

分析过程报告

本报告尽可能详尽地给出数据分析过程，在给出的368个样本，27个特征中，根据其特征详细说明，确定了各个特征属性，其中数值属性有7个，分别为：rectal temperature，pulse，respiratory rate，nasogastric reflux PH，packed cell volume，total protein，abdomcentesis total protein；标称属性16个，分别为surgerynan，Age，temperature of extremities，peripheral pulse，mucous membranes，capillary refill time，pain，peristalsis，abdominal distension，nasogastric tube，nasogastric reflux，rectal examination–feces，abdomen，abdominocentesis appearance，outcome和cp\_data；其余4个为固定代码，不予处理。

一、问题的描述

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，将该数据集按要求予以处理。

二、数据摘要和可视化

（一）数据摘要。

1、对于16个标称属性，给出每个可能取值的频数，相关预处理后的数据如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Age | | | | | |
|  | | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 | 1 | 340 | 92.4 | 92.4 | 92.4 |
| 9 | 28 | 7.6 | 7.6 | 100.0 |
| 合计 | 368 | 100.0 | 100.0 |  |
| peripheral pulse | | | | | |
|  | | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 | 1 | 151 | 41.0 | 41.0 | 41.0 |
| 2 | 6 | 1.6 | 1.6 | 42.7 |
| 3 | 116 | 31.5 | 31.5 | 74.2 |
| 4 | 12 | 3.3 | 3.3 | 77.4 |
| nan | 83 | 22.6 | 22.6 | 100.0 |
| 合计 | 368 | 100.0 | 100.0 |  |
| surgerynan | | | | | |
|  | | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 | 1 | 214 | 58.2 | 58.2 | 58.2 |
| 2 | 152 | 41.3 | 41.3 | 99.5 |
| nan | 2 | .5 | .5 | 100.0 |
| 合计 | 368 | 100.0 | 100.0 |  |
| temperature of extremities | | | | | |
|  | | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 | 1 | 95 | 25.8 | 25.8 | 25.8 |
| 2 | 39 | 10.6 | 10.6 | 36.4 |
| 3 | 135 | 36.7 | 36.7 | 73.1 |
| 4 | 34 | 9.2 | 9.2 | 82.3 |
| nan | 65 | 17.7 | 17.7 | 100.0 |
| 合计 | 368 | 100.0 | 100.0 |  |
| mucous membranes | | | | | |
|  | | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 | 1 | 98 | 26.6 | 26.6 | 26.6 |
| 2 | 38 | 10.3 | 10.3 | 37.0 |
| 3 | 81 | 22.0 | 22.0 | 59.0 |
| 4 | 50 | 13.6 | 13.6 | 72.6 |
| 5 | 28 | 7.6 | 7.6 | 80.2 |
| 6 | 25 | 6.8 | 6.8 | 87.0 |
| nan | 48 | 13.0 | 13.0 | 100.0 |
| 合计 | 368 | 100.0 | 100.0 |  |
| pain | | | | | |
|  | | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 | 1 | 49 | 13.3 | 13.3 | 13.3 |
| 2 | 77 | 20.9 | 20.9 | 34.2 |
| 3 | 82 | 22.3 | 22.3 | 56.5 |
| 4 | 47 | 12.8 | 12.8 | 69.3 |
| 5 | 50 | 13.6 | 13.6 | 82.9 |
| nan | 63 | 17.1 | 17.1 | 100.0 |
| 合计 | 368 | 100.0 | 100.0 |  |
| capillary refill time | | | | | |
|  | | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 | 1 | 232 | 63.0 | 63.0 | 63.0 |
| 2 | 96 | 26.1 | 26.1 | 89.1 |
| 3 | 2 | .5 | .5 | 89.7 |
| nan | 38 | 10.3 | 10.3 | 100.0 |
| 合计 | 368 | 100.0 | 100.0 |  |
| abdominal distension | | | | | |
|  | | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 | 1 | 101 | 27.4 | 27.4 | 27.4 |
| 2 | 75 | 20.4 | 20.4 | 47.8 |
| 3 | 85 | 23.1 | 23.1 | 70.9 |
| 4 | 42 | 11.4 | 11.4 | 82.3 |
| nan | 65 | 17.7 | 17.7 | 100.0 |
| 合计 | 368 | 100.0 | 100.0 |  |
| peristalsis | | | | | |
|  | | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 | 1 | 49 | 13.3 | 13.3 | 13.3 |
| 2 | 22 | 6.0 | 6.0 | 19.3 |
| 3 | 154 | 41.8 | 41.8 | 61.1 |
| 4 | 91 | 24.7 | 24.7 | 85.9 |
| nan | 52 | 14.1 | 14.1 | 100.0 |
| 合计 | 368 | 100.0 | 100.0 |  |
| nasogastric reflux | | | | | |
|  | | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 | 1 | 141 | 38.3 | 38.3 | 38.3 |
| 2 | 45 | 12.2 | 12.2 | 50.5 |
| 3 | 49 | 13.3 | 13.3 | 63.9 |
| nan | 133 | 36.1 | 36.1 | 100.0 |
| 合计 | 368 | 100.0 | 100.0 |  |
| nasogastric tube | | | | | |
|  | | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 | 1 | 89 | 24.2 | 24.2 | 24.2 |
| 2 | 121 | 32.9 | 32.9 | 57.1 |
| 3 | 27 | 7.3 | 7.3 | 64.4 |
| nan | 131 | 35.6 | 35.6 | 100.0 |
| 合计 | 368 | 100.0 | 100.0 |  |
| rectal examination - feces | | | | | |
|  | | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 | 1 | 68 | 18.5 | 18.5 | 18.5 |
| 2 | 14 | 3.8 | 3.8 | 22.3 |
| 3 | 61 | 16.6 | 16.6 | 38.9 |
| 4 | 97 | 26.4 | 26.4 | 65.2 |
| nan | 128 | 34.8 | 34.8 | 100.0 |
| 合计 | 368 | 100.0 | 100.0 |  |
| abdomen | | | | | |
|  | | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 | 1 | 31 | 8.4 | 8.4 | 8.4 |
| 2 | 24 | 6.5 | 6.5 | 14.9 |
| 3 | 19 | 5.2 | 5.2 | 20.1 |
| 4 | 55 | 14.9 | 14.9 | 35.1 |
| 5 | 96 | 26.1 | 26.1 | 61.1 |
| nan | 143 | 38.9 | 38.9 | 100.0 |
| 合计 | 368 | 100.0 | 100.0 |  |
| outcome | | | | | |
|  | | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 | 1 | 225 | 61.1 | 61.1 | 61.1 |
| 2 | 89 | 24.2 | 24.2 | 85.3 |
| 3 | 52 | 14.1 | 14.1 | 99.5 |
| nan | 2 | .5 | .5 | 100.0 |
| 合计 | 368 | 100.0 | 100.0 |  |
| abdominocentesis appearance | | | | | |
|  | | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 | 1 | 52 | 14.1 | 14.1 | 14.1 |
| 2 | 62 | 16.8 | 16.8 | 31.0 |
| 3 | 60 | 16.3 | 16.3 | 47.3 |
| nan | 194 | 52.7 | 52.7 | 100.0 |
| 合计 | 368 | 100.0 | 100.0 |  |
| cp\_ data | | | | | |
|  | | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 | 1 | 124 | 33.7 | 33.7 | 33.7 |
| 2 | 244 | 66.3 | 66.3 | 100.0 |
| 合计 | 368 | 100.0 | 100.0 |  |

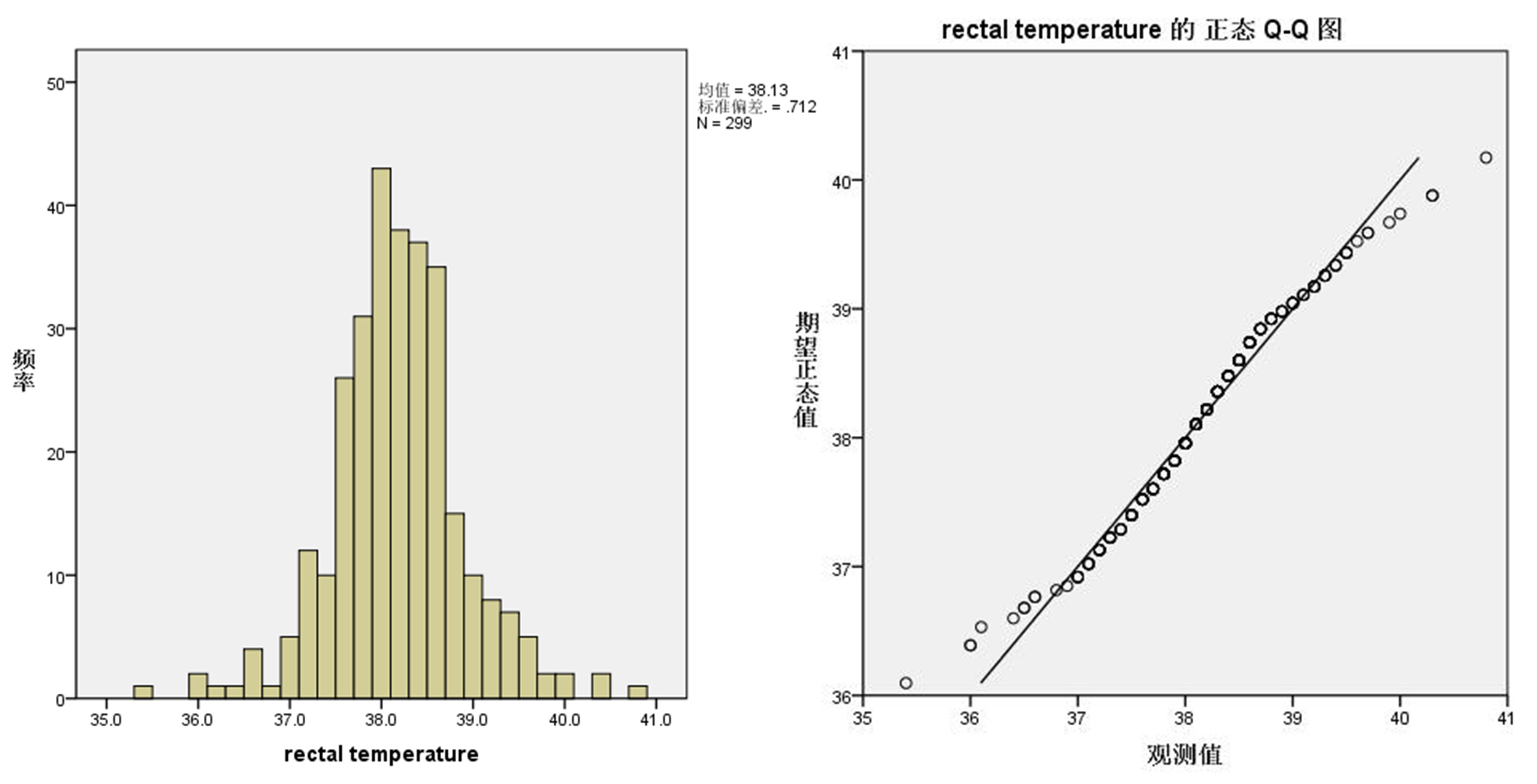
2、对于7个数值属性，给出最大、最小、均值、中位数、四分位数及缺失值的个数，相关预处理后的数据如下：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | rectal temper-ature | pulse | Respirat-ory rate | Nasogastr-ic reflux PH | packed cell volume | total protein | abdomcentesis total protein |
| N | 有效 | 299 | 342 | 297 | 69 | 331 | 325 | 133 |
| 缺失 | 69 | 26 | 71 | 299 | 37 | 43 | 235 |
| 均值 | | 38.134 | 70.76 | 30.52 | 4.962 | 45.66 | 24.771 | 2.948 |
| 中位数 | | 38.100 | 60.00 | 28.00 | 5.400 | 44.00 | 7.500 | 2.100 |
| 最小值 | | 35.4 | 30 | 8 | 1.0 | 4 | 3.3 | .1 |
| 最大值 | | 40.8 | 184 | 96 | 8.5 | 75 | 89.0 | 10.1 |
| 百分位数 | 25 | 37.800 | 48.00 | 18.00 | 3.250 | 37.00 | 6.500 | 1.900 |
| 50 | 38.100 | 60.00 | 28.00 | 5.400 | 44.00 | 7.500 | 2.100 |
| 75 | 38.500 | 88.00 | 36.00 | 6.500 | 52.00 | 58.000 | 3.900 |

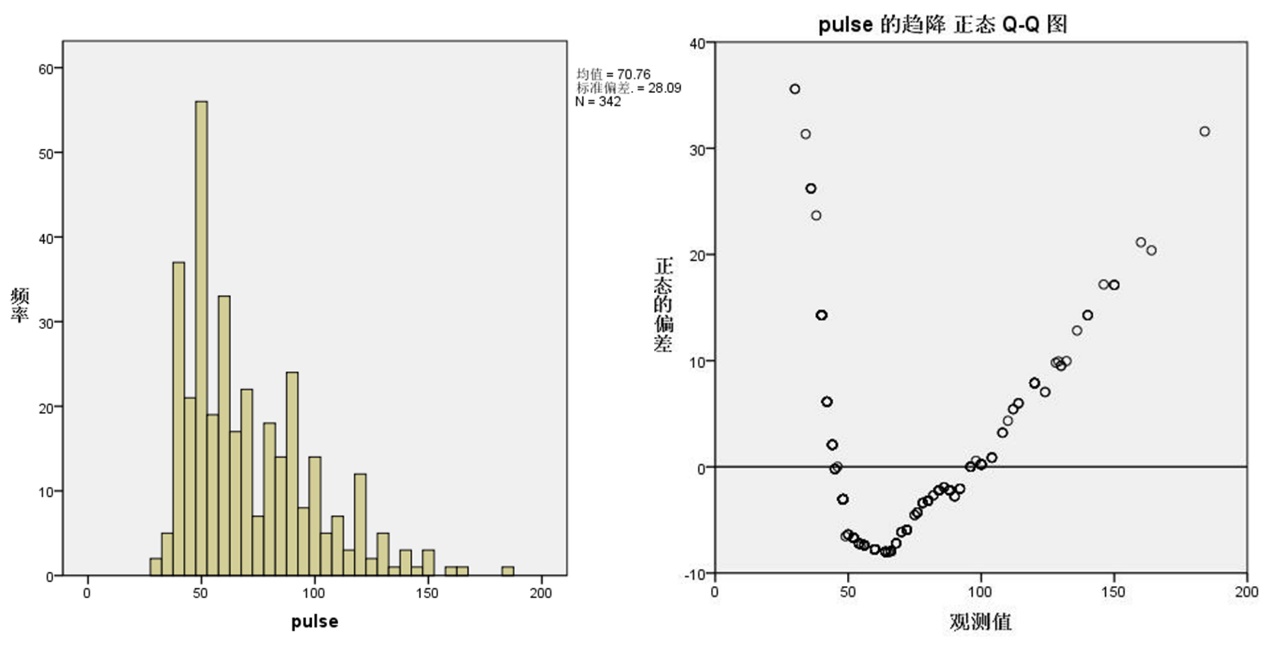
（二）数据的可视化。针对数值属性：

1、绘制直方图，用qq图检验其分布是否为正态分布。

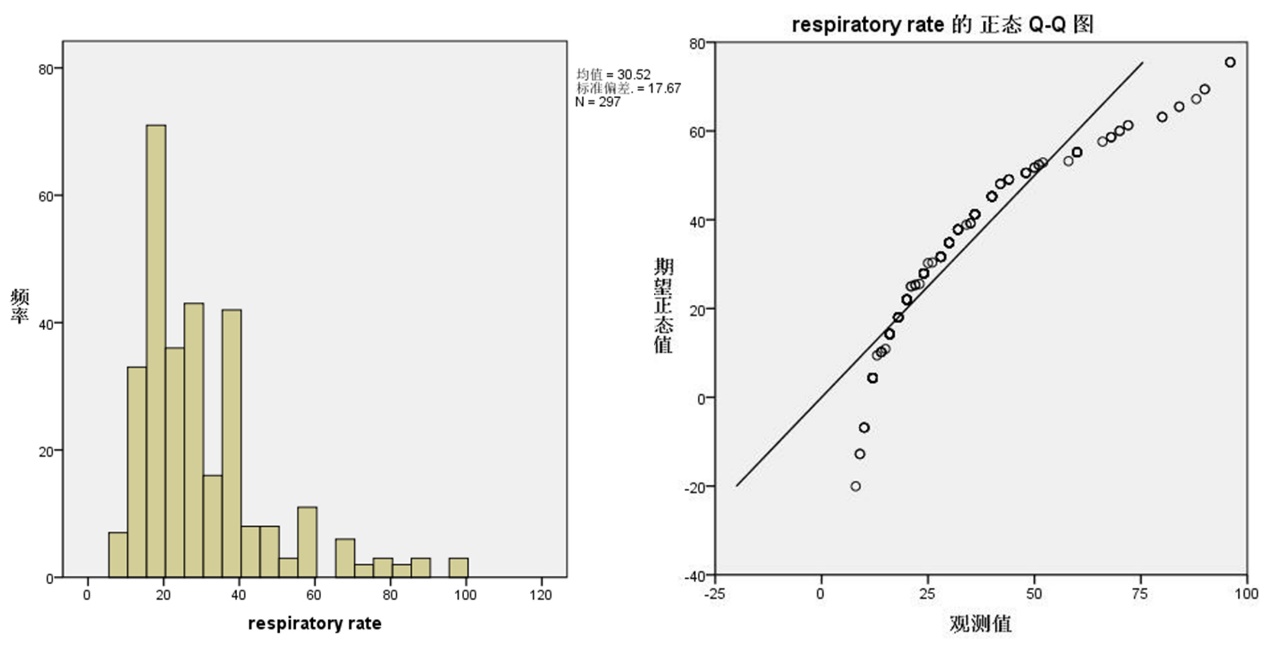
（1）rectal temperature的直方图与验证QQ图



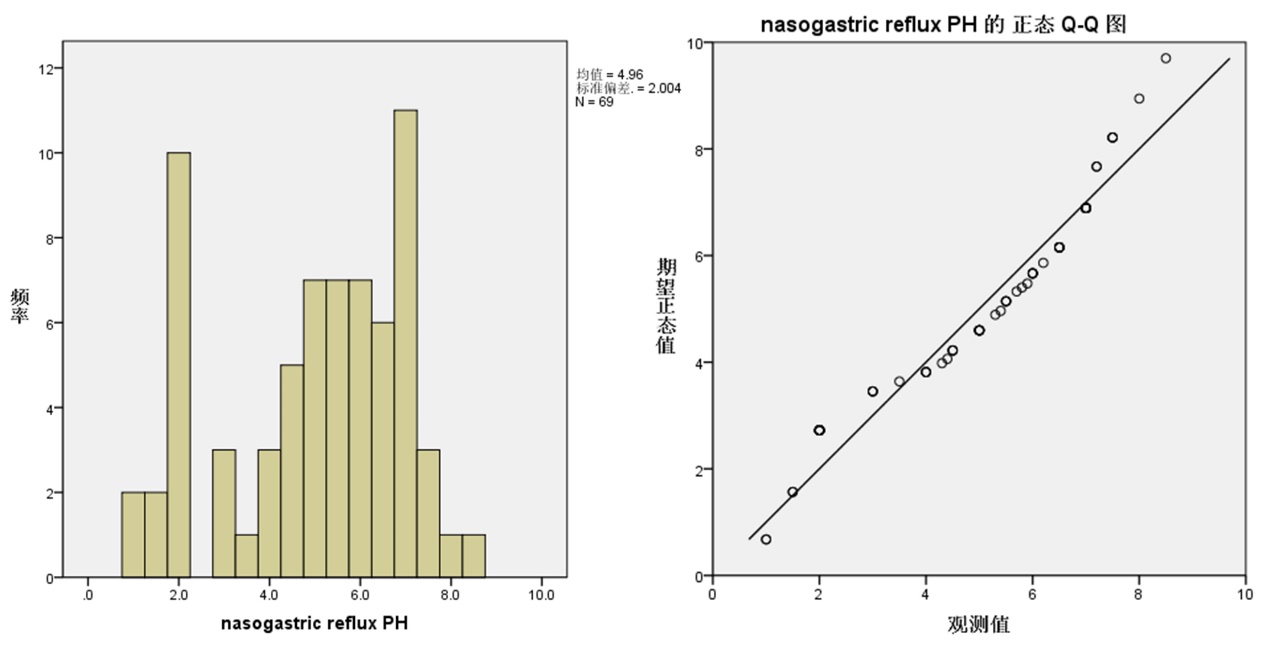
（2）pulse的直方图与验证QQ图



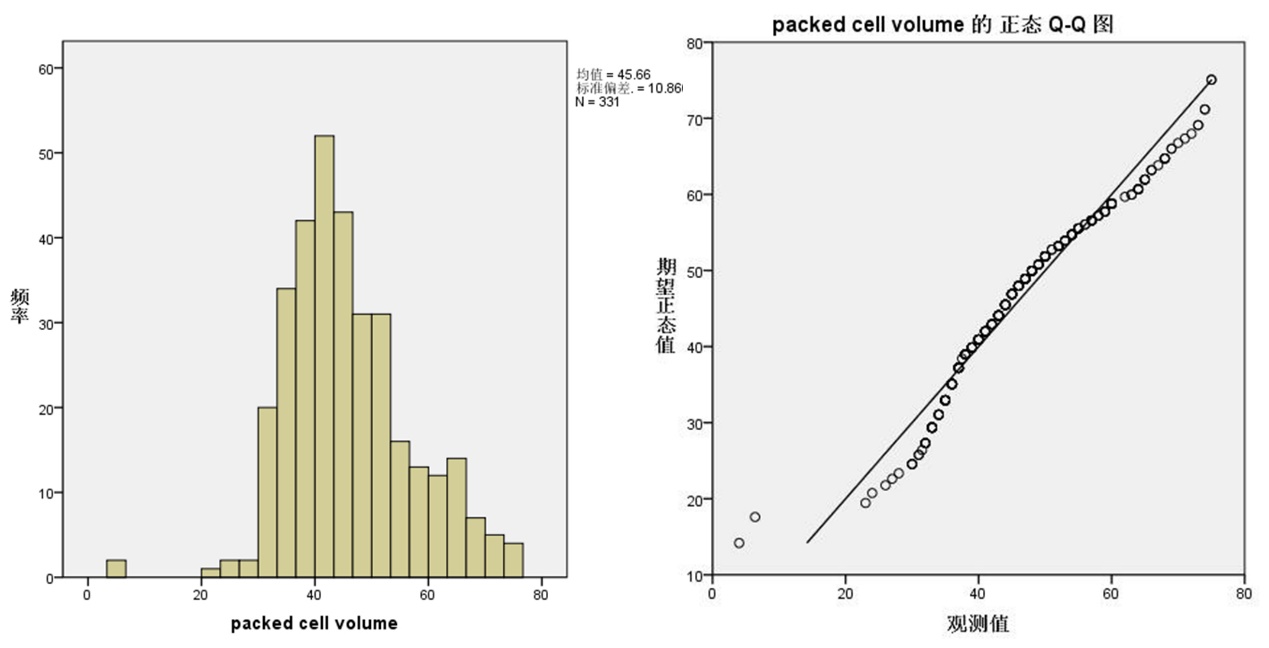
（3）respiratory rate的直方图与验证QQ图



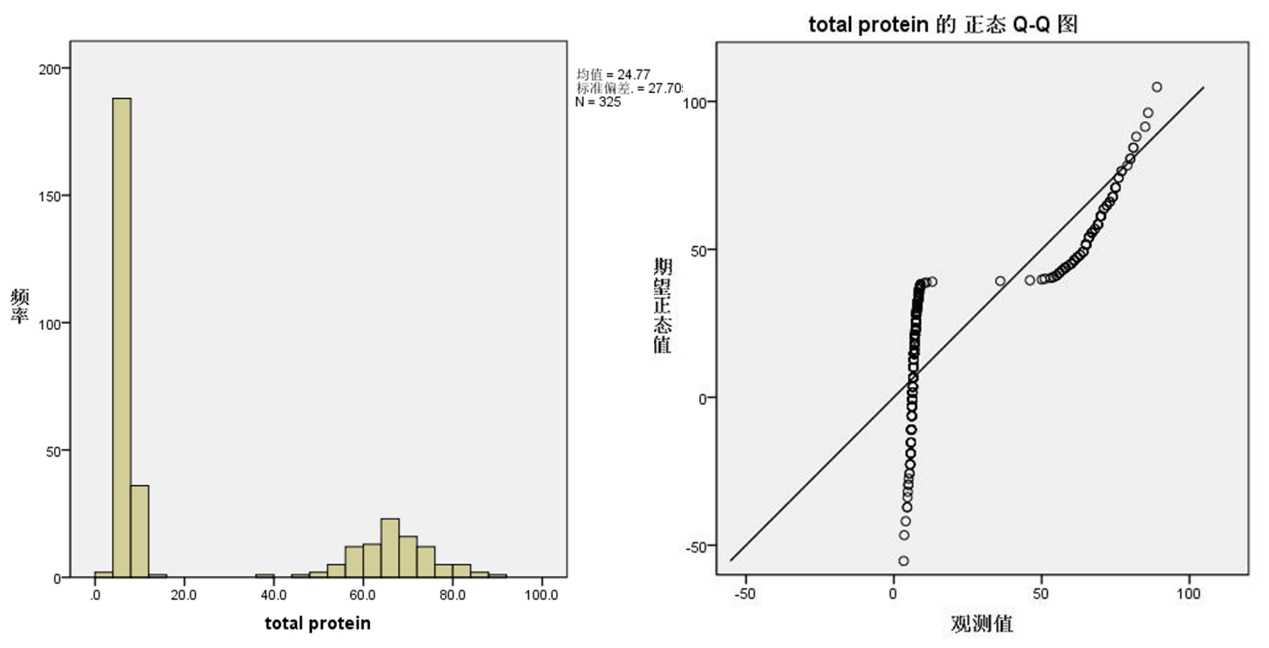
（4）nasogastric reflux PH的直方图与验证QQ图



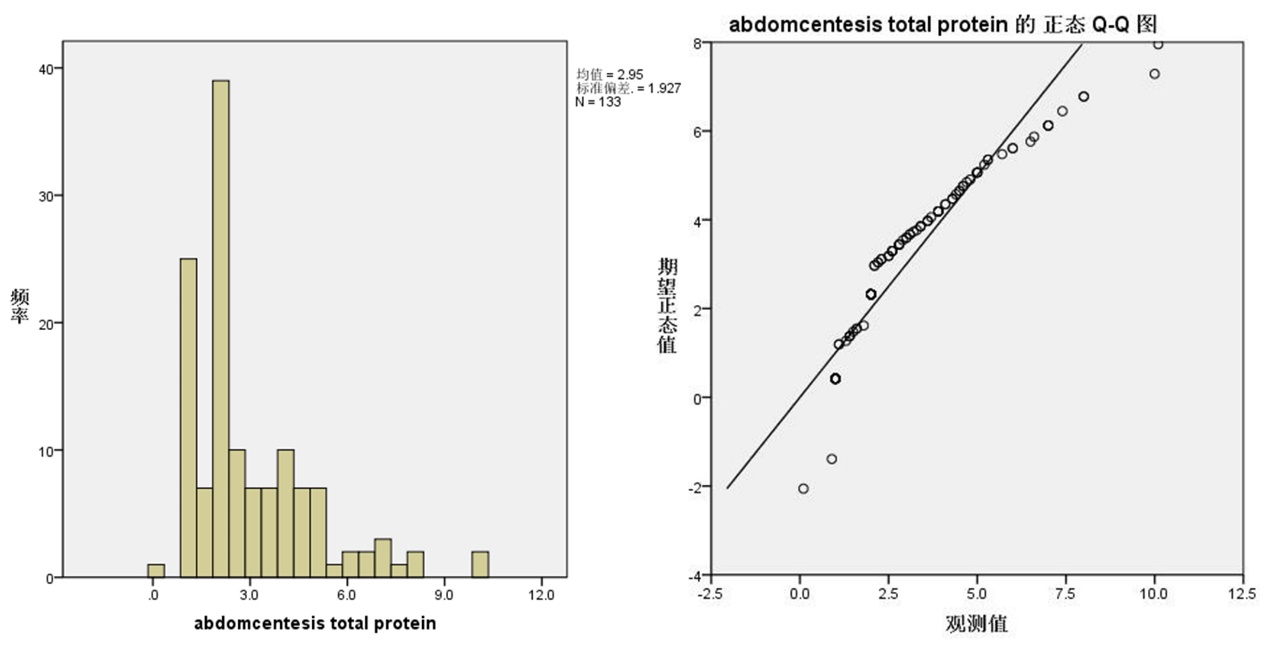
（5）packed cell volume的直方图与验证QQ图



（6）total protein的直方图与验证QQ图



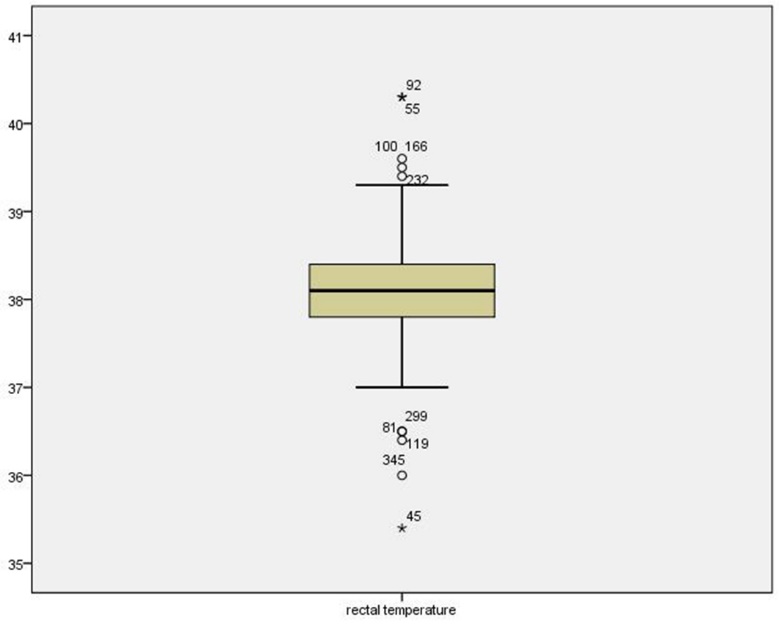
（7）abdomcentesis total protein的直方图与验证QQ图



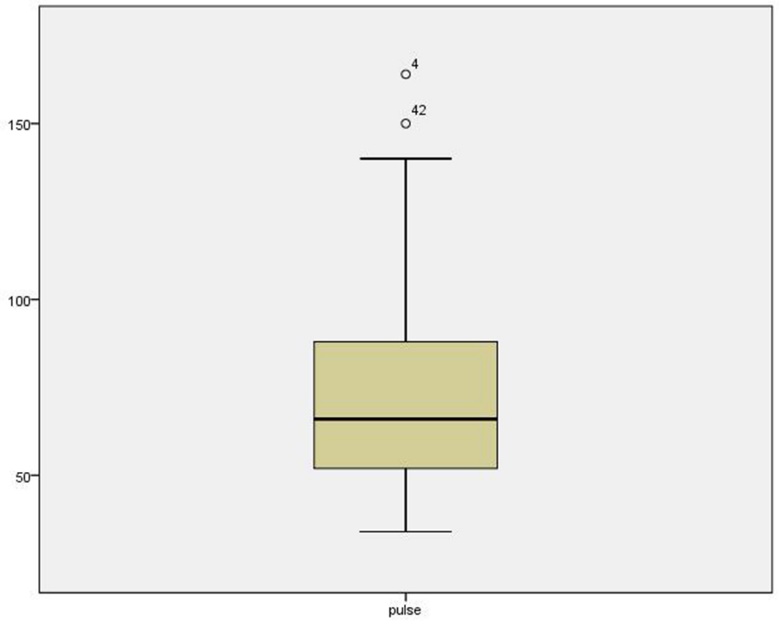
对比可知，数值属性的各个数据点分布服从正态分布。

2、绘制盒图，对离群值进行识别

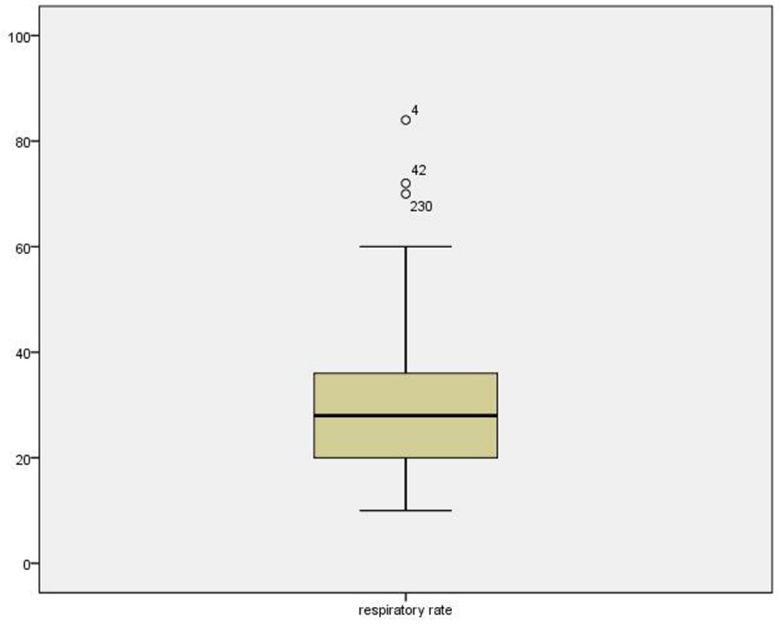
（1）rectal temperature的盒图



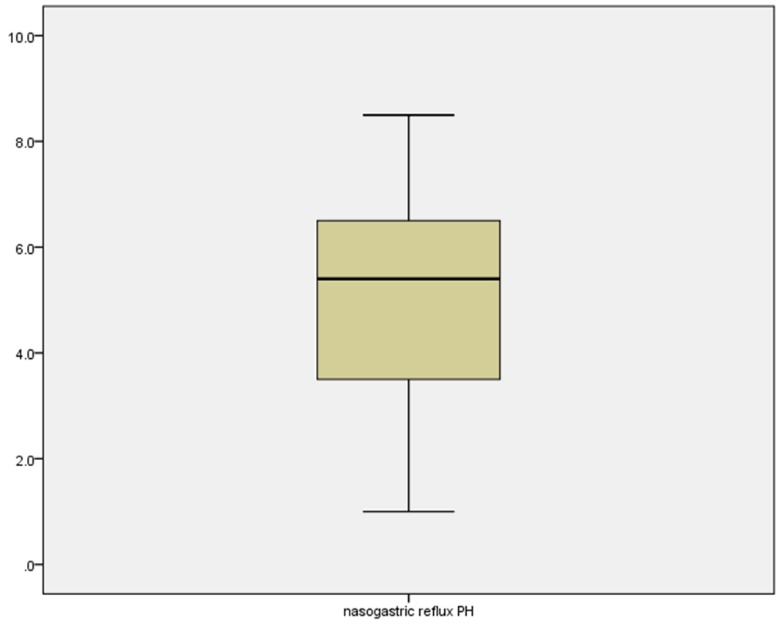
（2）pulse的盒图



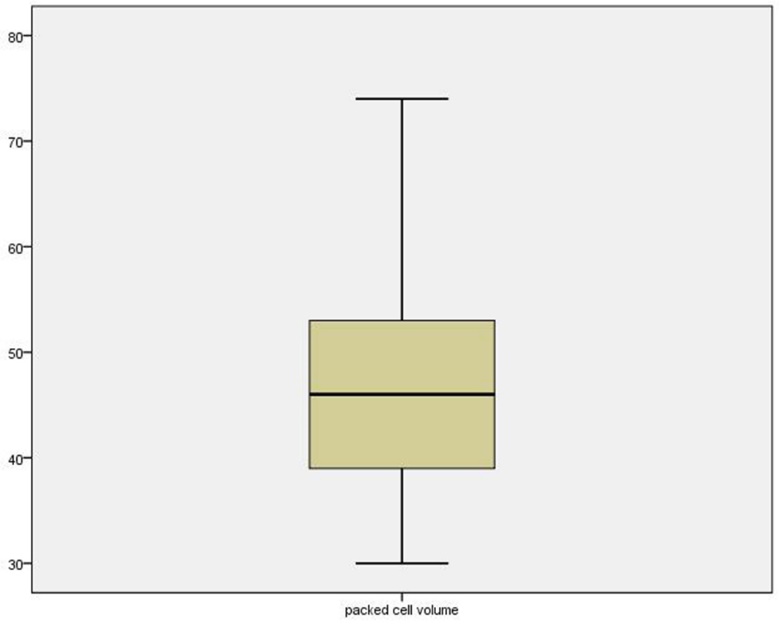
（3）respiratory rate的盒图



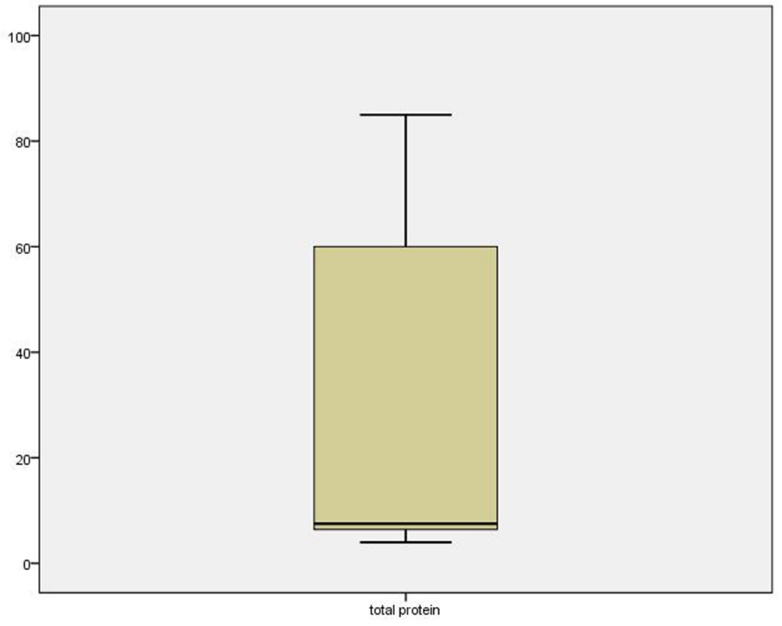
（4）nasogastric reflux PH的盒图



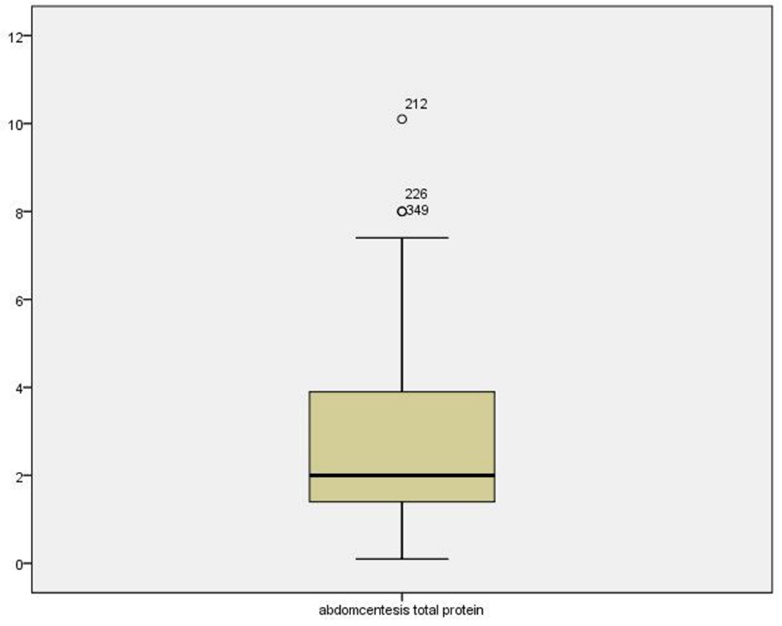
（5）packed cell volume的盒图



（6）total protein的盒图



（7）abdomcentesis total protein的盒图



三、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:

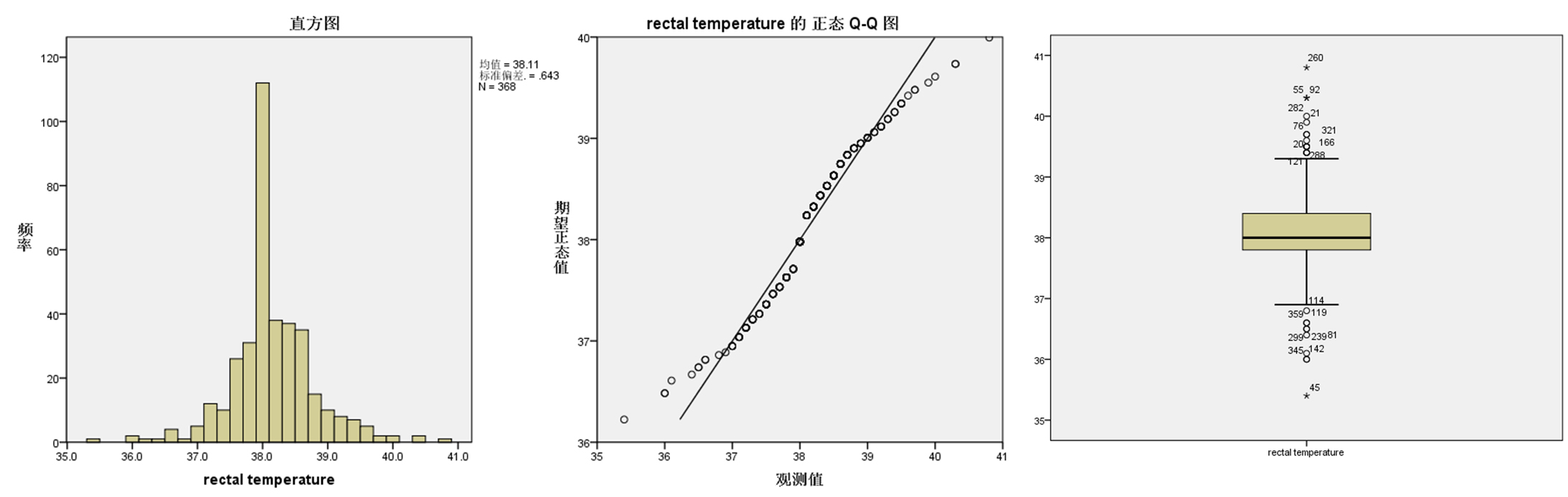
（一）将缺失部分剔除

去除掉所有样本28个属性中大于等于10个为NaN值的数据进行清除，共清除了34条（12.23%）数据。在使用SPSS工具绘制图形时，对空值的处理等价于直接剔除，结果与上一部分所绘图形一致，在此不再赘述。

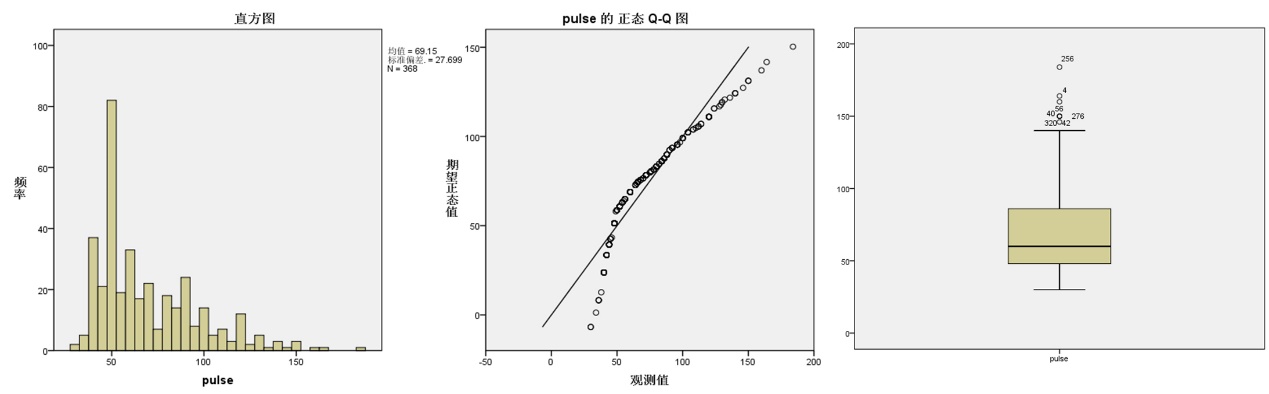
（二）用最高频率值来填补缺失值

这里的最高频率值，使用用最大频数的数据（标称属性）或者中位数（数值数据）来填补缺失值，结果如下：

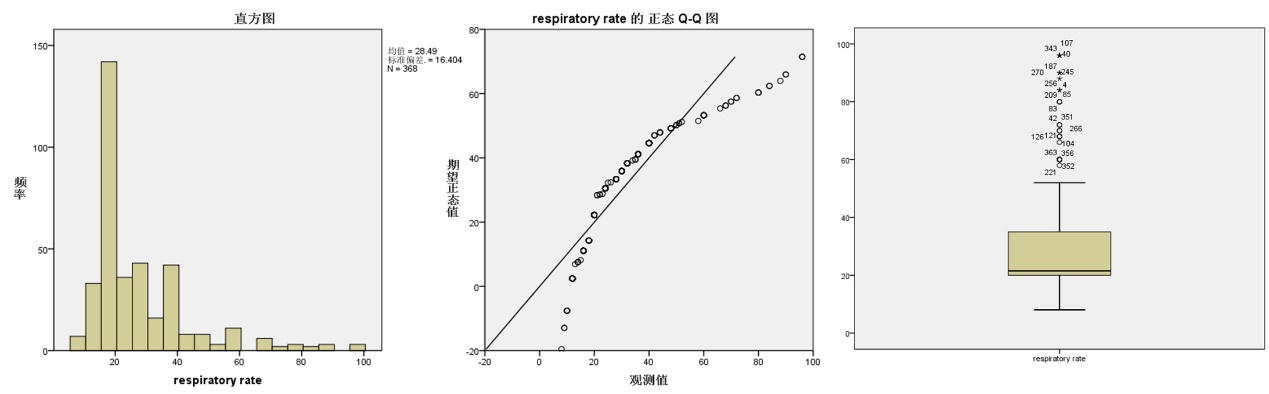
（1）rectal temperature的直方图、验证QQ图与盒图



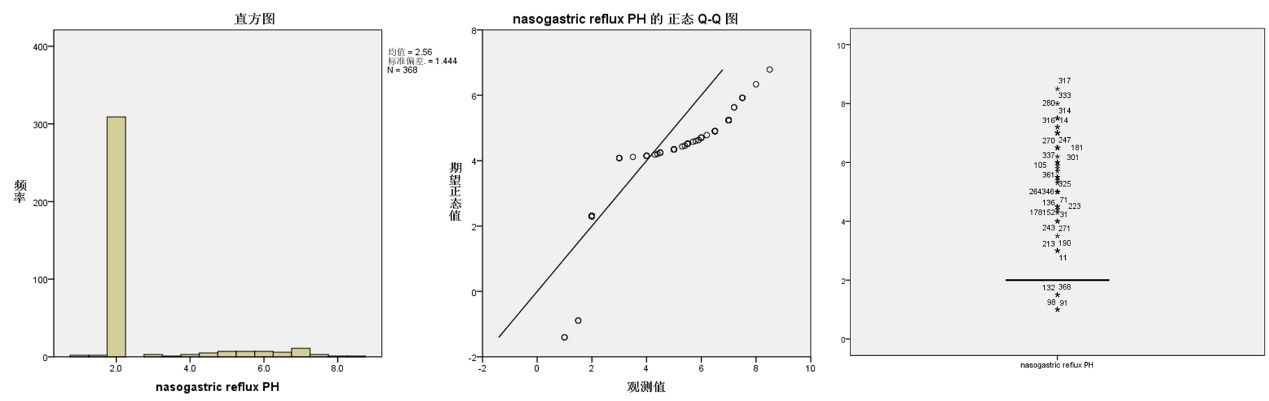
（2）pulse的直方图、验证QQ图与盒图



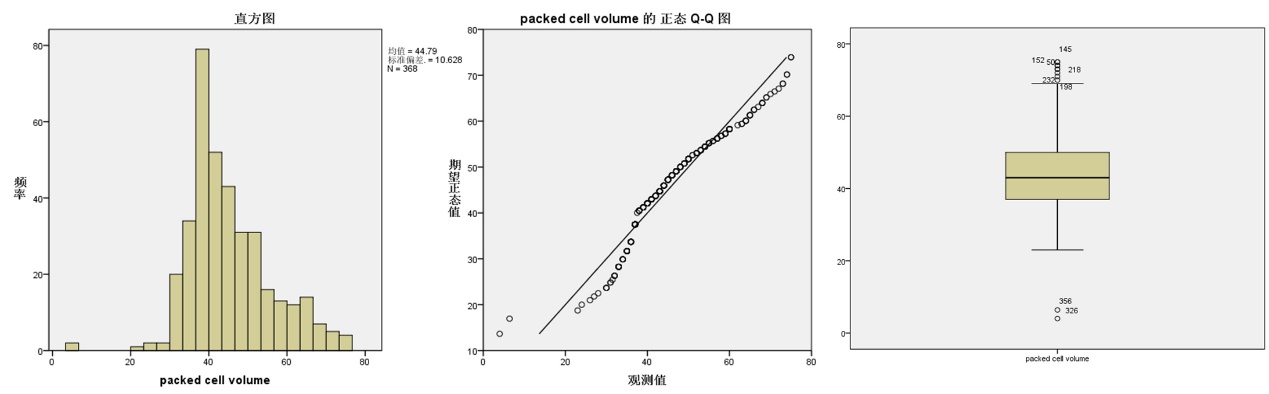
（3）respiratory rate的直方图、验证QQ图与盒图



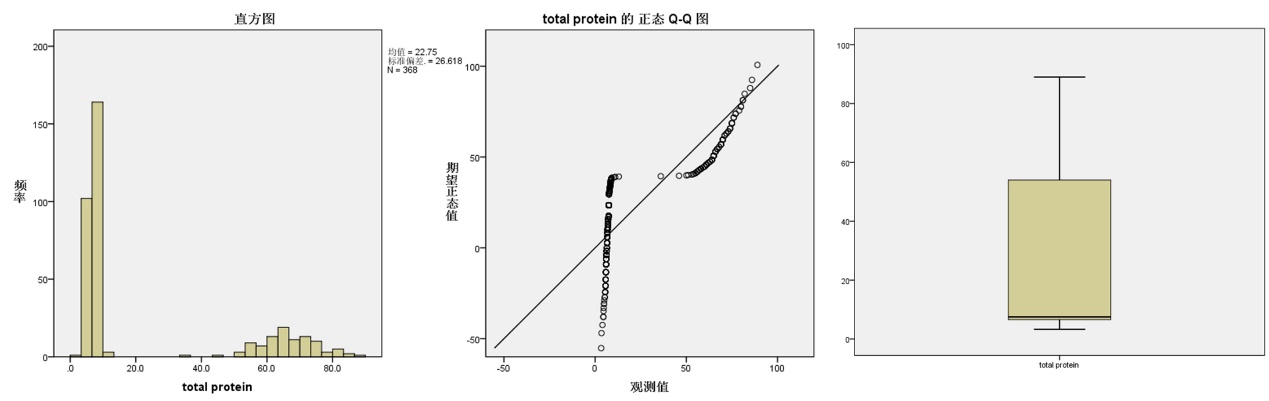
（4）nasogastric reflux PH的直方图、验证QQ图与盒图



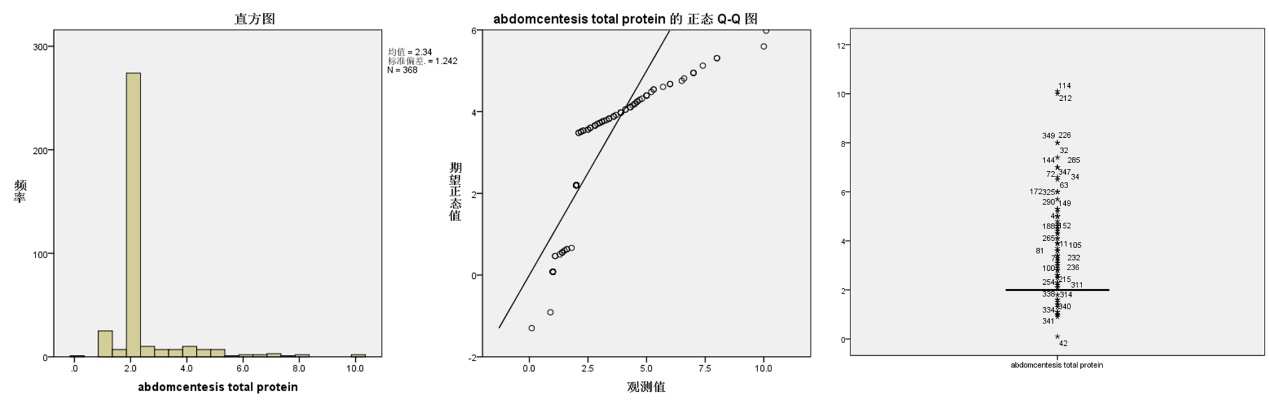
（5）packed cell volume的直方图、验证QQ图与盒图



（6）total protein的直方图、验证QQ图与盒图



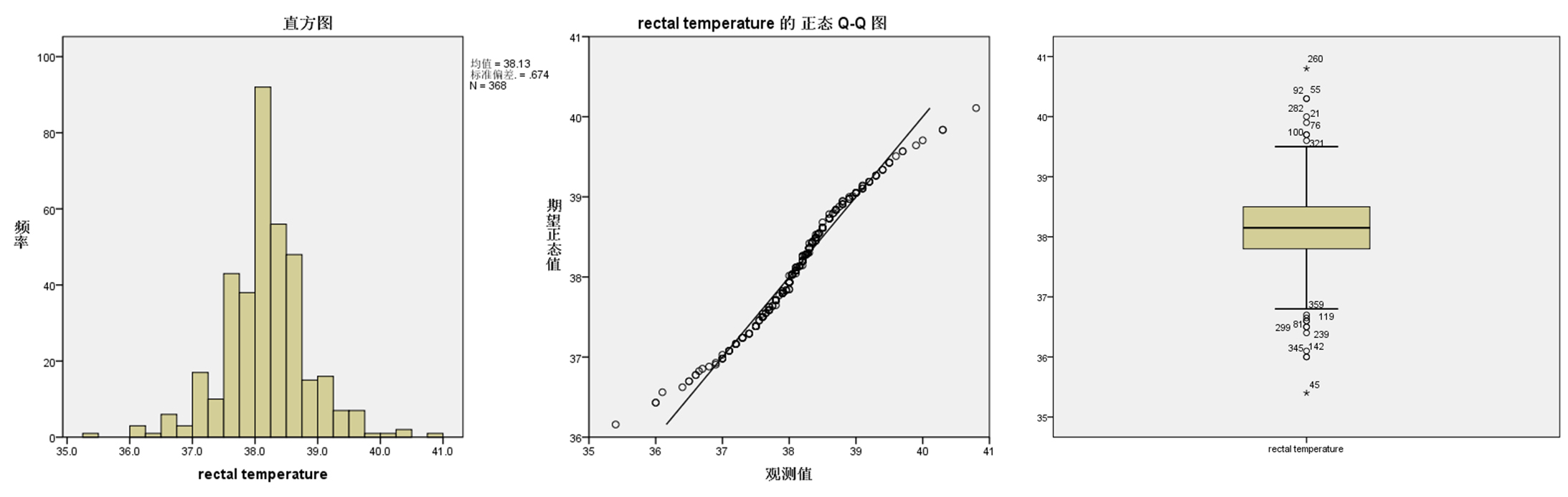
（7）abdomcentesis total protein的直方图、验证QQ图与盒图



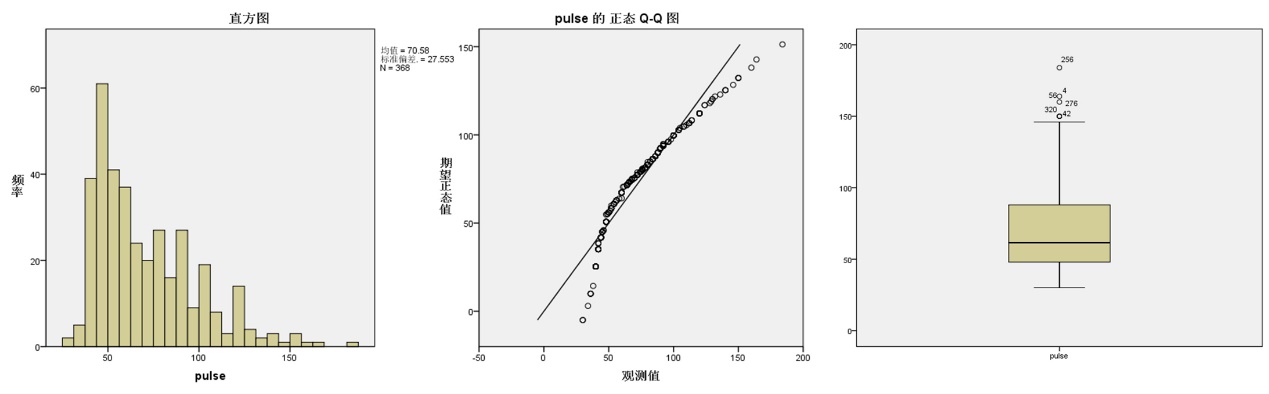
（三）通过属性的相关关系来填补缺失值

按照属性相关关系填补缺失值后，结果如下：

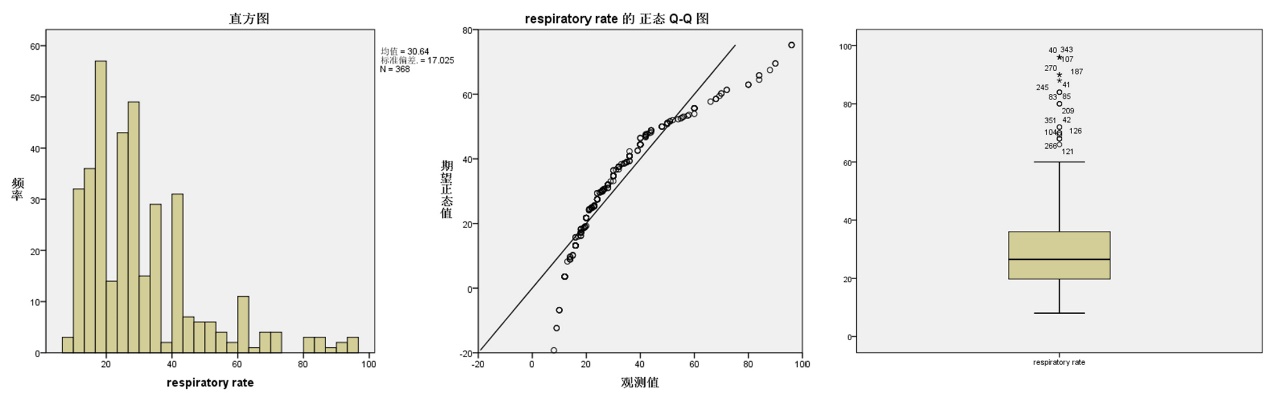
（1）rectal temperature的直方图、验证QQ图与盒图



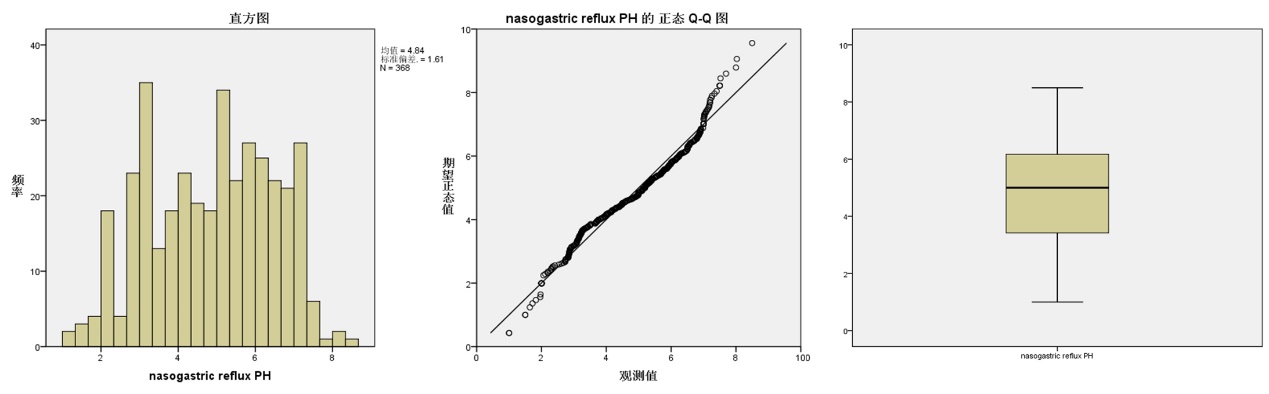
（2）pulse的直方图、验证QQ图与盒图



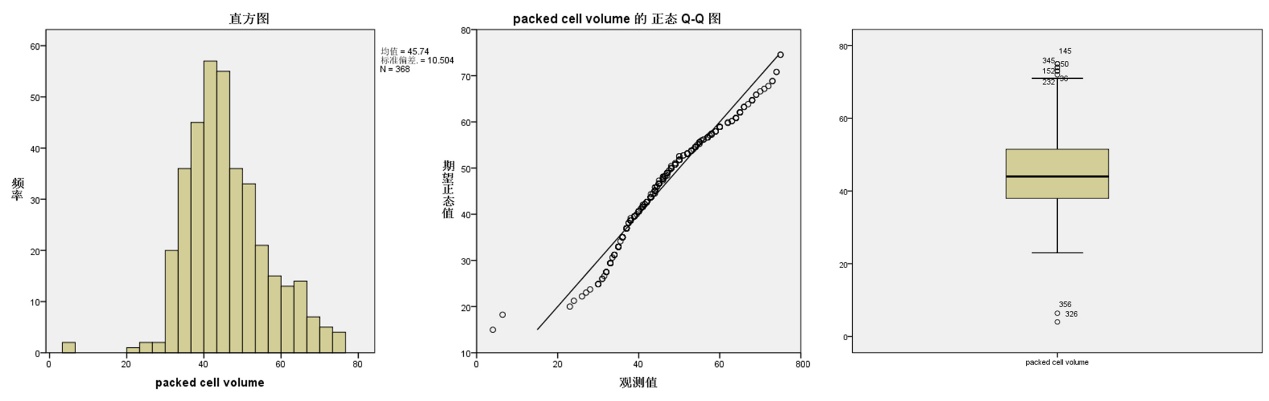
（3）respiratory rate的直方图、验证QQ图与盒图



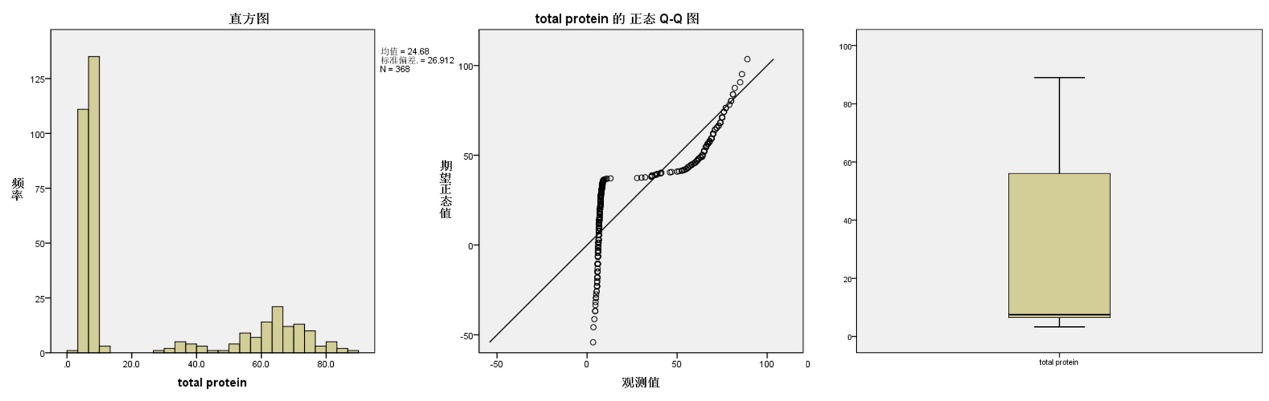
（4）nasogastric reflux PH的直方图、验证QQ图与盒图



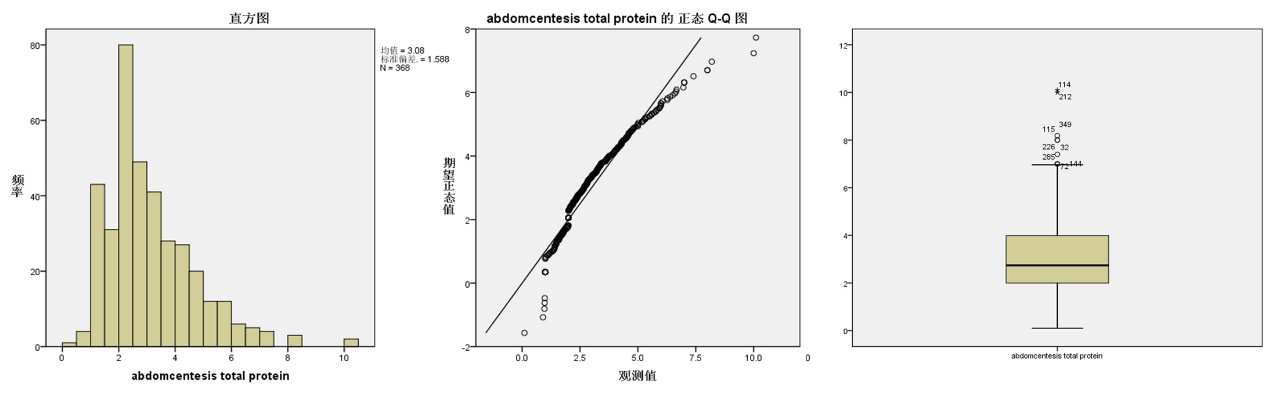
（5）packed cell volume的直方图、验证QQ图与盒图



（6）total protein的直方图、验证QQ图与盒图



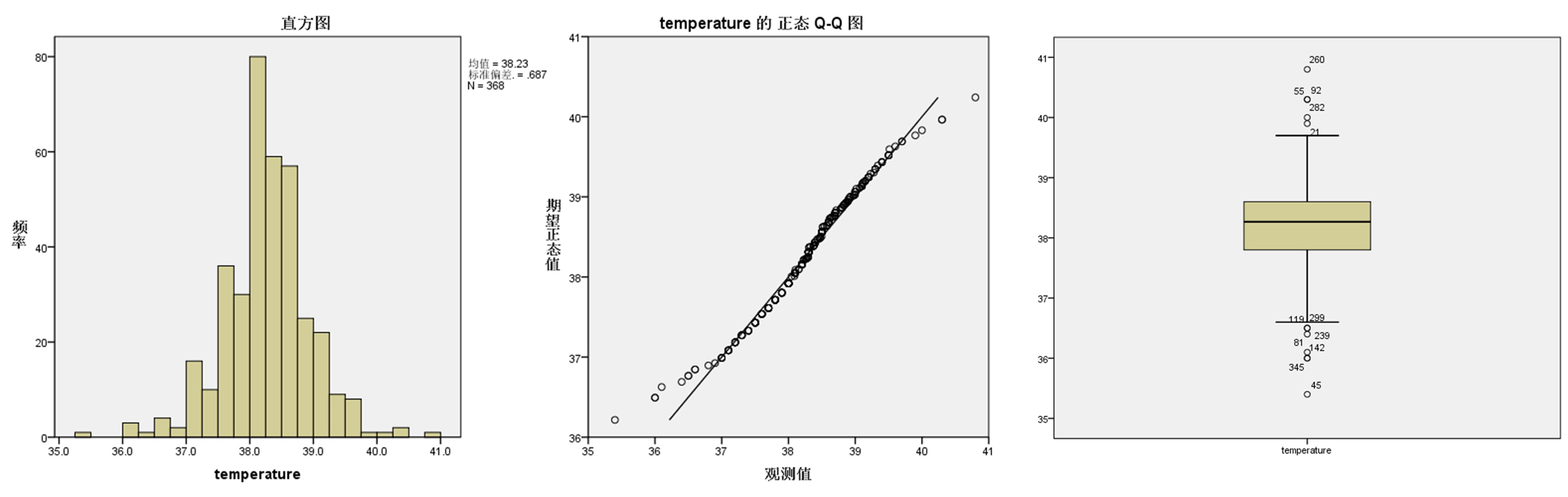
（7）abdomcentesis total protein的直方图、验证QQ图与盒图



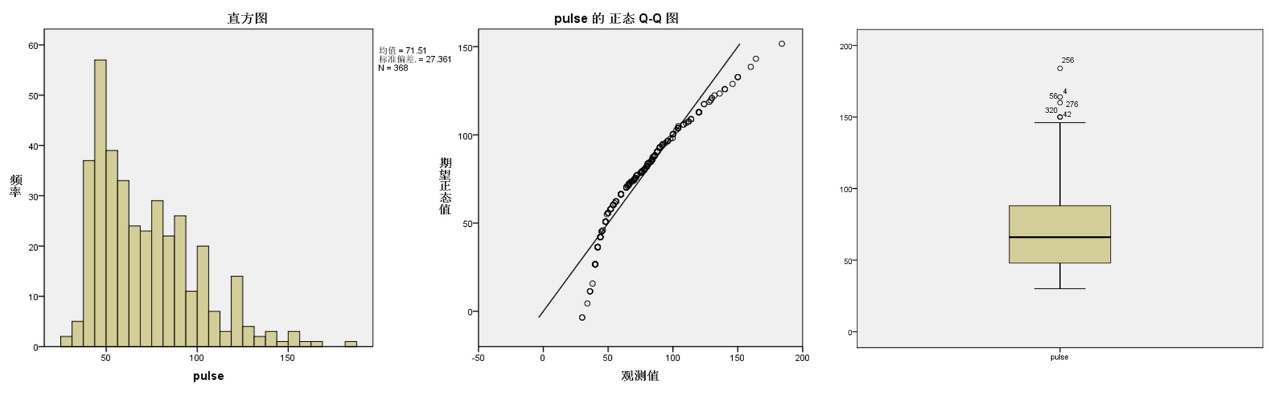
（四）通过数据对象之间的相似性来填补缺失值

数据对象之间的相似性我们采用了K-means聚类算法，将所有样本中这些属性都不为空的样本作为标准，计算要填充样本与他们的距离，寻找出最短的，将NaN补充。为了防止样本过于单一，我们每填充一个样本，就把它放到标准集中。

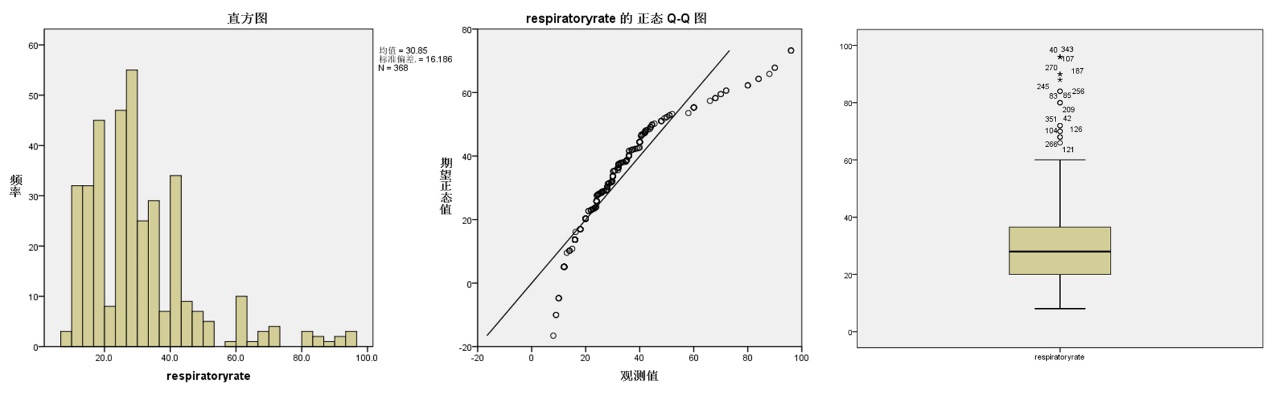
（1）rectal temperature的直方图、验证QQ图与盒图



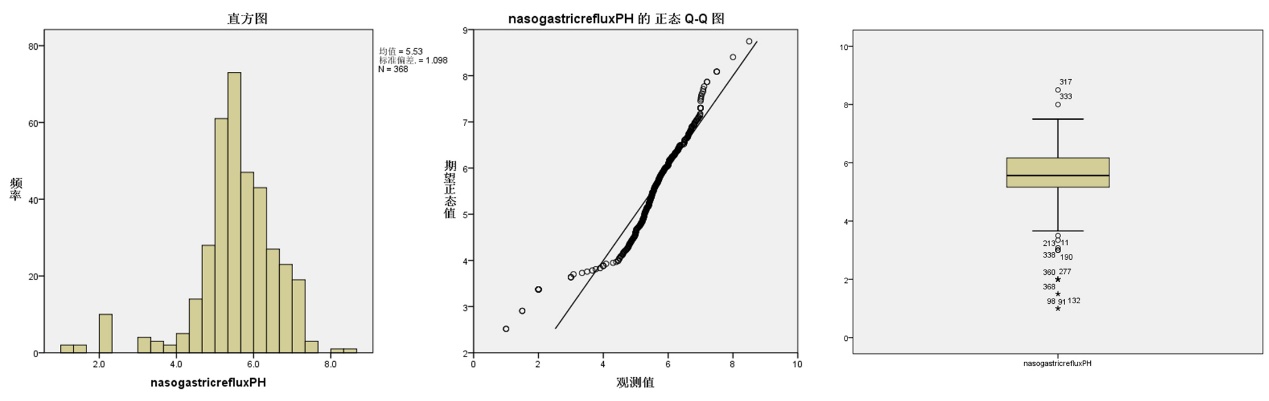
（2）pulse的直方图、验证QQ图与盒图



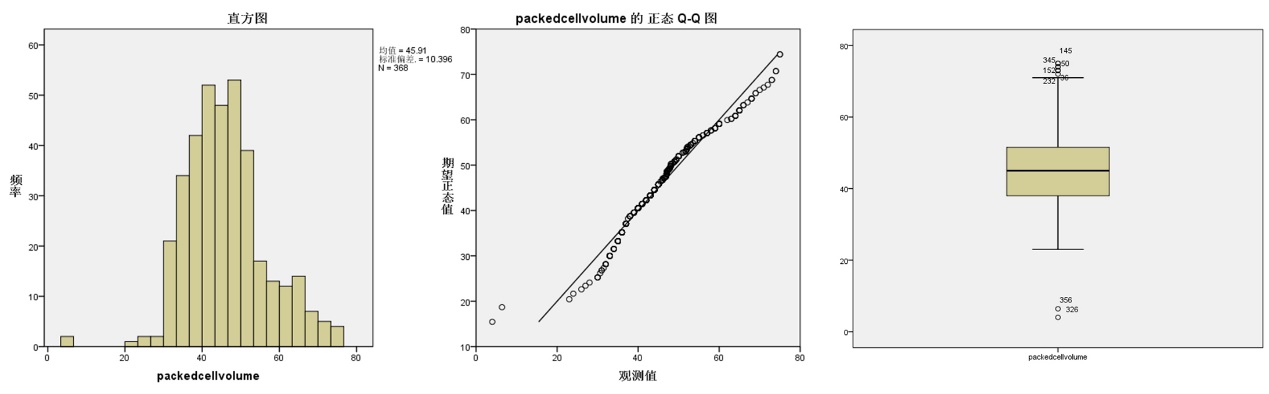
（3）respiratory rate的直方图、验证QQ图与盒图



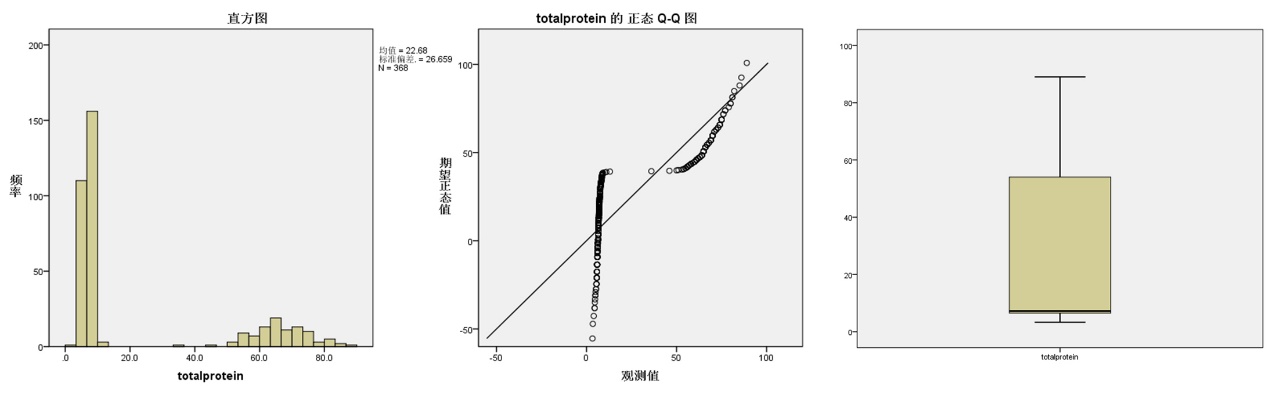
（4）nasogastric reflux PH的直方图、验证QQ图与盒图



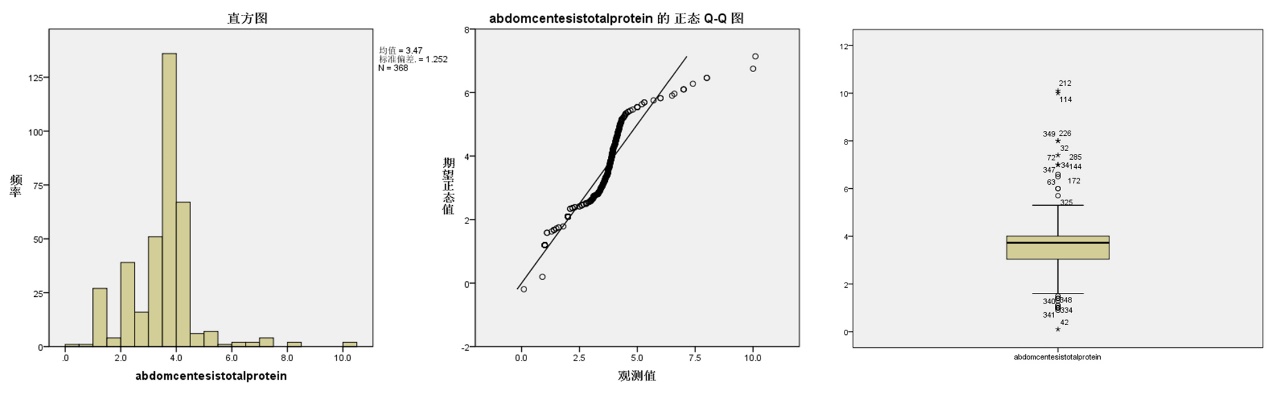
（5）packed cell volume的直方图、验证QQ图与盒图



（6）total protein的直方图、验证QQ图与盒图



（7）abdomcentesis total protein的直方图、验证QQ图与盒图



处理后，通过可视化地对比新旧数据集，我们可以发现数据集得到了有效的补充，并且新数据集和旧数据集保持了很好的一致性。

分析程序

本作业中数据处理主要借助Python工具和R语言来进行。其中，数据摘要部分通过Python工具调用Pandas库来完成，数据可视化部分使用SPSS（Statistical Product and Service Solutions）工具生成。在“数据缺失处理”部分“通过数据对象之间的相似性来填补缺失值”环节，使用R语言来解决，设计思路有两个：一是通过数据分析找出最相近个体，如果该个体中有样本缺失值，则用该个体数值填充；二是使用K-means聚类算法，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大，簇是由距离靠近的对象组成的，因此把得到紧凑且独立的簇作为最终目标，找到相似个体后填充缺失值。在本作业中采用第二种算法来解决数据对象之间的相似性填充问题。

具体分析程序代码见同本作业一起提交的“数据挖掘第一次作业”文件夹内“分析程序”相关文档。

预处理后的数据

在本作业的分析报告中，已经尽可能多地给出了预处理后的数据，限于篇幅，一些预处理后的数据参见同本作业一起提交的“数据挖掘第一次作业”文件夹内“预处理后的数据”相关文档。

附录一：作业题目

马的疝病分析

一、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。

二、数据说明

共368个样本，27个特征。关于特征的详细说明见下载链接。原始数据集及说明见同本作业一起提交的“数据挖掘第一次作业”文件夹内“数据集及特征说明”相关文档。

三、数据分析要求

（一）数据可视化和摘要

**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。

**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

（二）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。

四、提交内容

• 分析过程的报告

• 分析程序

• 预处理后的数据集

附录二：作业提交要求

一、作业提交截止时间：2017年4月17日前。

二、作业提交的形式及要求：

作业应独立完成，并将相关的文档及代码放入个人的Github仓库中；完成后将Github仓库地址发送到18010192975@163.com。

三、作业互评：

在课堂上将随机抽取3-5人进行课堂展示，其余的同学将在课后进行互评，每人至少需要评价3份作业。