作业一 马的疝病分析

姓名：常明

学号：2120160580

**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。

**三. 数据分析要求**

**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

**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。

**四、实验环境及语言**

语言： python 3.5

依赖的包：requests,matplotlib, statsmodel, numpy

requests:数据获得

numpy:数据处理

matplotlib, statsmodel：数据可视化时用于生成图

**五、实现方法**

0 数据下载

程序：dataDownload.py, pre\_processing.py

使用dataDownload.py脚本对数据下载，保存为.txt文件，使用脚本pre\_processing.py将数据转为numpy数组形式，便于处理。

1 数据摘要

程序：dataSummary.py

数据有28个属性其中第4, 5, 6, 16, 19, 20, 22属性为数值属性，其余为标称属性。

函数介绍：

nominal\_attribute()处理标称属性

median()求中位数

numeric\_attribute()处理数值属性

数据处理时默认提出了缺失值，最后得到的数据保存为json文件。

示例：

{

"1": {"2": 152, "1": 214}, "2": {"9": 28, "1": 340}

……

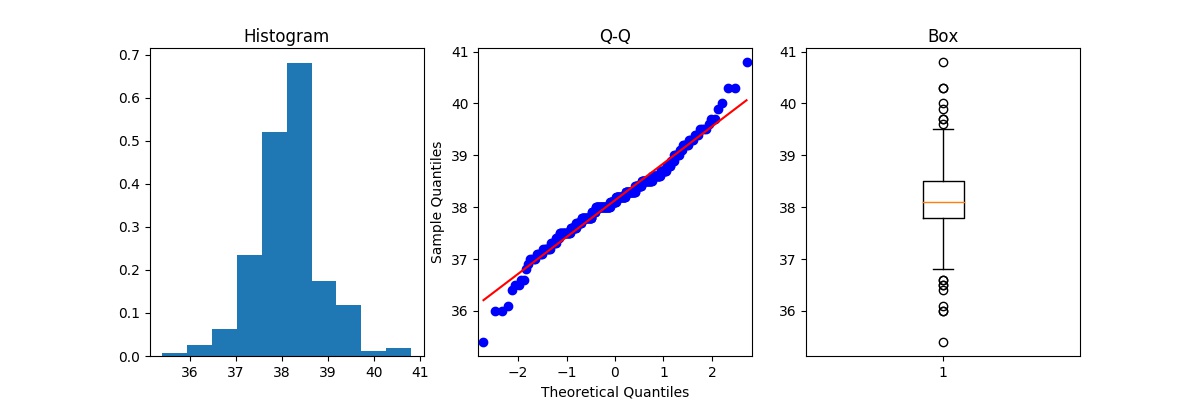
}

其中外层键代表第几属性。

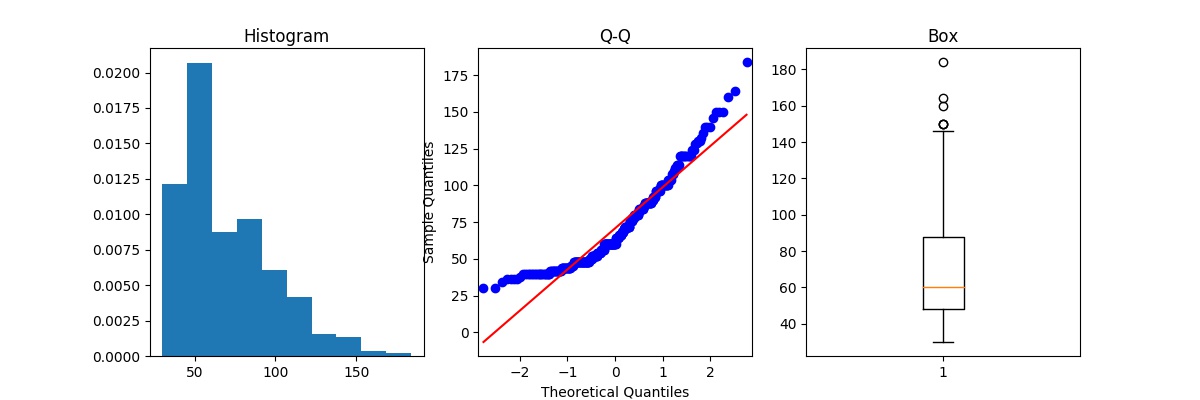
2 数据可视化

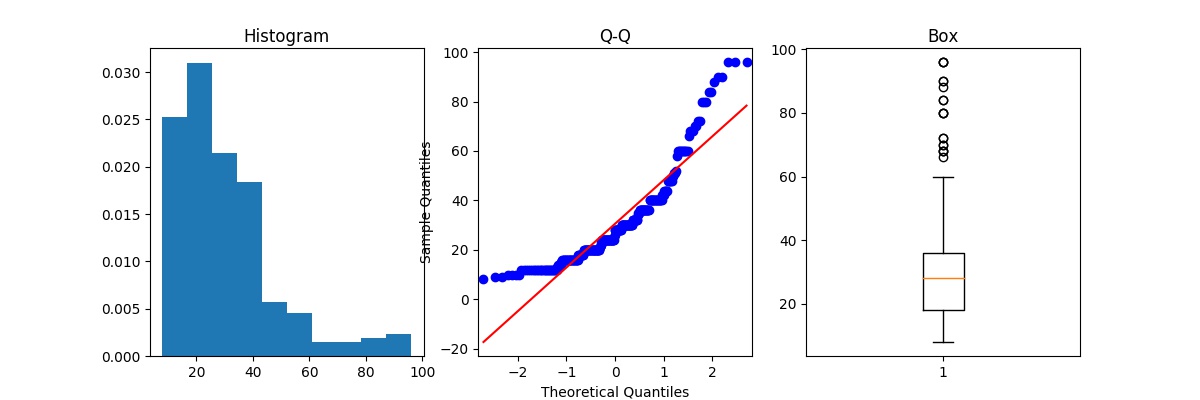
程序：visualize.py

对7个数值属性进行了可视化，数据处理时同样对缺失值进行了剔除。

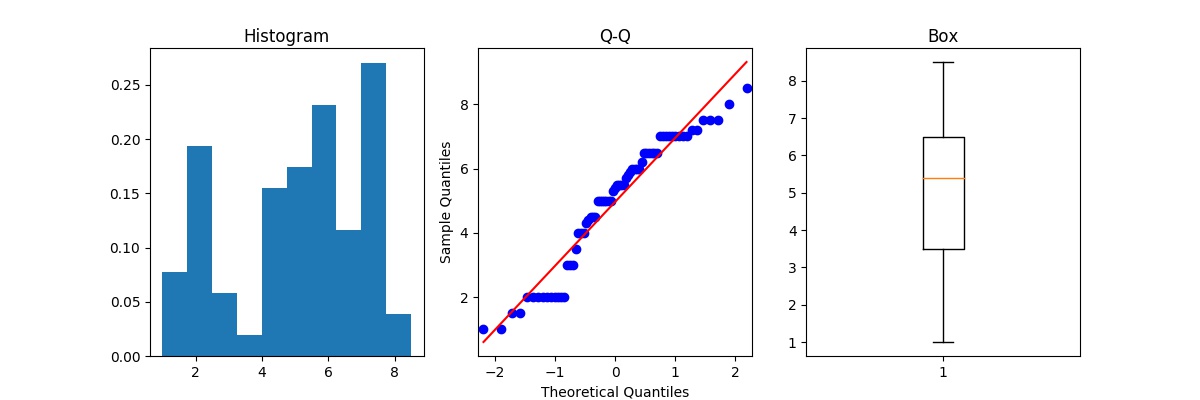
可视化结果保存为.jpg文件，效果如下：

属性4

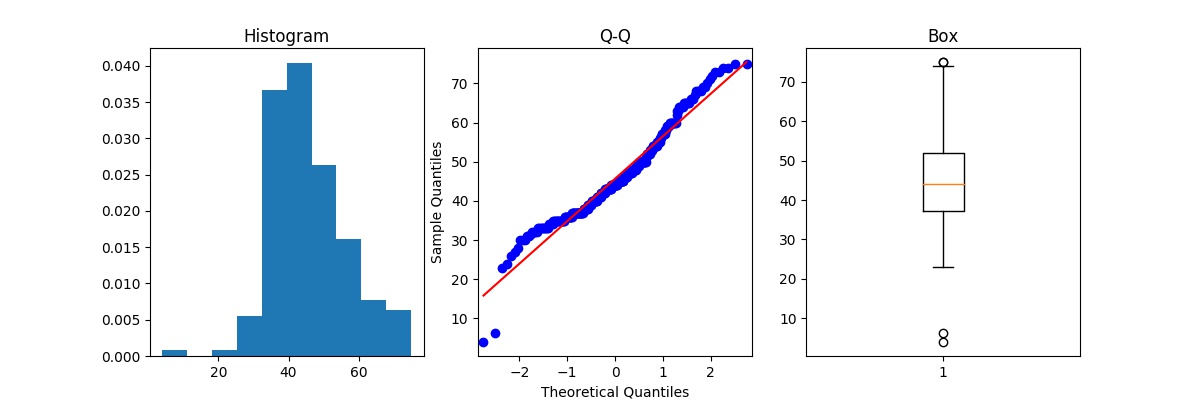
属性5

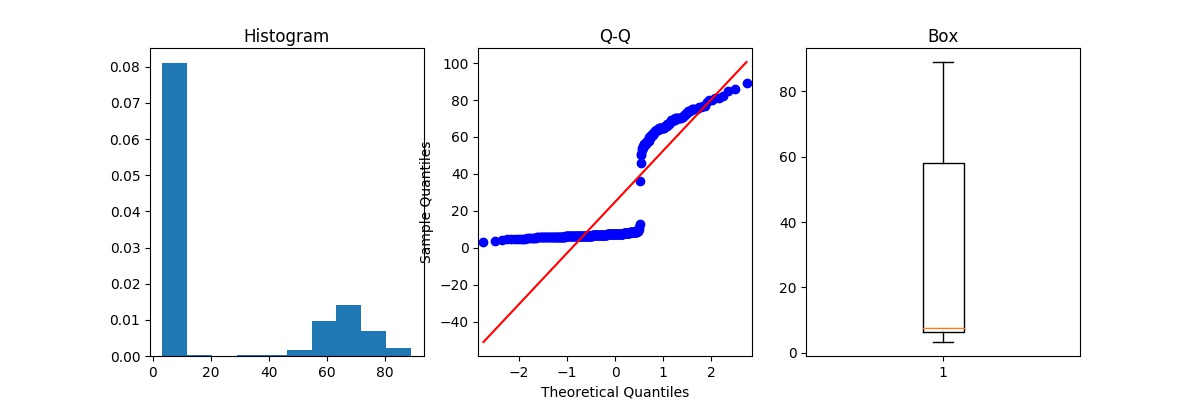


属性6

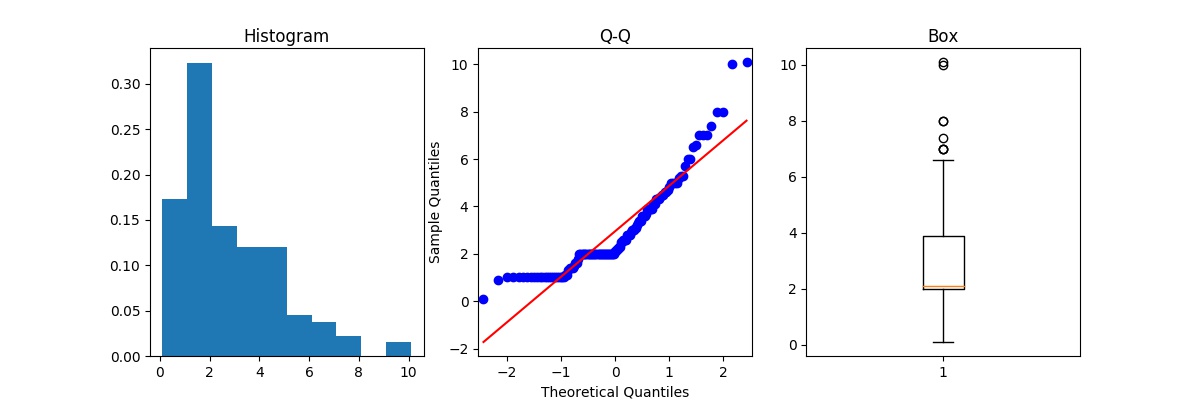


属性16

属性19



属性20



属性22

从图中可以看出只有属性4和属性19（直肠温度，填充细胞体积）符合正态分布。

**3 缺失值处理**

程序：correlate.py, dissimilarity.py, visualize.py

程序说明：

correlate.py用来计算各属性的相关系数

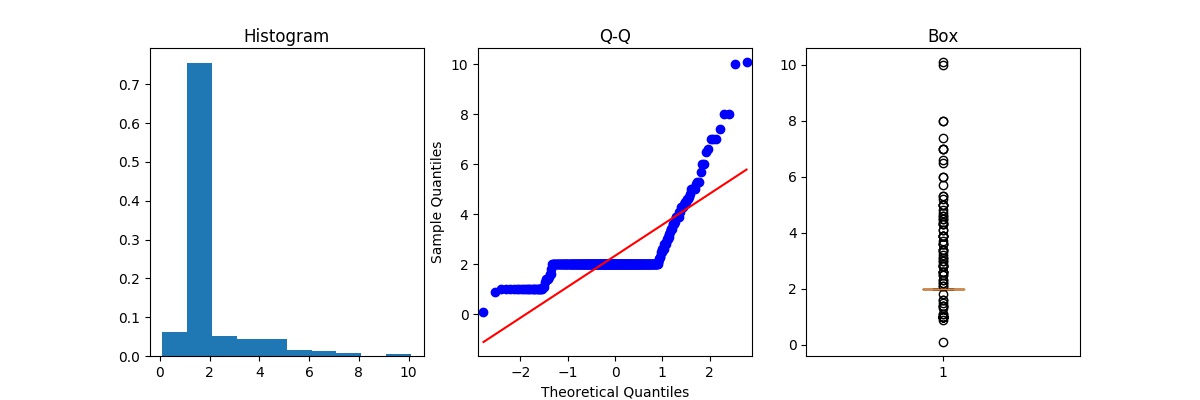
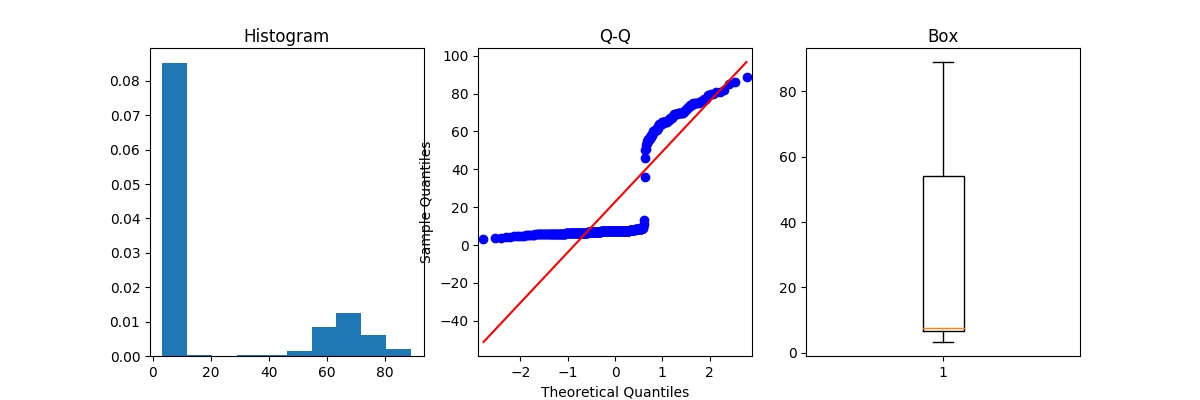
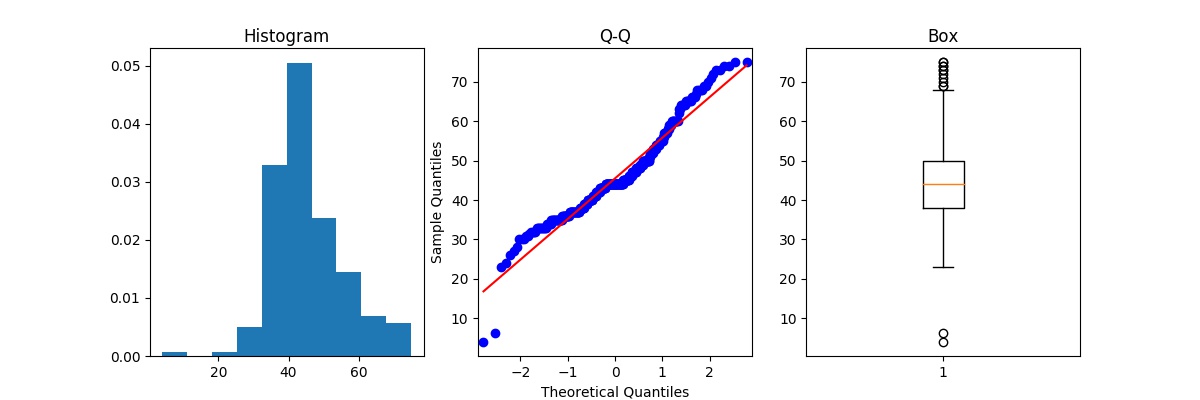
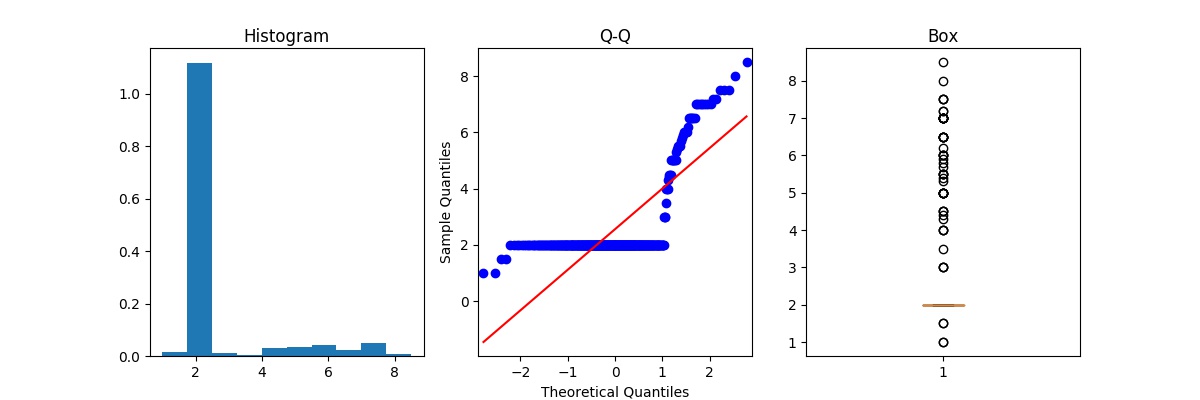
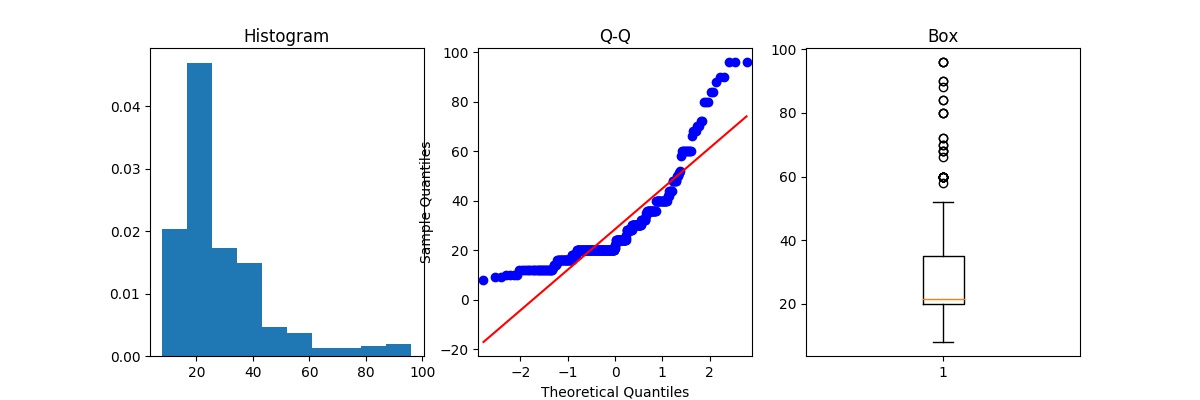
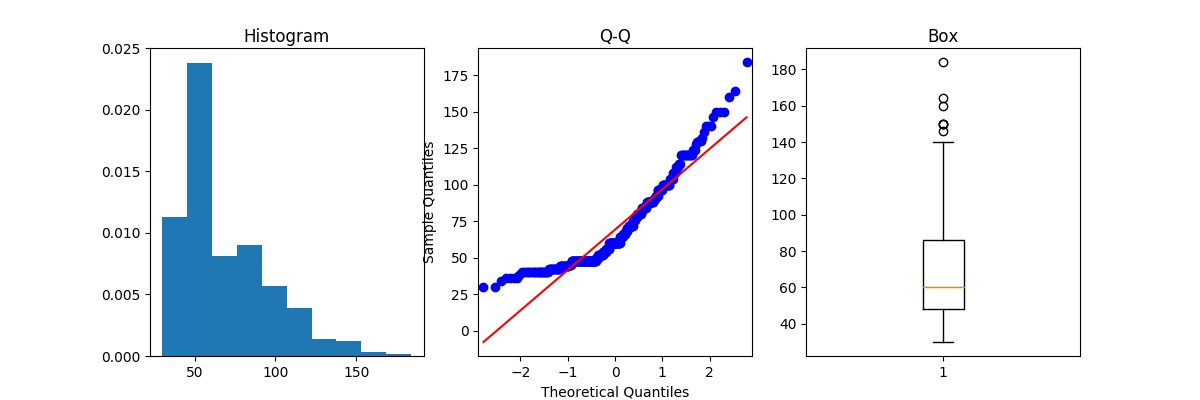
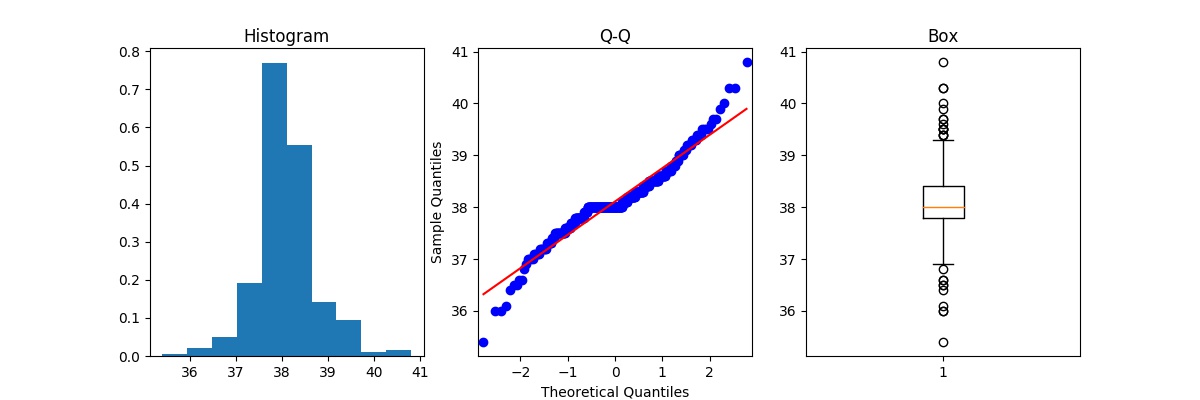
dissimilarity.py用来计算各对象的相似性

visualize.py用来可视化数据，其中的函数visualize\_data()可以传入参数决定数据处理时缺失值的处理方法。

3.1剔除缺失值

上步数据可视化时采用的数据处理方式就是剔除缺失值，这里就不再赘述了。

3.2 用最高频率的值代替缺失值



可视化结果保存为了jpg文件，命名形式为“filled\_fre\_X”（X为属性的序数（4，5，6，19，20，22））。

3.3 通过属性的相关关系来填补缺失值

通过计算各属性的相关系数，可以得到一个7x7的矩阵

[[ 1. 0.21363955 0.25166039 0.23524114 0.07258469 -0.01567165

-0.01395907]

[ 0.21363955 1. 0.44552019 0.04593649 0.40913557 -0.0775091

0.07710956]

[ 0.25166039 0.44552019 1. 0.091003 0.07714125 -0.08620983

-0.05150615]

[ 0.23524114 0.04593649 0.091003 1. -0.09649049 -0.69112507

0.27626632]

[ 0.07258469 0.40913557 0.07714125 -0.09649049 1. -0.08706951

0.15478027]

[-0.01567165 -0.0775091 -0.08620983 -0.69112507 -0.08706951 1.

-0.49817627]

[-0.01395907 0.07710956 -0.05150615 0.27626632 0.15478027 -0.49817627

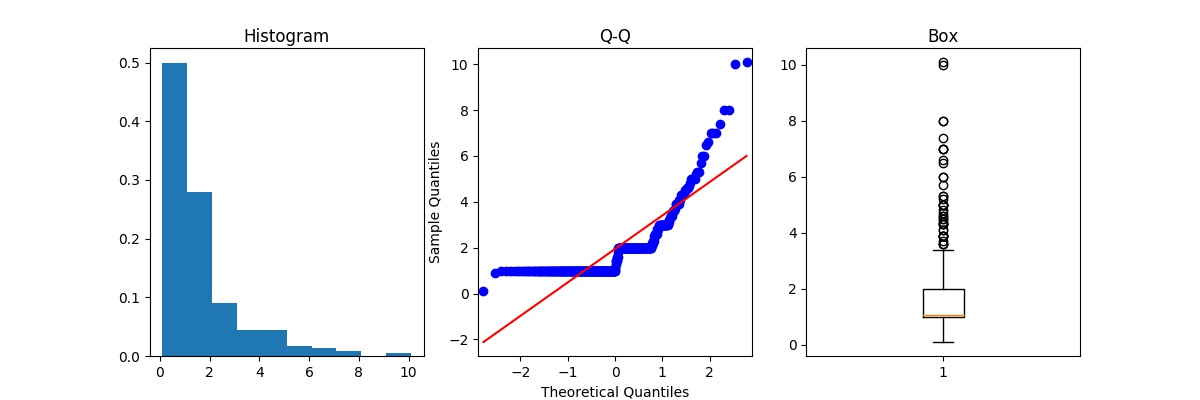
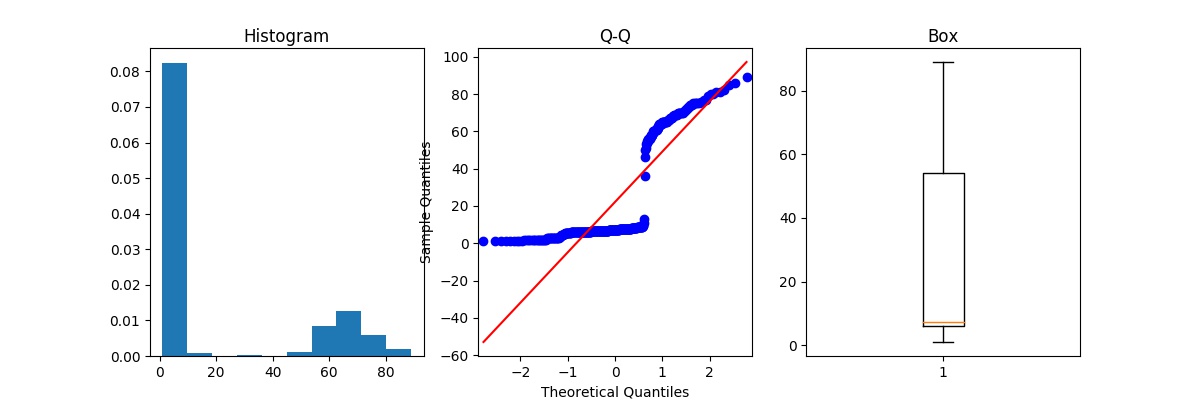
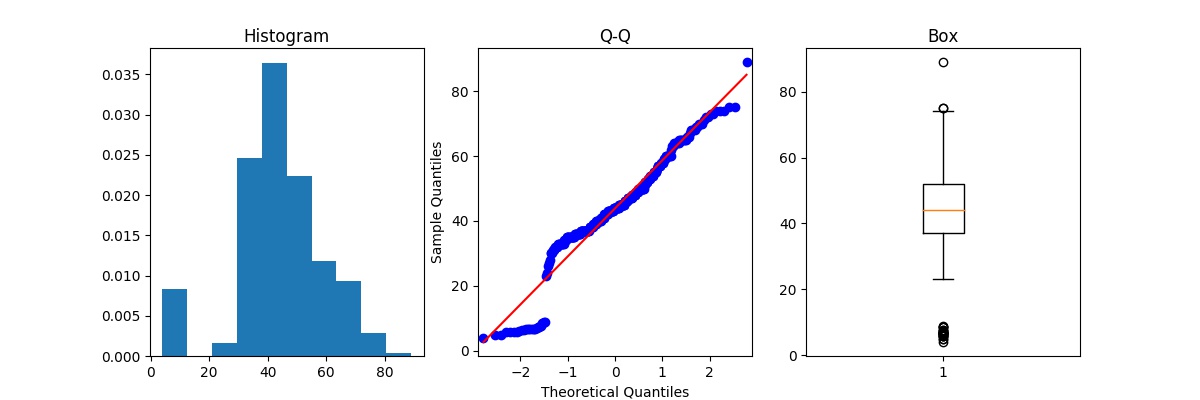
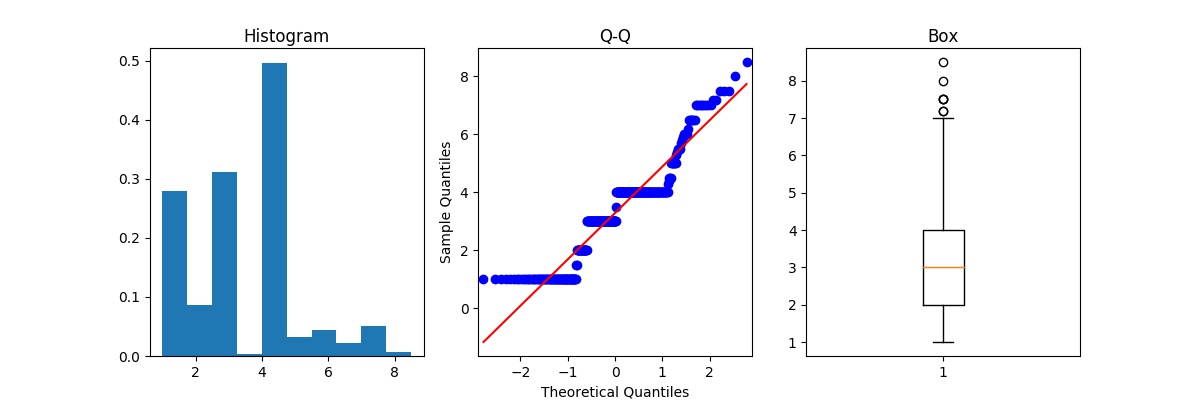
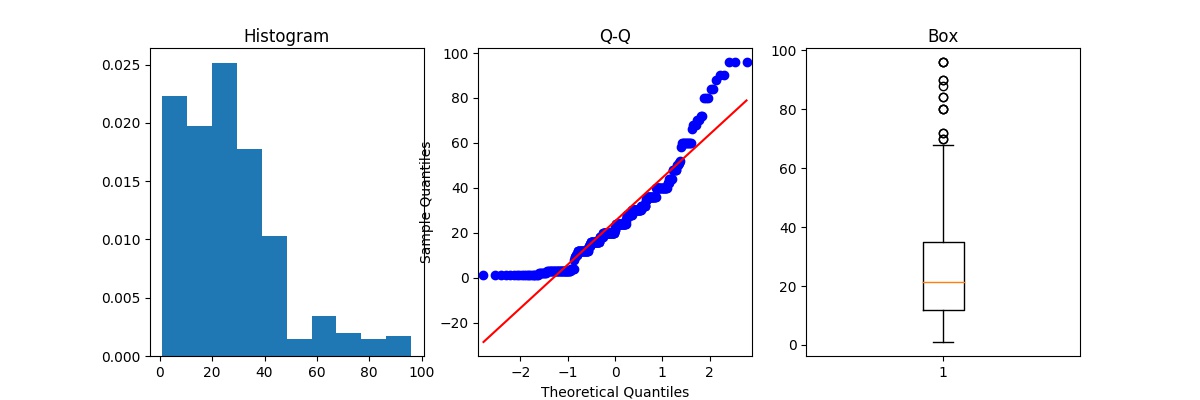
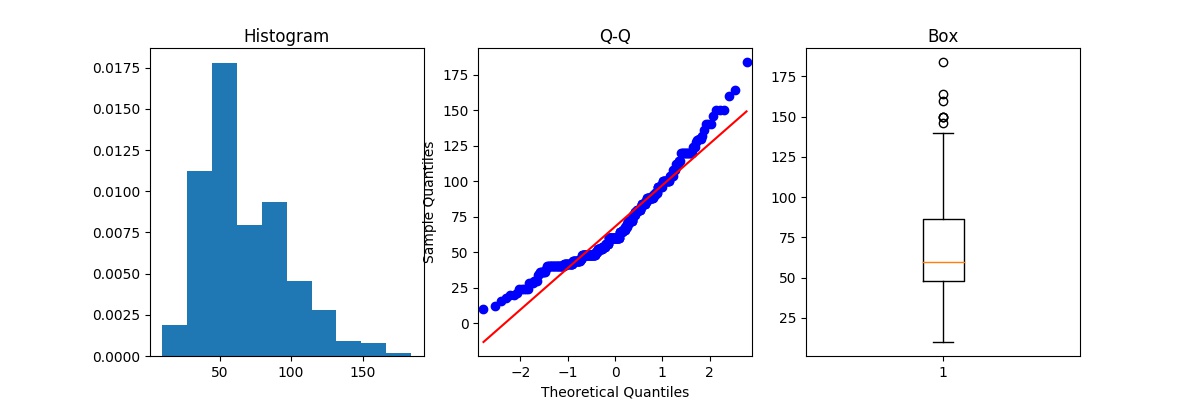
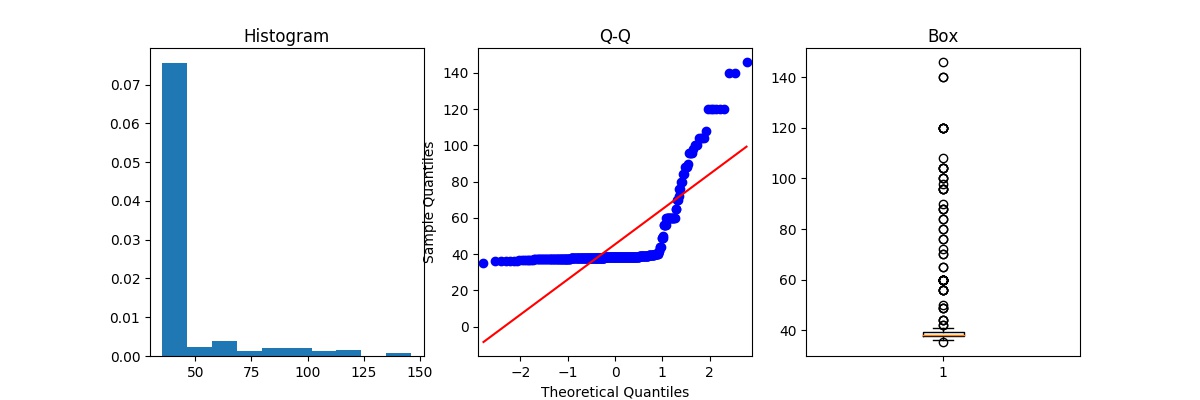
1. ]]

可以看出，各属性的相关系数都很小，相关关系很弱，所以不适合通过相关关系对缺失值进行填充。

3.4 通过数据对象之间的相似性来填补缺失值

根据书上给出的计算混合对象相似性的方法，可以得出各个对象的相似程度，用一个368x368的矩阵来表示相似性。通过dissimilarity.py计算出该矩阵，并存储在sim\_array.pkl文件中。

可视化结果如下：



可视化结果保存为了jpg文件，命名形式为“filled\_sim\_X”（X为属性的序数（4，5，6，19，20，22））。