马的疝病分析报告

姜林 2120161001

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

共 368 个样本，每个样本具有 28 个特征，包括：21 个标称型变量和 7 个数值型变量。

3. 数据分析要求

3.1 数据可视化和摘要

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。

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. 数据分析过程

4.1 数据摘要

对标称属性，给出每个可能取值的频数：

```
def getFrequency(attribute):
    setAttribute=set(attribute)
    res={}
    for item in setAttribute:
        res[item]=attribute.count(item)
    return res
```

部分结果：

Surgery The frequency of each value: {'1': 214, '2': 152}

Age The frequency of each value: {'1': 340, '9': 28}

Peripheral pulse The frequency of each value: {'1': 151, '3': 116, '2': 6, '4': 12}

Mucous membranes The frequency of each value: {'1': 98, '3': 81, '2': 38, '5': 28, '4': 50, '6': 25}

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：

```
def getMax(attribute):
    return max(attribute)

def getMin(attribute):
    return min(attribute)

def getAver(attribute):
    return sum(attribute)/len(attribute)

def getMedandQ(attribute):
    l=len(attribute)
    attribute.sort()
    MedandQ=[]
    if l%2==0:#median
        MedandQ.append((attribute[l/2]+attribute[l/2+1])/2)
    else:
        MedandQ.append(attribute[l/2+1])
    if (l+1)%4==0:#Q
        MedandQ.append(attribute[(l+1)/4])
        MedandQ.append(attribute[3*(l+1)/4])
    else:
        MedandQ.append((attribute[(l+1)/4]+attribute[(l+1)/4+1])/2)
        MedandQ.append((attribute[3*(l+1)/4]+attribute[3*(l+1)/4+1])/2)
    return MedandQ

def getMissCount(attribute):
    return 368-len(attribute)
```

部分结果：

Pulse Max value: 184.0 Min value: 30.0 Average value: 70.7573099415 Median, Q1,Q3: [60.0, 48.0, 88.0] The number of Miss value: 26

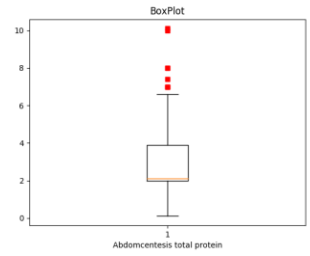
Respiratory rate Max value: 96.0 Min value: 8.0 Average value: 30.5218855219 Median,Q1,Q3: [28.0, 18.0, 36.0] The number of Miss value: 71

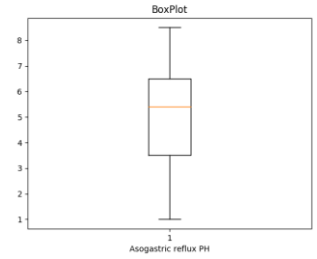
4.2 数据可视化

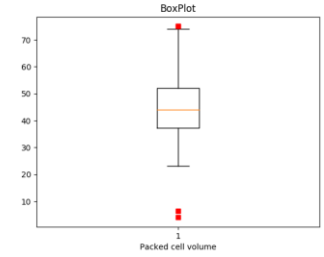
针对数值属性，绘制直方图、qq 图、盒图：

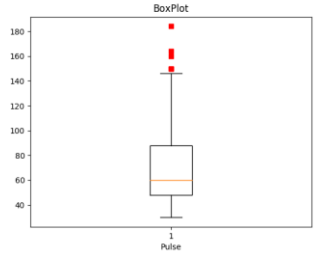
```
def drawHistogram(attribute,name):
    pl.figure()
    pl.hist(attribute,normed=True)
    stringname="plots/histogram"+name+".png"
    pl.savefig(stringname)
    pl.close()
def drawQQ(attribute,name):
    sm.qqplot(np.log(attribute),line='s')
    stringname="plots/qq"+name+".png"
    plt.savefig(stringname)
    plt.close()
def drawBox(attribute,name):
    plt.boxplot(attribute,notch=False,sym='rs',vert=True)
    plt.xlabel(name)
    plt.title('BoxPlot')
    stringname="plots/box"+name+".png"
    plt.savefig(stringname)
    plt.close()
```

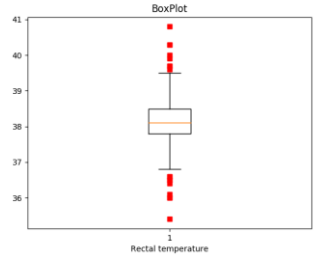
部分结果：

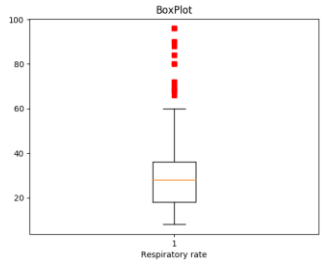


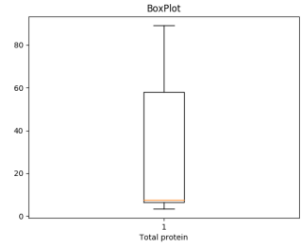


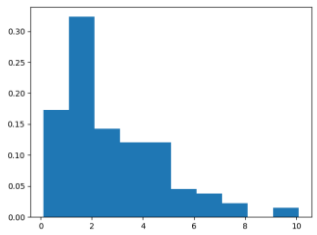


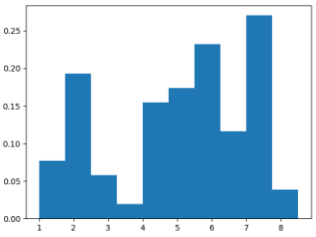


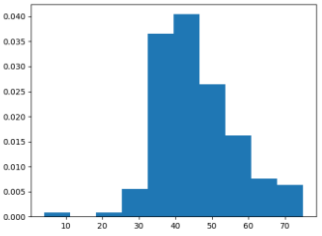


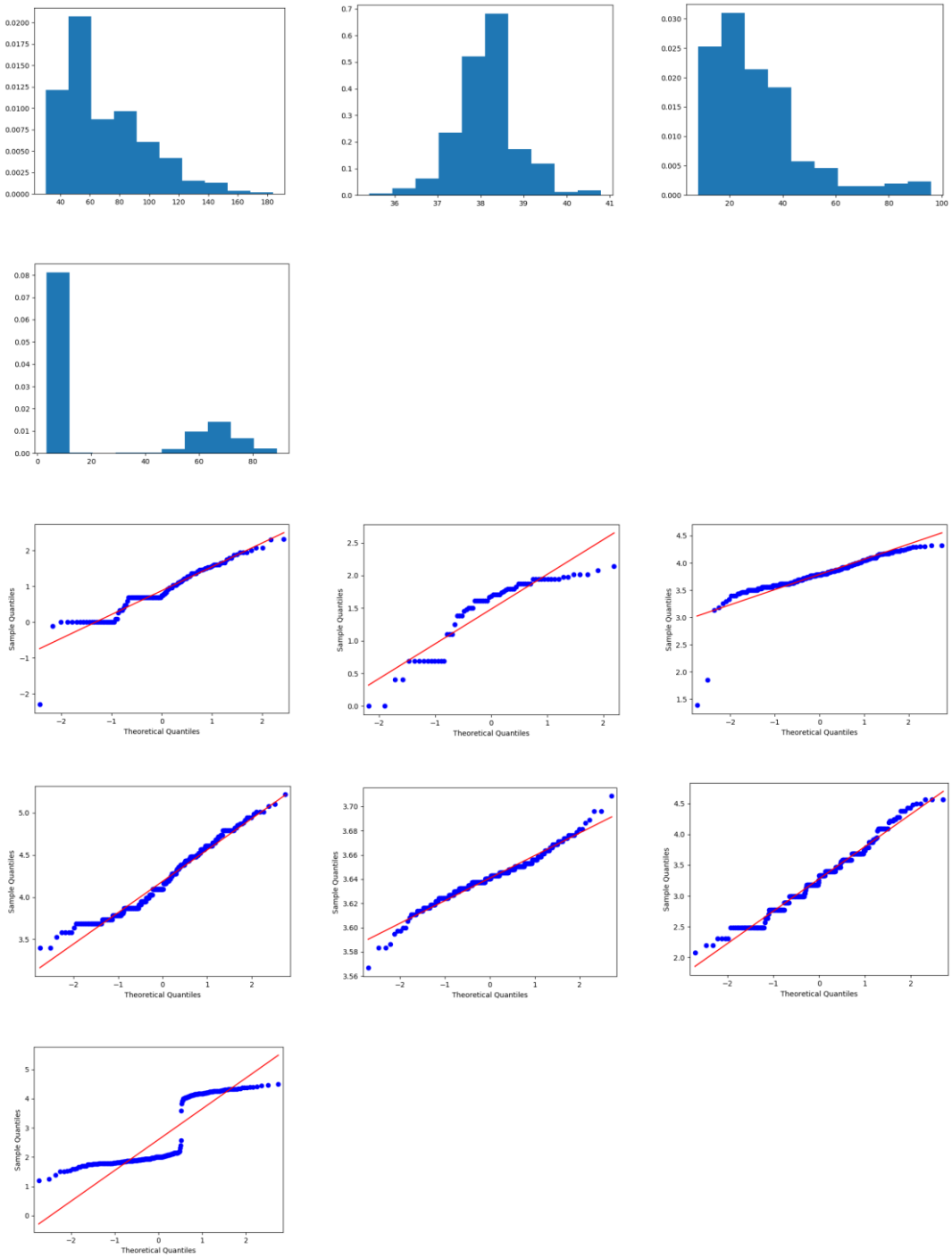












4.3 数据缺失处理

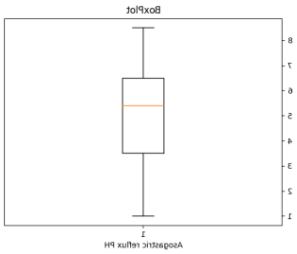
1) 将缺失部分剔除

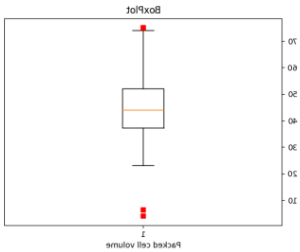
剔除含有缺失属性值的样本，剔除后，数据集只剩下 7 个完整的样本。

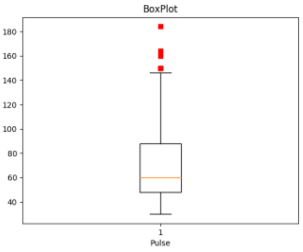
```
for line in f:
    attribute=line.split(' ')
    if '?' not in attribute:
        f2.write(line)
```

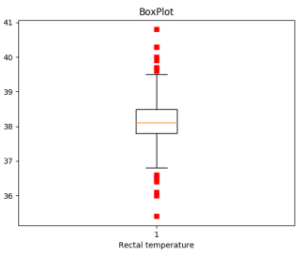
部分结果：

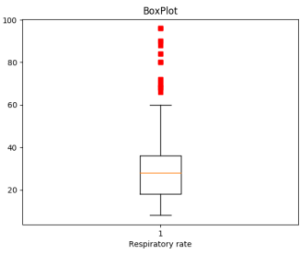

BoxPlot
Abdomcentesis total protein

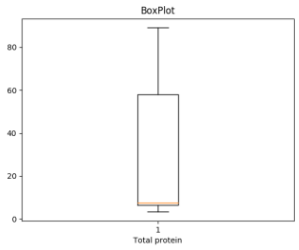

BoxPlot

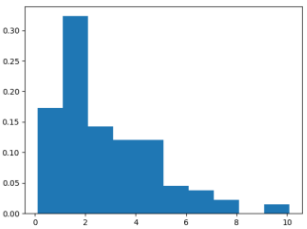

BoxPlot

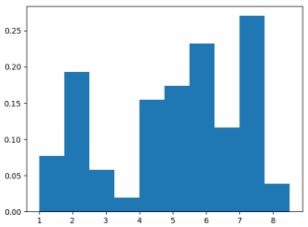

BoxPlot
Pulse

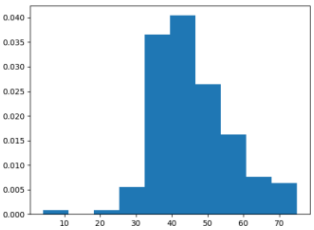

BoxPlot
Rectal temperature

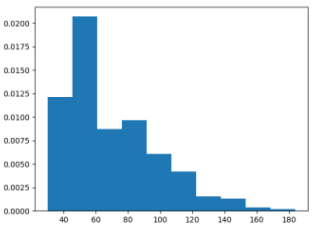

BoxPlot
Respiratory rate

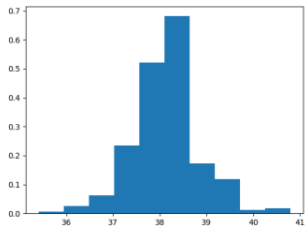

BoxPlot
Total protein

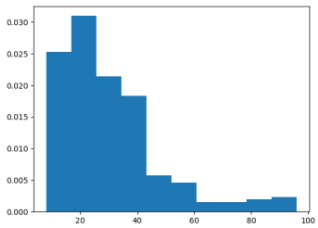


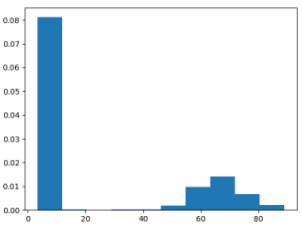


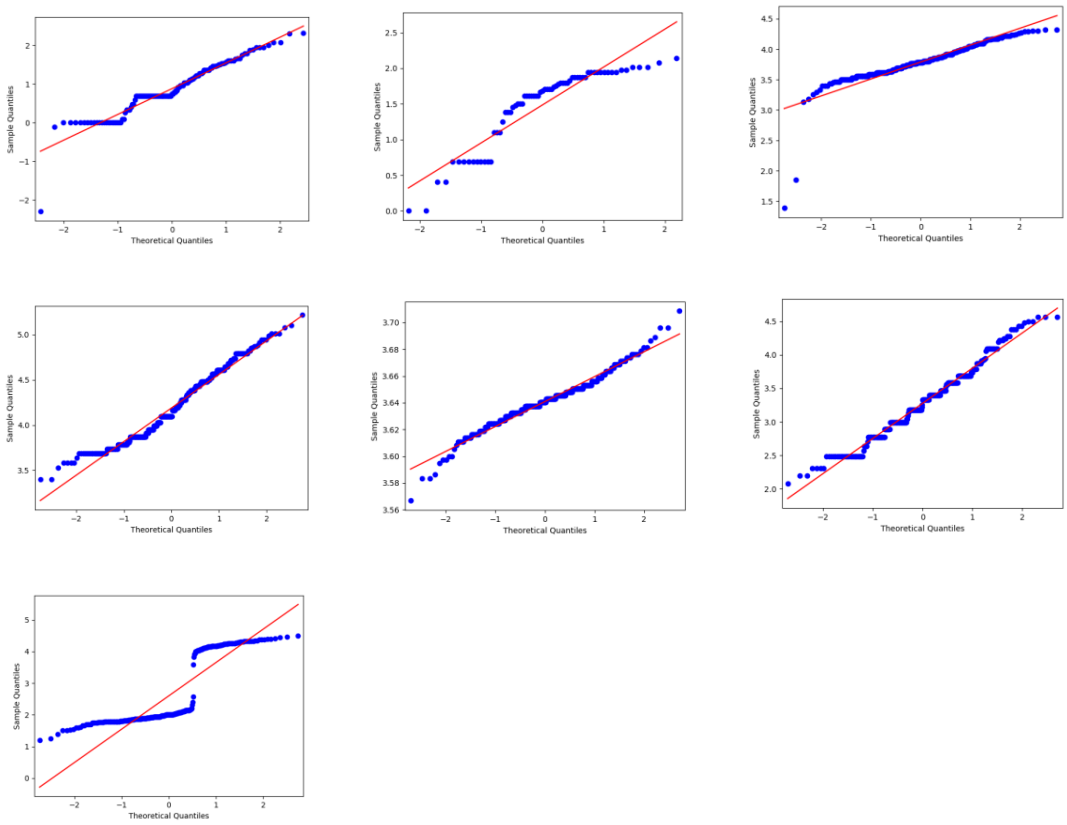










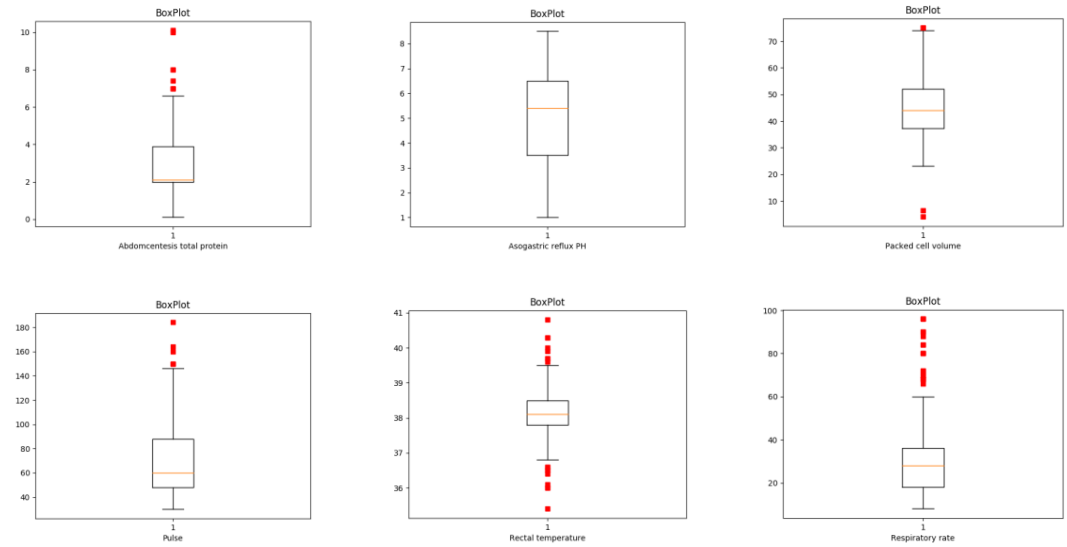


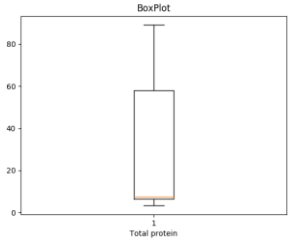
2) 用最高频率值来填补缺失值

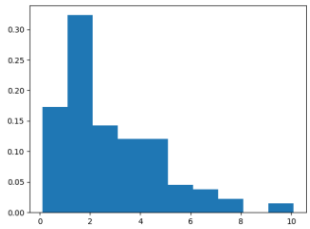
计算各属性的最高频率值，并将其填入该属性有缺失的样本中。

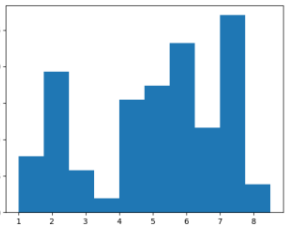
```
def getMostFrequencyList(Attribute):
    item=[]
    for i in range(28):
        item.append(getMostFrequency(Attribute[i]))
    return item
```

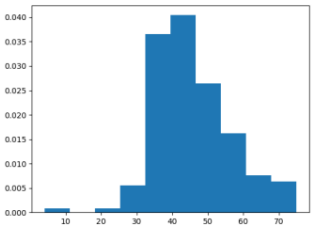
部分结果：

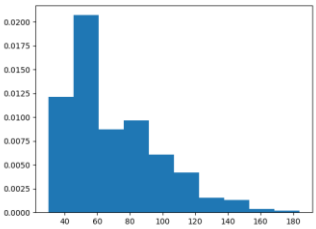


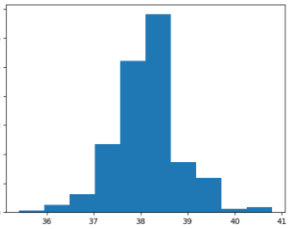

BoxPlot
0
20
40
60
80
1
Total protein

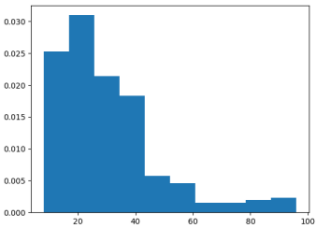

0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
2
4
6
8
10

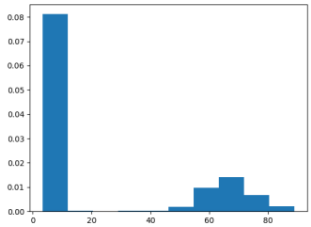

0.00
0.05
0.10
0.15
0.20
0.25
1
2
3
4
5
6
7
8

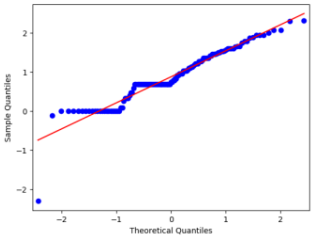

0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
10
20
30
40
50
60
70

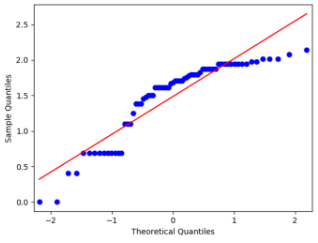

0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
0.0200
40
60
80
100
120
140
160
180

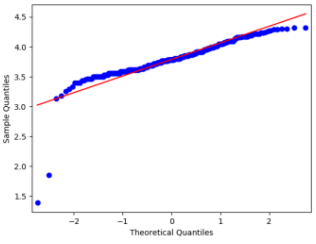

0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
36
37
38
39
40
41

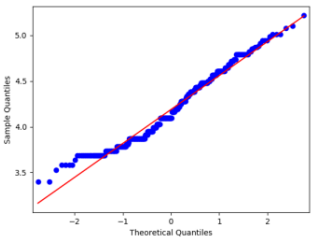

0.000
0.005
0.010
0.015
0.020
0.025
0.030
20
40
60
80
100

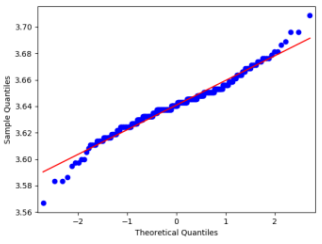

0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0
20
40
60
80

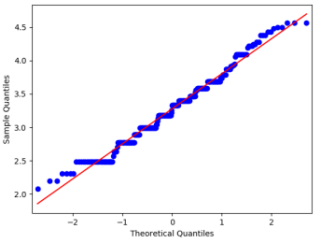

Sample Quantiles
−2
−1
0
1
2
−2
−1
0
1
2
Theoretical Quantiles

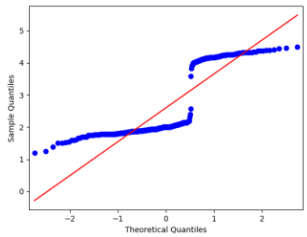

Sample Quantiles
0.0
0.5
1.0
1.5
2.0
2.5
−2
−1
0
1
2
Theoretical Quantiles


Sample Quantiles
1.5
2.0
2.5
3.0
3.5
4.0
4.5
−2
−1
0
1
2
Theoretical Quantiles


Sample Quantiles
3.5
4.0
4.5
5.0
−2
−1
0
1
2
Theoretical Quantiles


Sample Quantiles
3.56
3.58
3.60
3.62
3.64
3.66
3.68
3.70
−2
−1
0
1
2
Theoretical Quantiles


Sample Quantiles
2.0
2.5
3.0
3.5
4.0
4.5
−2
−1
0
1
2
Theoretical Quantiles

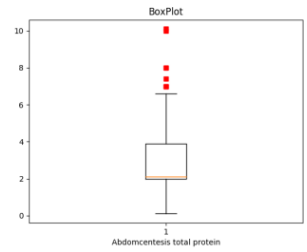


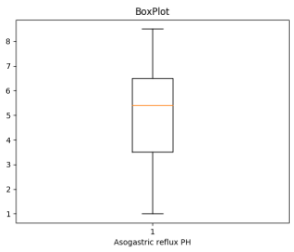
3) 通过属性的相关关系来填补缺失值

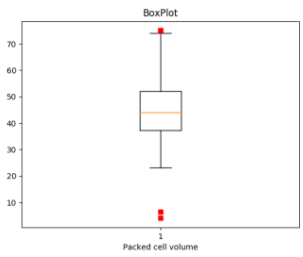
首先选取出三个重要的标称属性（age, temperature of extremities, abdominal distension），计算出在这三个属性确定的情况下，其他属性的各个可能取值出现的概率，将概率最大的取值填入缺失样本。

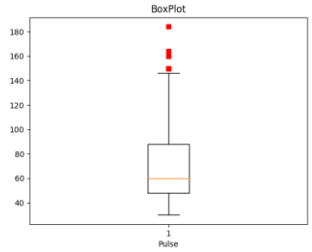
```
for i in range(28):
    attribute=horses[H].getIthAtribute(i)
    if attribute=='?':
        atts=[]
        for h in range(len(horsesSameAttributes)):
            att=horsesSameAttributes[h].getIthAtribute(i)
            if att!='?':
                atts.append(att)
        setAtts=set(atts)
        mostFrequency=0
        mostR=""
        sign0=0
        for r in setAtts:
            frequency=atts.count(r)
            if sign0==0:
                mostFrequency=frequency
                mostR=r
                sign0=1
            elif frequency>mostFrequency:
                mostFrequency=frequency
                mostR=r
        attribute=mostR
    if i==0:
        line=attribute
    else:
        line=line+' '+attribute
```

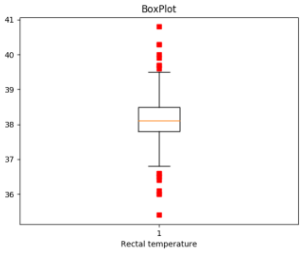
部分结果：

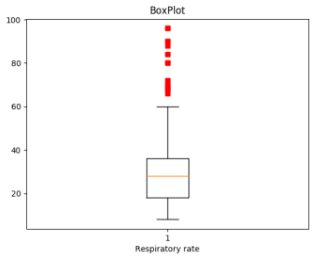


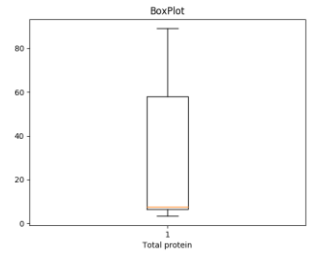


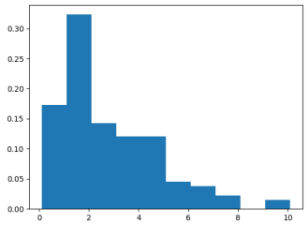


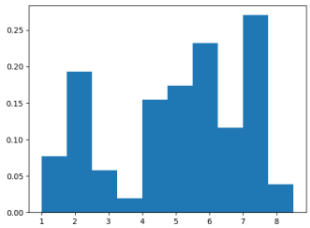

BoxPlot
Pulse

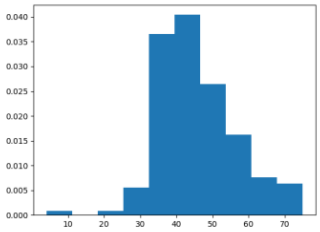

BoxPlot
Rectal temperature

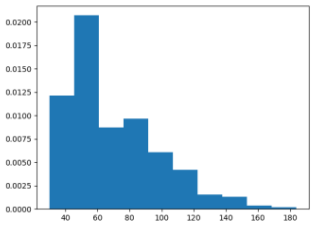

BoxPlot
Respiratory rate

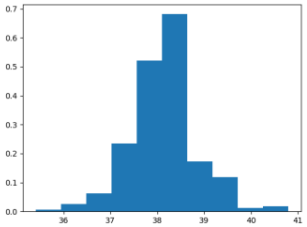

BoxPlot
Total protein

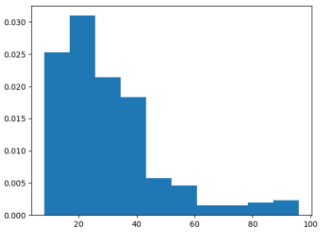


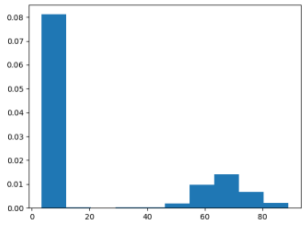


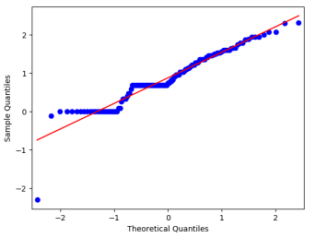


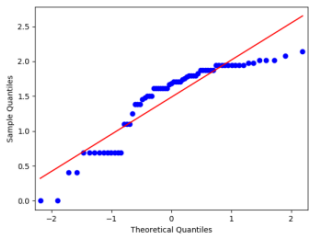


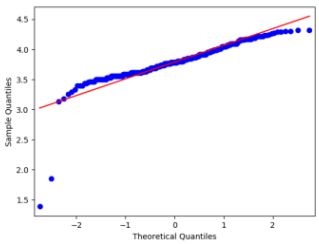


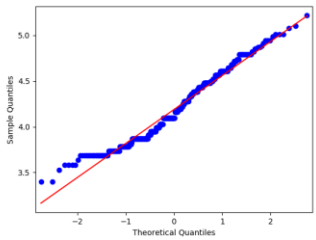


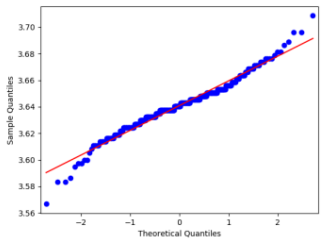


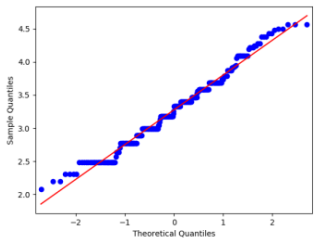

Sample Quantiles
Theoretical Quantiles

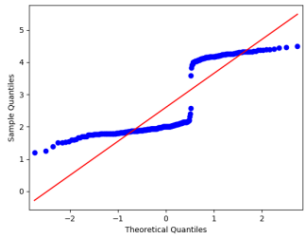

Sample Quantiles
Theoretical Quantiles


Sample Quantiles
Theoretical Quantiles







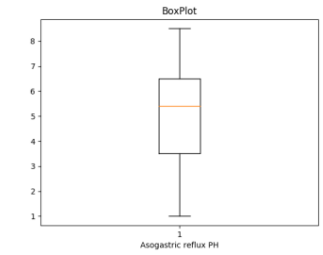


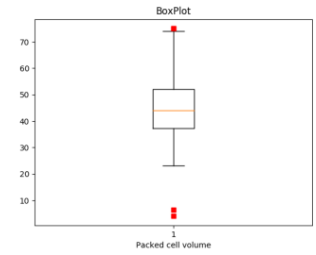
4) 通过数据对象之间的相似性来填补缺失值

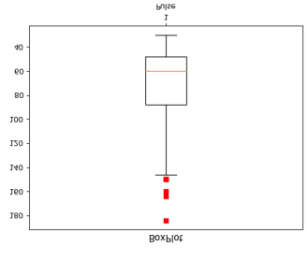
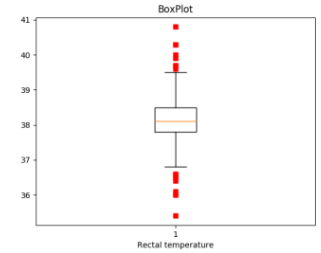
对数据集中的其他所有样本，计算标称属性的 Hamming 距离，数值属性的欧式距离，最后选出与待填补样本最相似的样本进行填补。

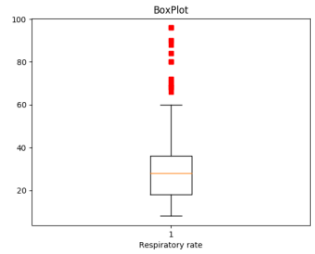
```
def hammingDist(x,y):
    if len(x) != len(y):
        return 1000
    count=0
    for i in range(len(x)):
        if x[i] != y[i]:
            count=count+1
    return count
```

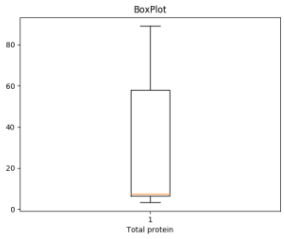
部分结果：

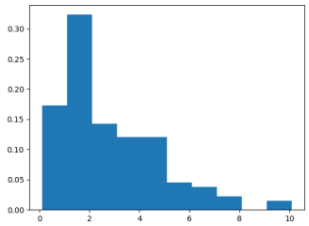


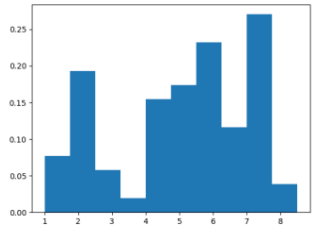


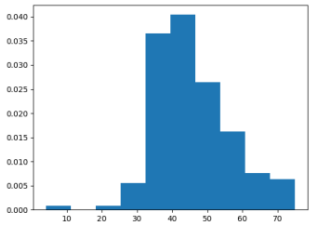



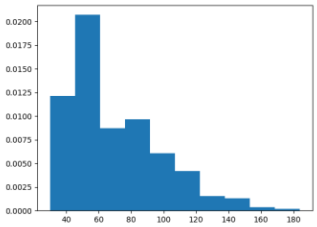


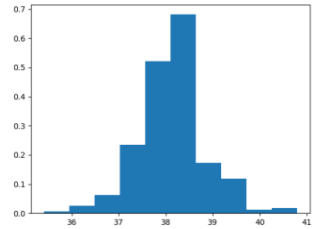

BoxPlot
80
60
40
20
0
1
Total protein

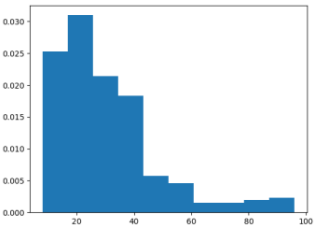

0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
10

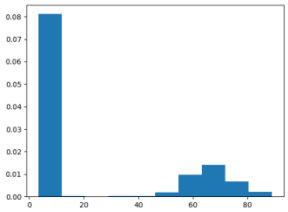

0.25
0.20
0.15
0.10
0.05
0.00
1
2
3
4
5
6
7
8

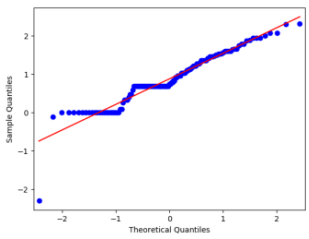

0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
10
20
30
40
50
60
70

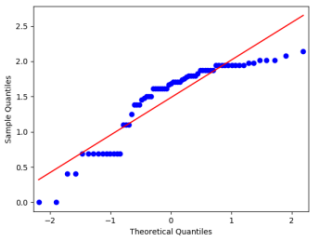

0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
40
60
80
100
120
140
160
180

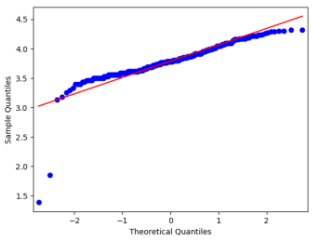

0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
36
37
38
39
40
41

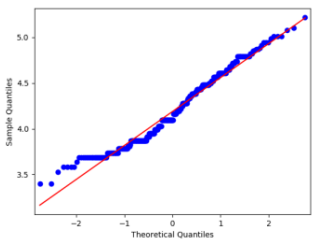

0.030
0.025
0.020
0.015
0.010
0.005
0.000
20
40
60
80
100

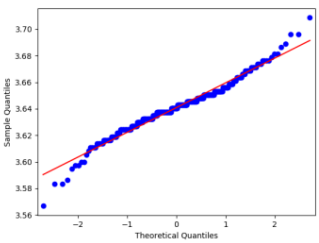

0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
20
40
60
80

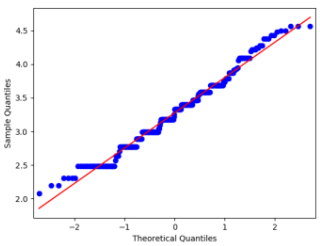

2
1
0
−1
−2
Sample Quantiles
−2
−1
0
1
2
Theoretical Quantiles

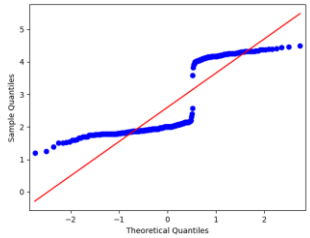

2.5
2.0
1.5
1.0
0.5
0.0
Sample Quantiles
−2
−1
0
1
2
Theoretical Quantiles


4.5
4.0
3.5
3.0
2.5
2.0
1.5
Sample Quantiles
−2
−1
0
1
2
Theoretical Quantiles


5.0
4.5
4.0
3.5
Sample Quantiles
−2
−1
0
1
2
Theoretical Quantiles


3.70
3.68
3.66
3.64
3.62
3.60
3.58
3.56
Sample Quantiles
−2
−1
0
1
2
Theoretical Quantiles


4.5
4.0
3.5
3.0
2.5
2.0
Sample Quantiles
−2
−1
0
1
2
Theoretical Quantiles



5. 数据分析代码目录

- code ----------源码文件夹
  - Data Summary and Visualization----------数据摘要及可视化
    - dataSummaryAndVisualization.py----------生成数据摘要及图形的代码
    - initialDataset.txt----------初始数据集
    - results----------处理结果文件夹
      - box*.png----------盒图文件
      - histogram*.png----------直方图文件
      - qq*.png----------qq 图文件
      - dataSummary.txt----------数据摘要文件
  - Data Processing----------数据缺失处理
    - Remove the missing part----------将缺失部分剔除
      - removeTheMissingPart.py----------剔除缺失部分的代码
      - initialDataset.txt----------初始数据集
      - newDataset.txt----------处理后的新数据集
      - results----------处理结果文件夹
        - box*.png----------盒图文件
        - histogram*.png----------直方图文件
        - qq*.png----------qq 图文件
        - dataSummary.txt----------数据摘要文件
    - Highest frequency----------用最高频率值来填补缺失值
      - highestFrequency.py----------用最高频率填补的代码
      - initialDataset.txt----------初始数据集
      - newDataset.txt----------处理后的新数据集
      - results----------处理结果文件夹
        - box*.png----------盒图文件
        - histogram*.png----------直方图文件
        - qq*.png----------qq 图文件
        - dataSummary.txt----------数据摘要文件
    - Correlativity of attributes----------通过属性的相关关系来填补缺失值
      - correlativityOfAttributes.py----------通过相关关系填补的代码
      - initialDataset.txt----------初始数据集
      - newDataset.txt----------处理后的新数据集
      - results----------处理结果文件夹
        - box*.png----------盒图文件

        - histogram*.png----------直方图文件
        - qq*.png----------qq 图文件
        - dataSummary.txt----------数据摘要文件
- Similarity of data----------通过数据对象之间的相似性来填补缺失值
    - similarityOfData----------通过相似性填补的代码
    - initialDataset.txt----------初始数据集
    - newDataset.txt----------处理后的新数据集
    - results----------处理结果文件夹
        - box*.png----------盒图文件
        - histogram*.png----------直方图文件
        - qq*.png----------qq 图文件
        - dataSummary.txt----------数据摘要文件