## 作业一 马的疝病的分析报告

**姓名：刘艳**

**学号：2620160026**

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. 数据说明

**下载数据:** [地址](http://archive.ics.uci.edu/ml/datasets/Horse+Colic)

共368个样本，27个特征。关于特征的详细说明见下载链接。

3. 数据分析要求

## 3.1 数据可视化和摘要

3.1.1数据摘要

* 对标称属性，给出每个可能取值的频数，
* 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。

3.1.2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

## 3.2 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。

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
* 通过数据对象之间的相似性来填补缺失值

处理后，可视化地对比新旧数据集。

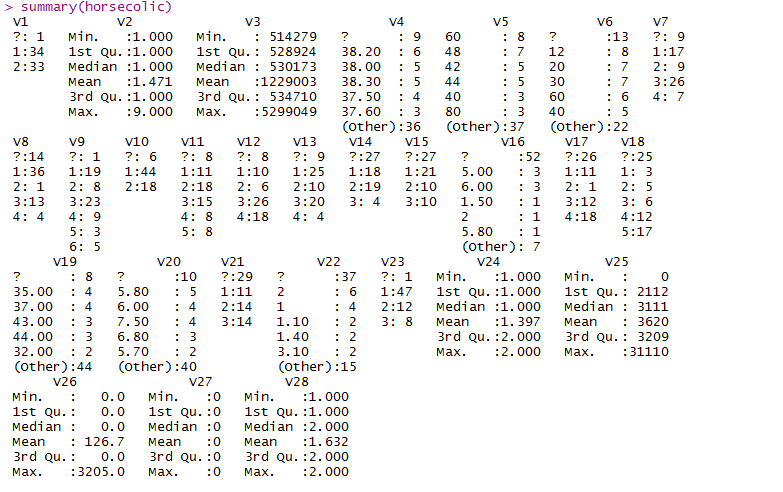
4. 实验环境及语言

在RStudio的环境下，使用R语言对数据进行预处理

5.分析过程

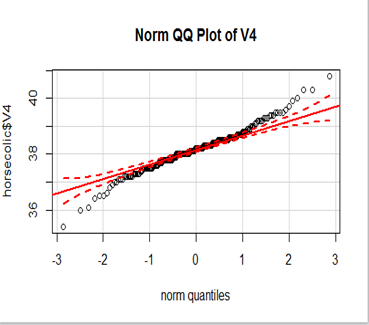
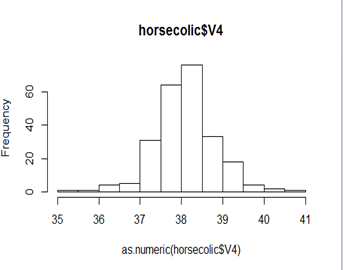
## 5.1数据摘要

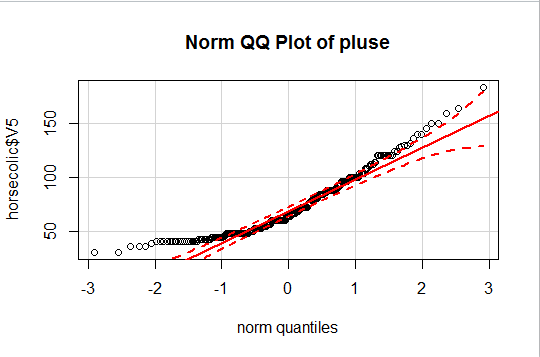
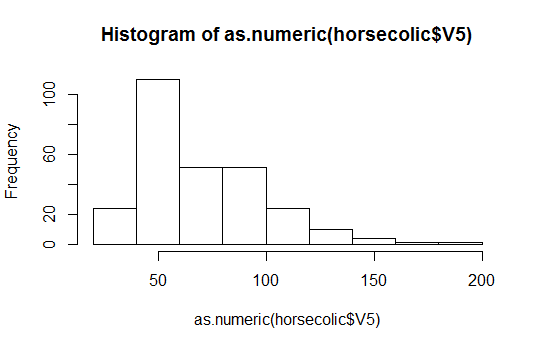
获取数据的描述性统计摘要是获取数据统计特性的一个重要方法。我们通过R语言的Summary函数对数据摘要进行分析， 这个函数给出了数据的统计特征概括，对于标称变量，他给出了每个取值变量的频数。对于数值型变量， R为我们提供了四分之一位数、中位数、均值、四分之三位数以及极值等一系列信息。所得结果如下：

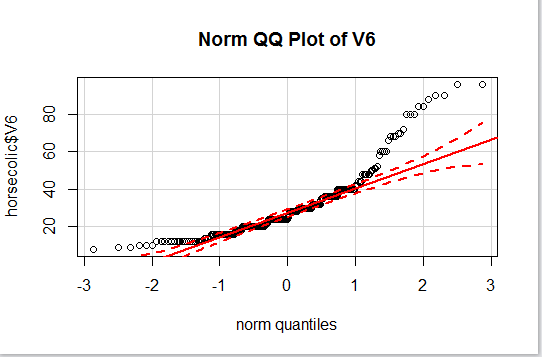
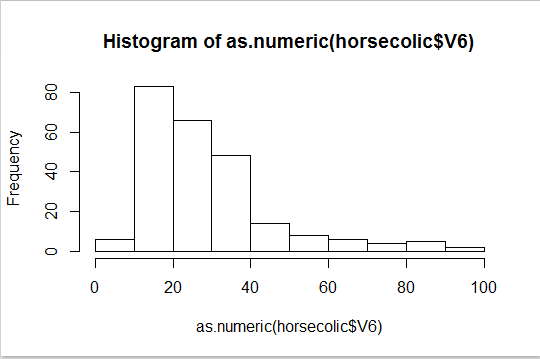


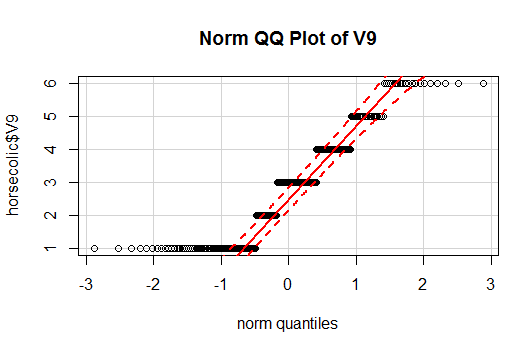
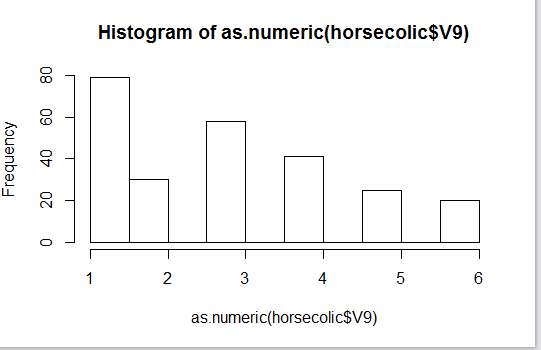
## 5.2数据可视化

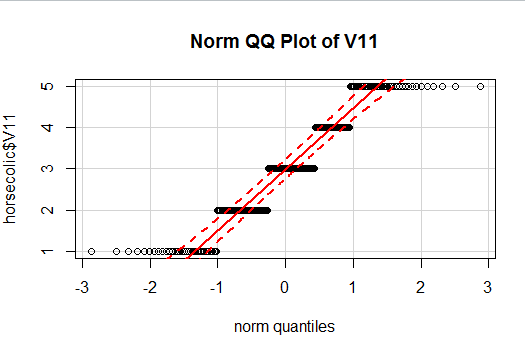
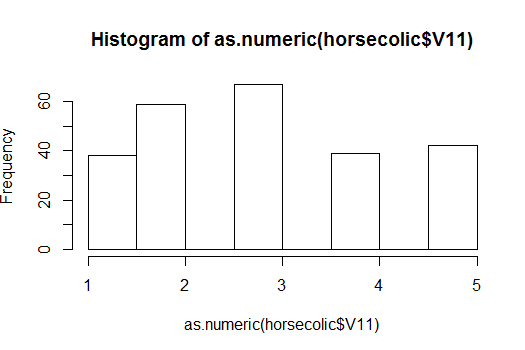
（1）数据的直方图和 QQ图如下（v4，v5，v6，v7，v9，v11，根据数据描述对一些数据不做直方图处理）。





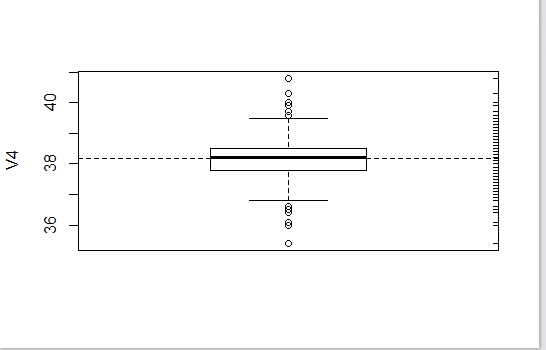


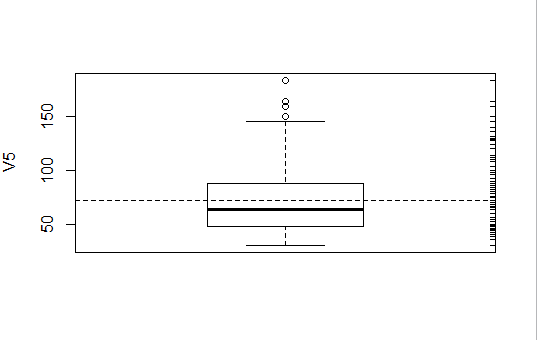


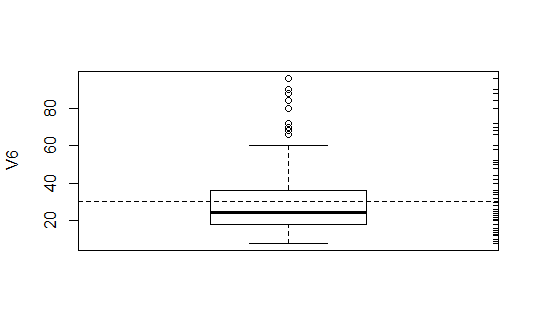


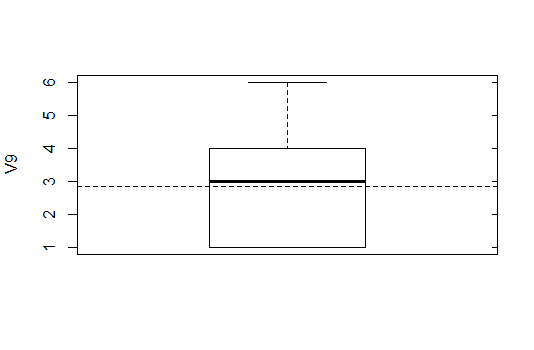
以上绘制的直方图横轴是其分布区间，纵轴是其频数； QQ图中，红色实线为其QQ线， 虚线为95%置信度的置信区间。 结果表明直方图显示变量mxPH的分布非常接近正态分布，它的值大多数都集中在变量的均值附近； QQ图绘制了变量值与正态分布的理论分位数的散点图，同时他给出正态分布的95%的置信区间的带状图， 除去有几个小的值据观察明显在95%置信区间之外，基本服从正态分布。

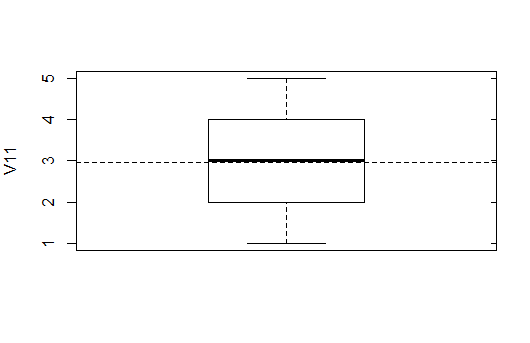
1. 数据的盒图如下：





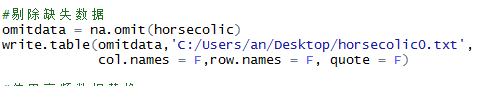




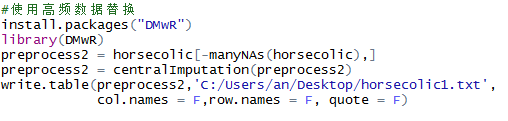


5.3缺失数据处理

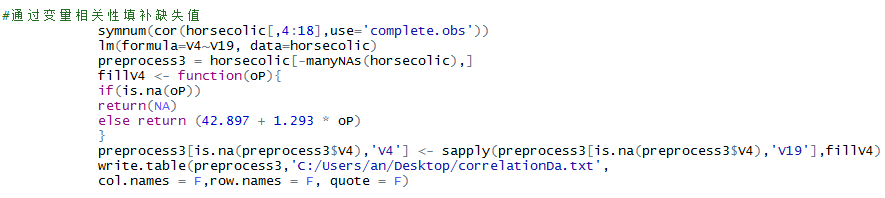
（1）剔除缺失部分数据

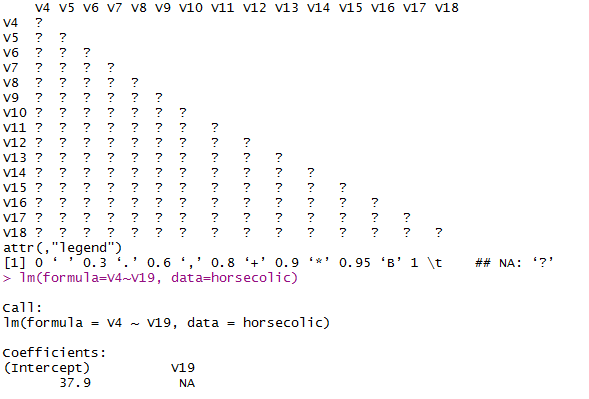


（2）利用最高频率值来填补缺失值

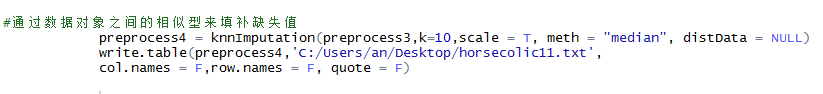


（3）通过属性的相关性来填补缺失值



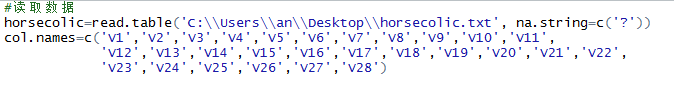


（4）通过数据对象之间的相似性来填补缺失值



6程序描述

6.1读取



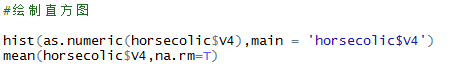
6.2摘要

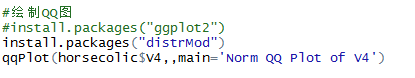
75D5N0~OT{_}MPW@2]O4[B6

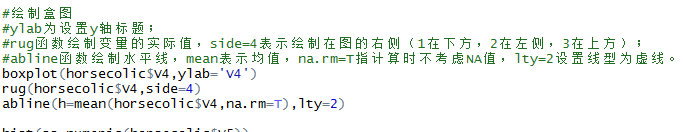
6.3可视化

@O5{@)LQDQ}(~81~O}F1VQS

6.4绘图







## 6.5处理缺失如“5.3”中所示

7处理后的数据

