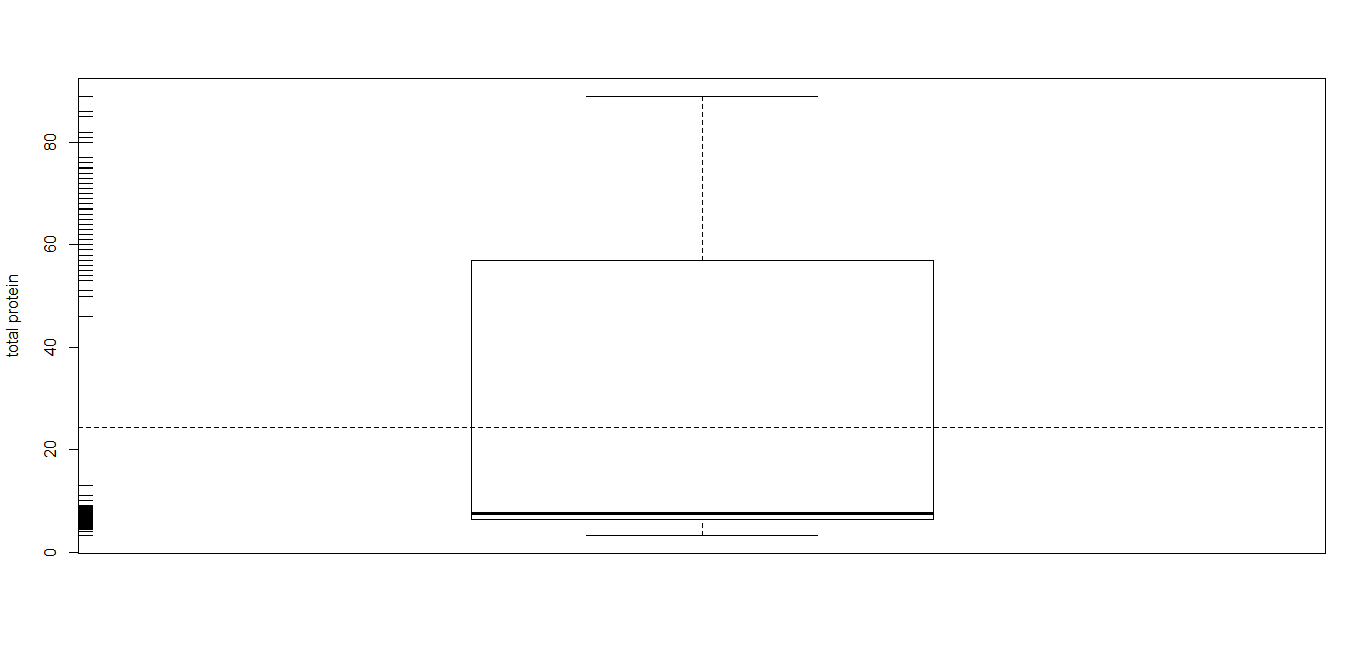
**nasogastric reflux PH**属性：



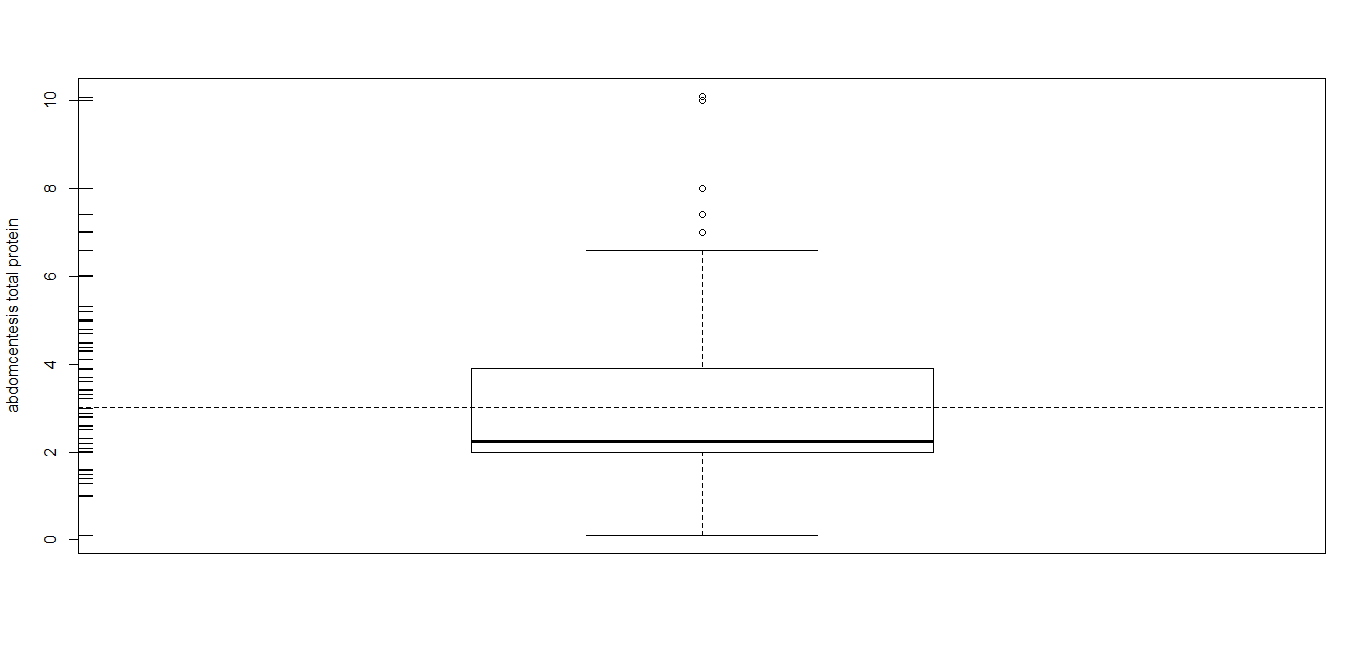
**packed cell volume**属性：



**total protein**属性：



**abdomcentesis total protein**属性：



1. 数据缺失的处理
2. 策略一：将缺失部分剔除

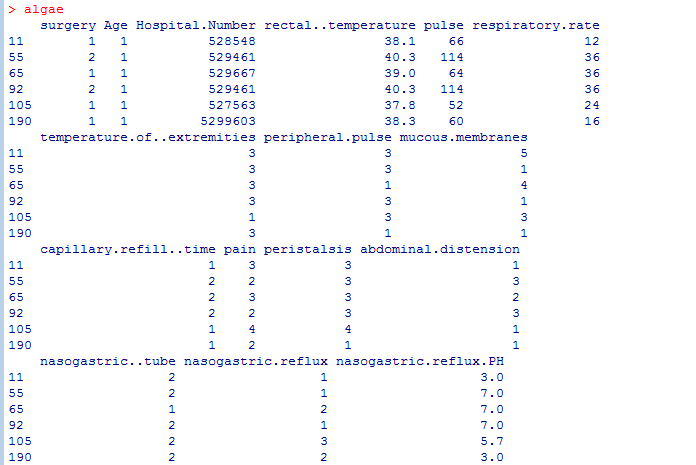
查看含有缺失值的记录的个数

含有缺失值的记录个数

剔除所有包含缺失值的记录

删除含缺失值的记录

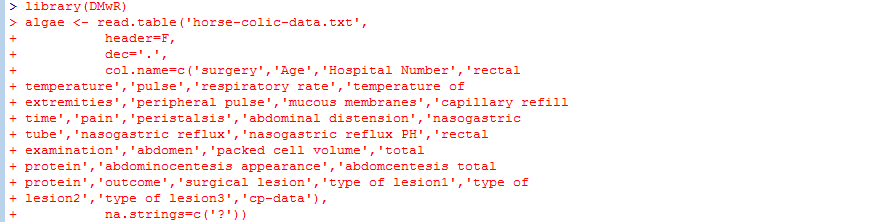
输出结果（部分）：



在本问题中，样本缺失值太多，剔除所有含缺失值记录的方法过于极端，处理后的数据已所剩无几，于是我们考虑仅剔除缺失值较多的几行数据。

剔除缺失值较多的几行记录

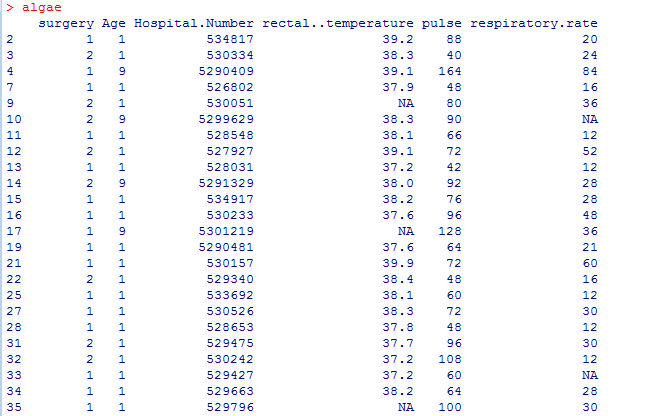
首先，由于上一步中对数据已改动，故需重新加载数据：



利用manyNAs( )函数找出缺失值个数大于列数20%的行，并剔除：

删除较多缺失值的记录

输出结果（部分）：



1. 策略二：用最高频率值来填补缺失值

重新加载数据

剔除缺失值个数大于列数20%的行

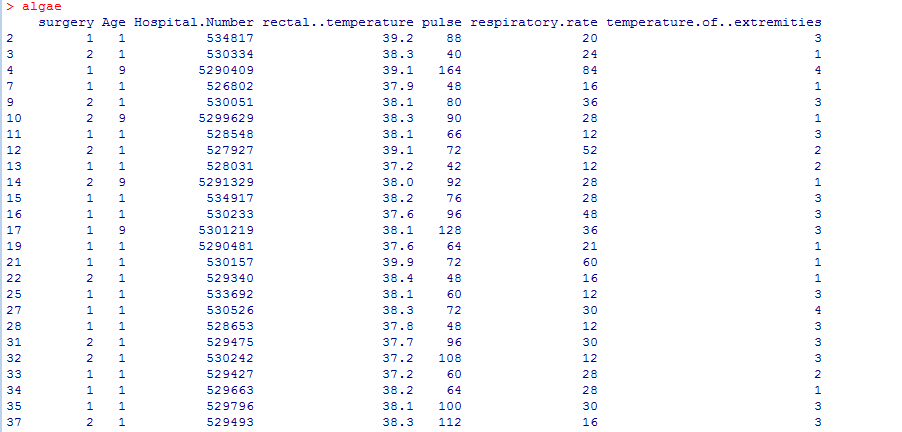
删除较多缺失值的记录

用中心趋势值填补缺失值

最高频率补语句

插件包中提供的函数centralImputation( )可以用数据的中心趋势值来填补数据集的所有缺失值。对数值型变量，该函数用中位数；对名义变量，它采用众数。

输出结果（部分）：



**四、实验感想**

本次试验让我更加进一步了解学习了数据挖掘的相关内容，期间遇到了很多问题通过小组讨论也有所解决，在接下来的学习中会进一步探索努力的。

1. **说明**
2. 本文仅详细介绍了训练集horse-colic-data.txt的实验步骤，由于测试集horse-colic-test.txt的处理过程与之类似，就不赘述，但结果保存在result/test文件中，供查看。
3. 参考文献

[1]Lurs Torgo著;李洪成等译.数据挖掘与R语言．机械工业出版社．2013.