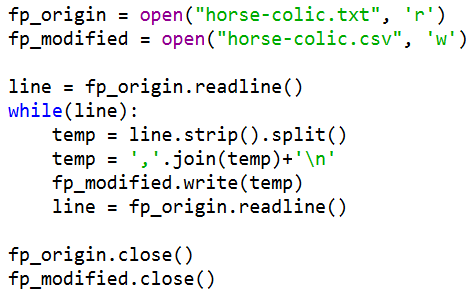
**马的疝病分析**

郭强

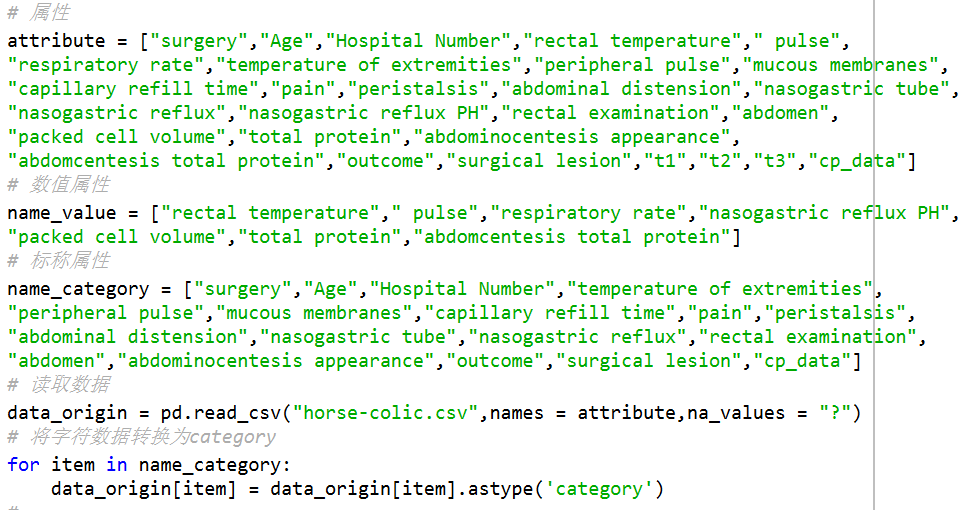
2120160993

**一 数据准备**

首先将原始的txt文件转换成csv文件，以便于对数据进行分析：



然后读取数据，并且将属性分为标称属性和数值属性：

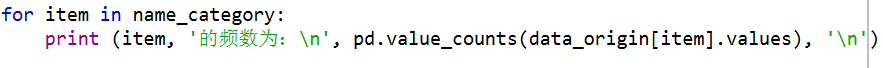


**二 数据摘要和可视化**

**2.1 数据摘要**

**2.1.1 标称属性**

对标称属性，给出每个可能取值的频数。使用pandas中的value\_counts函数统计每个标称属性的取值频数：



统计结果为：

Surgery的频数为：

1.0 214

2.0 152

Age的频数为：

1 340

9 28

temperature of extremities的频数为：

3.0 135

1.0 95

2.0 39

4.0 34

peripheral pulse的频数为：

1.0 151

3.0 116

4.0 12

2.0 6

mucous membranes的频数为：

1.0 98

3.0 81

4.0 50

2.0 38

5.0 28

6.0 25

capillary refill time的频数为：

1.0 232

2.0 96

3.0 2

pain 的频数为：

3.0 82

2.0 77

5.0 50

1.0 49

4.0 47

peristalsis 的频数为：

3.0 154

4.0 91

1.0 49

2.0 22

abdominal distension 的频数为：

1.0 101

3.0 85

2.0 75

4.0 42

nasogastric tube 的频数为：

2.0 121

1.0 89

3.0 27

nasogastric reflux 的频数为：

1.0 141

3.0 49

2.0 45

rectal examination 的频数为：

4.0 97

1.0 68

3.0 61

2.0 14

abdomen 的频数为：

5.0 96

4.0 55

1.0 31

2.0 24

3.0 19

abdominocentesis appearance 的频数为：

2.0 62

3.0 60

1.0 52

outcome 的频数为：

1.0 225

2.0 89

3.0 52

surgical lesion 的频数为：

1 232

2 136

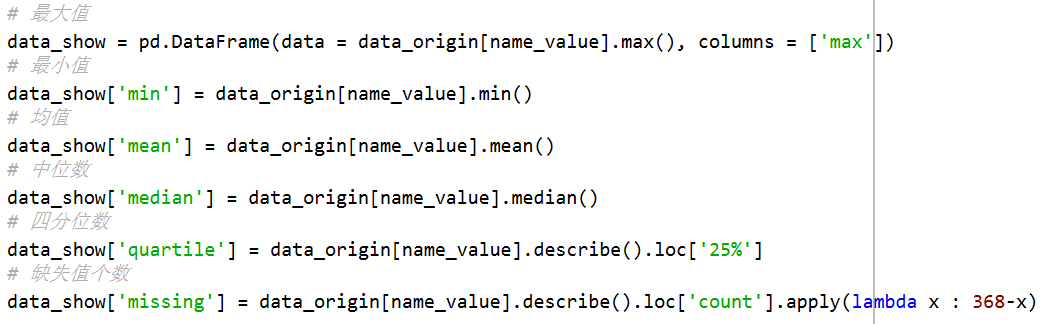
cp\_data 的频数为：

2 244

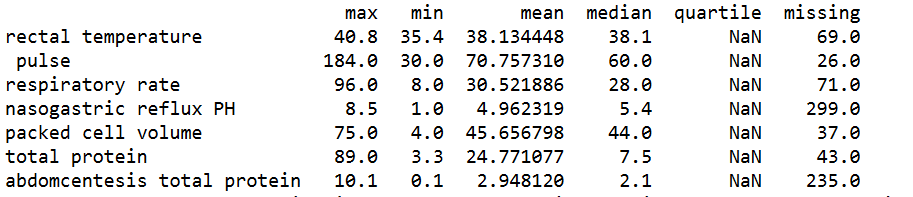
1. 124

**2.1.2 数值属性**

对于数值属性，给出最大、最小、均值、中位数、四分位数及缺失值的个数：



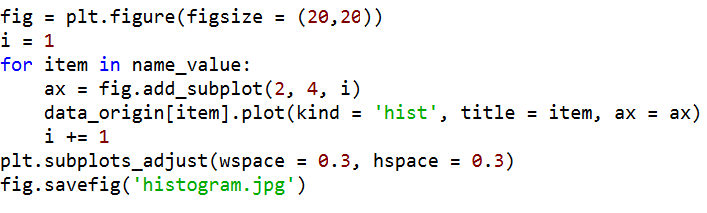
结果如下：

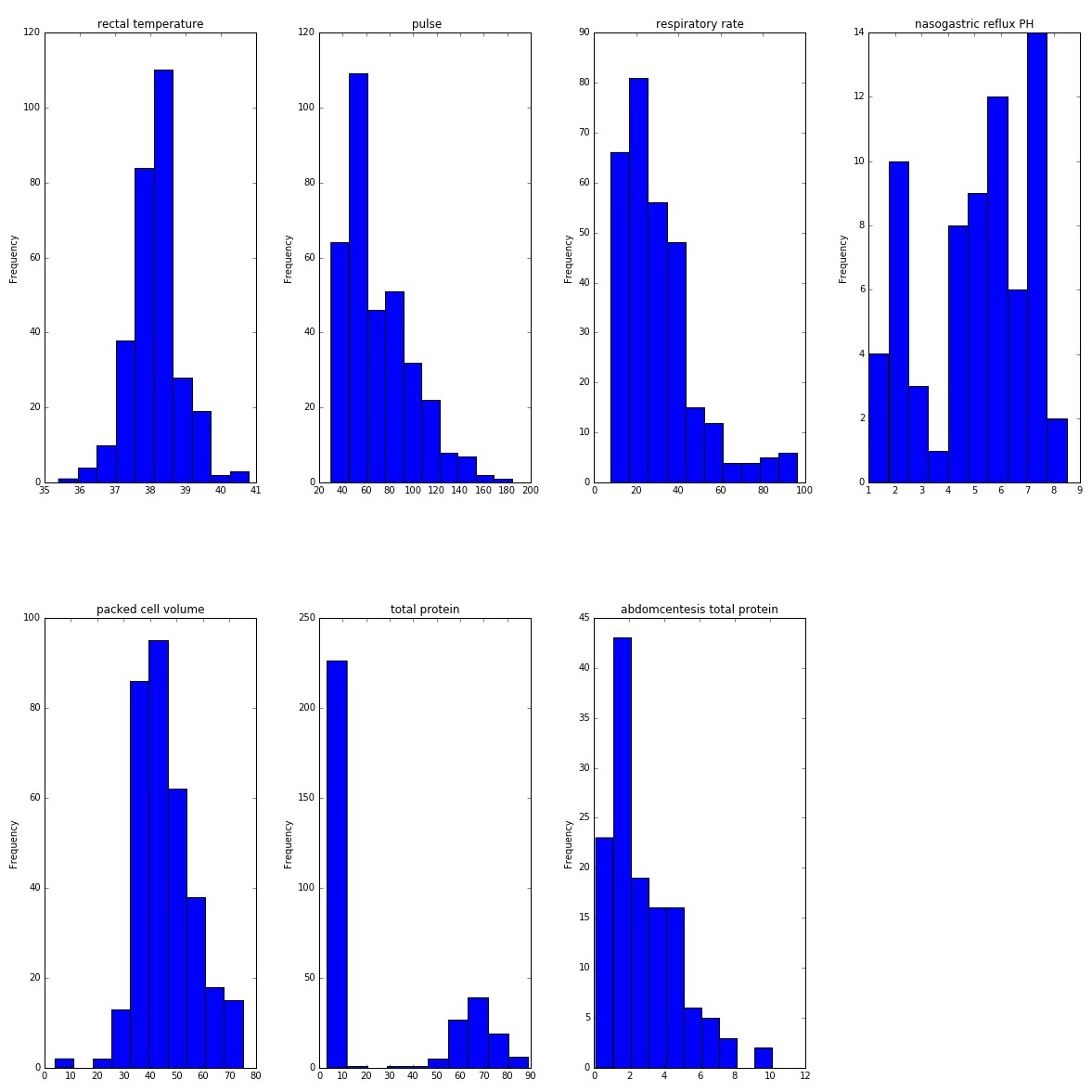


**2.2 数据可视化**

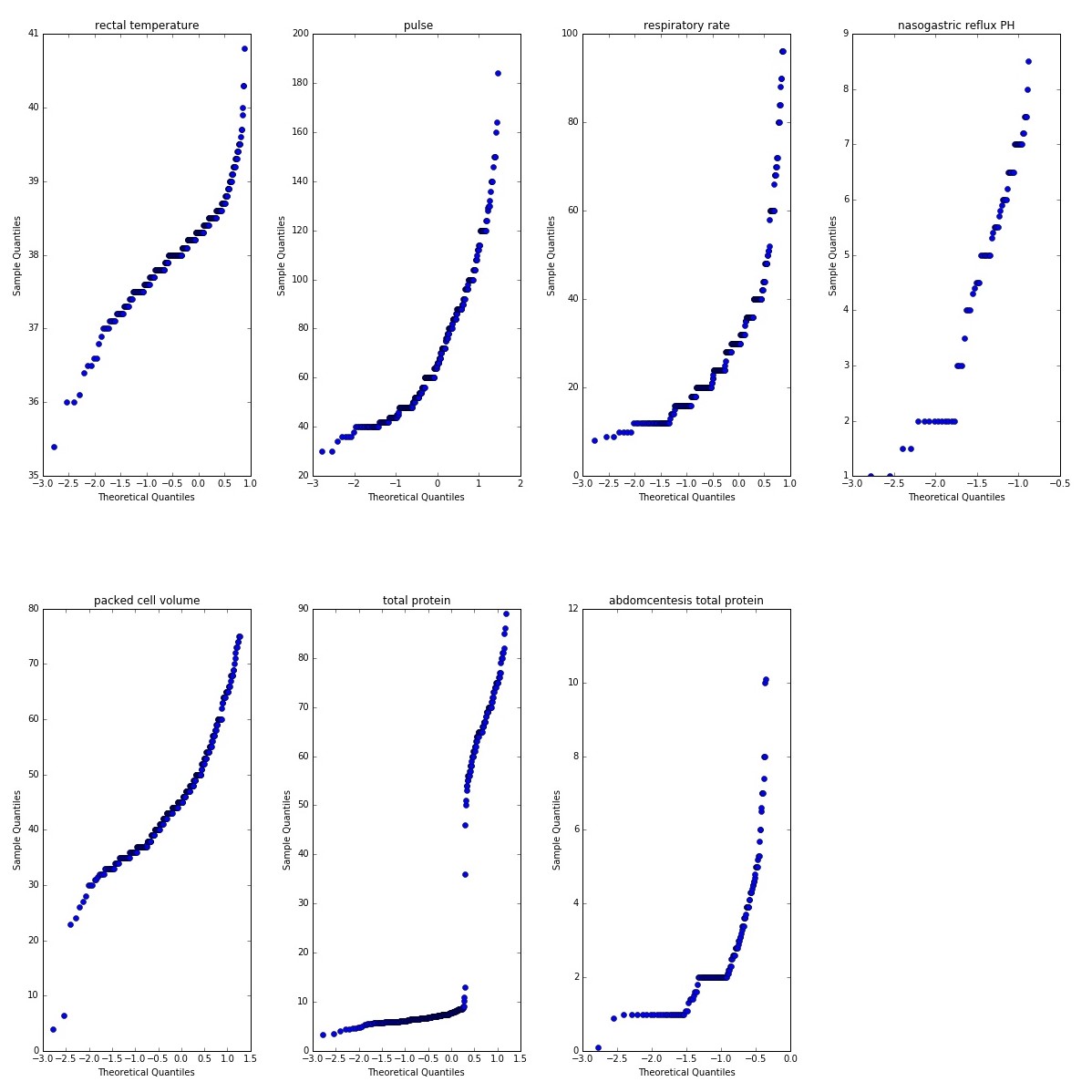
**2.2.1 直方图和qq图**

绘制直方图：





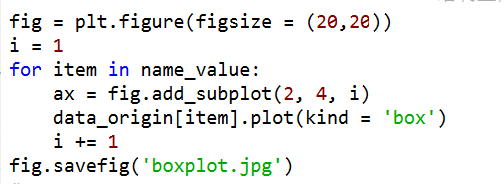
绘制qq图，并且根据qq图判断是否为正态分布：

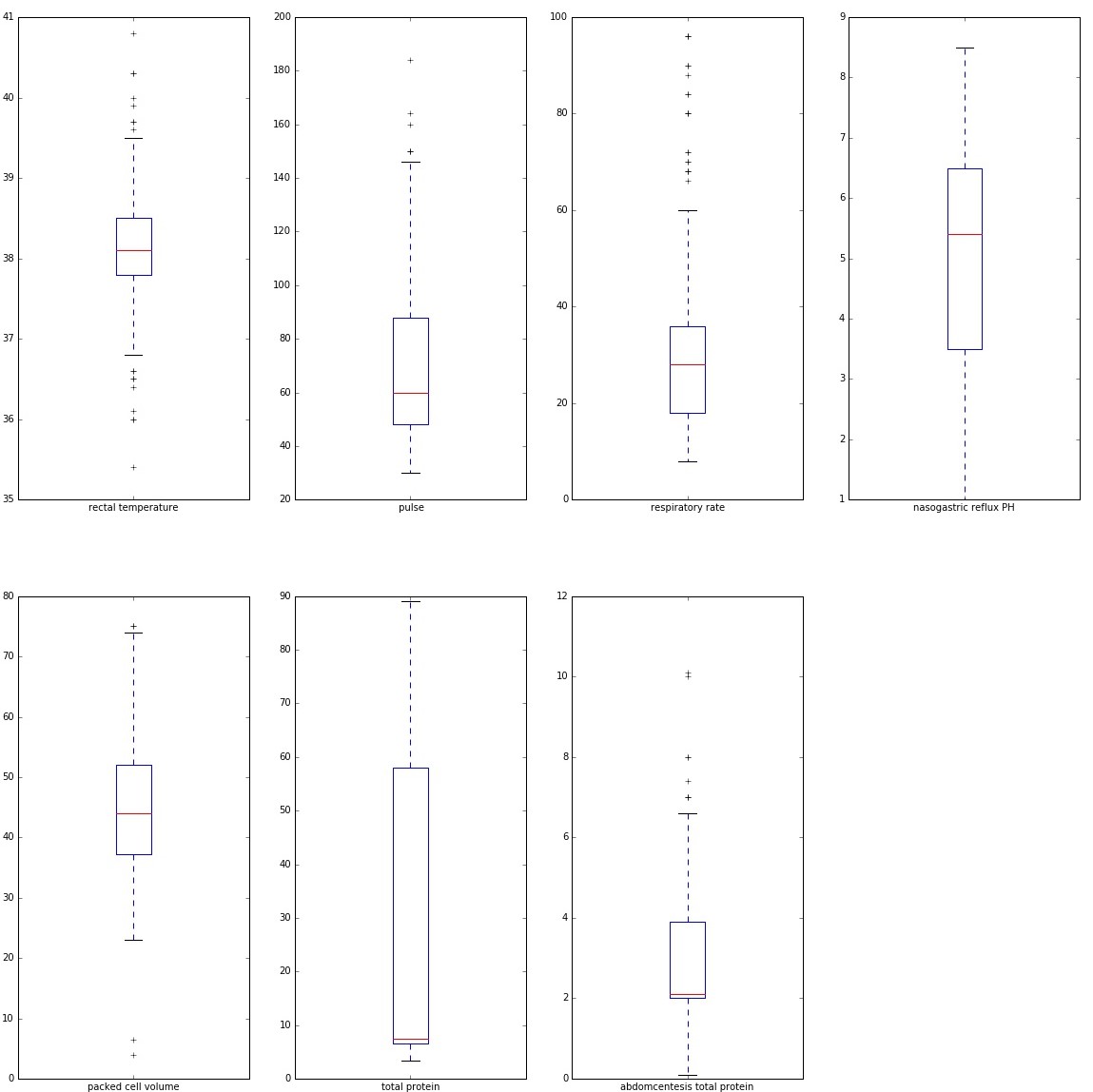


通过qq图可以看到rectal temperature属性和packed volume属性满足正态分布。

**2.2.2 盒图**

绘制盒图，对离群值进行识别：





**三 处理缺失数据**

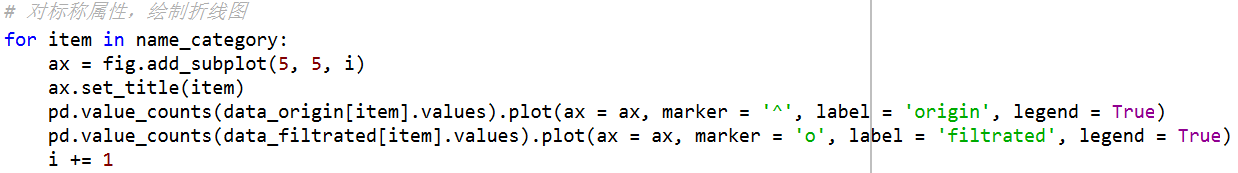
**3.1 将缺失部分剔除**

使用dropna函数，将有缺失值的数据整条删除，得到新的数据集：

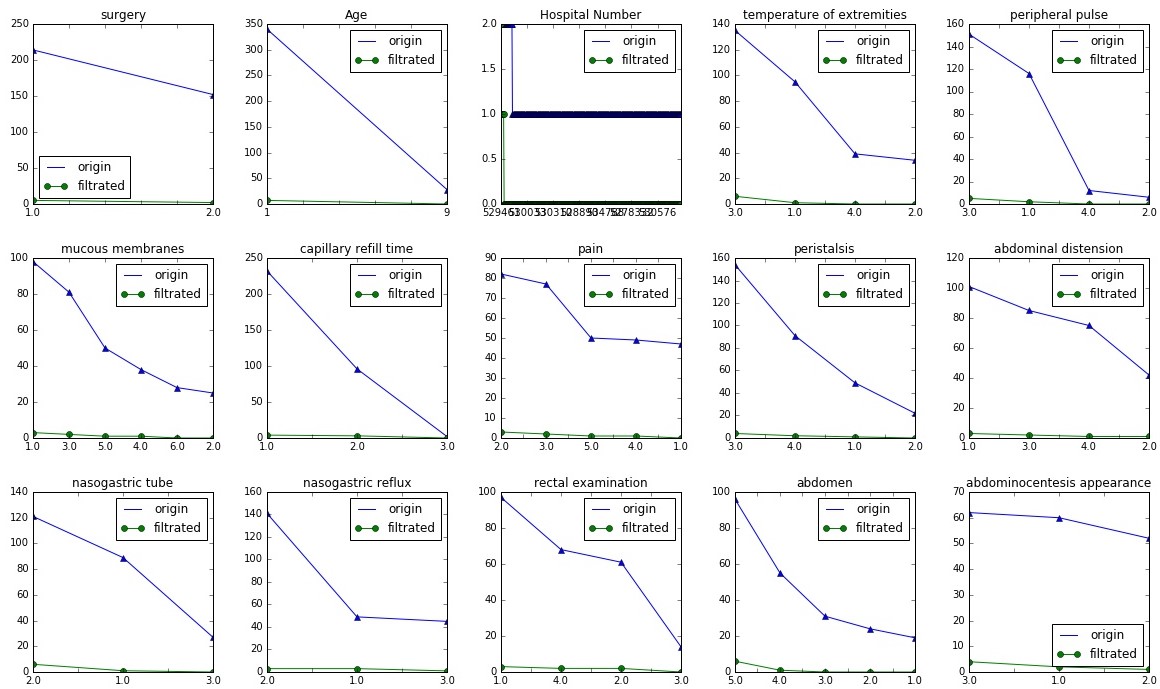
C:\Users\Roach\AppData\Roaming\Tencent\Users\1844977966\QQ\WinTemp\RichOle\U8I4_T_7$T}1JCZ`KYQU62L.png

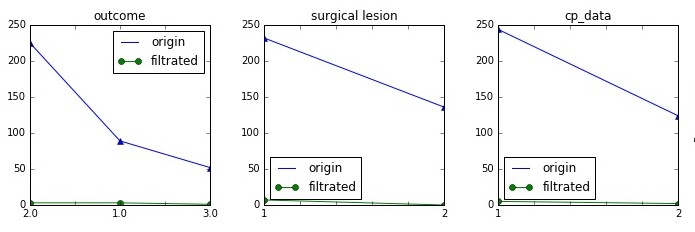
然后可视化的对新旧数据集进行比较。

通过折线图比较新旧数据集的标称属性：

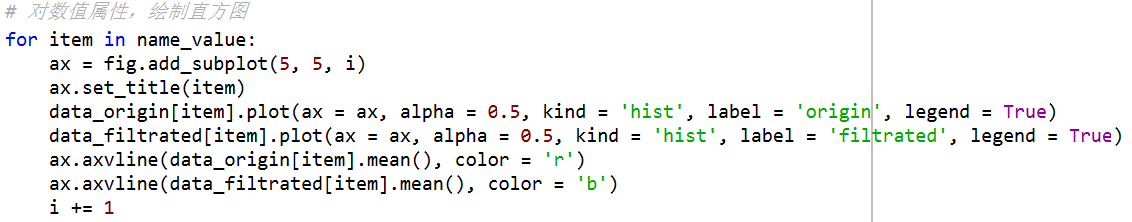


通过折线图可以看到，新的数据集数据量明显减少：

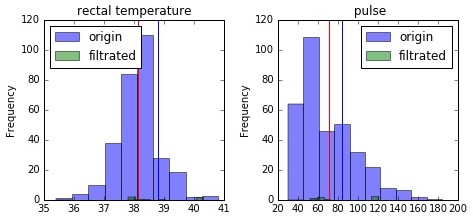


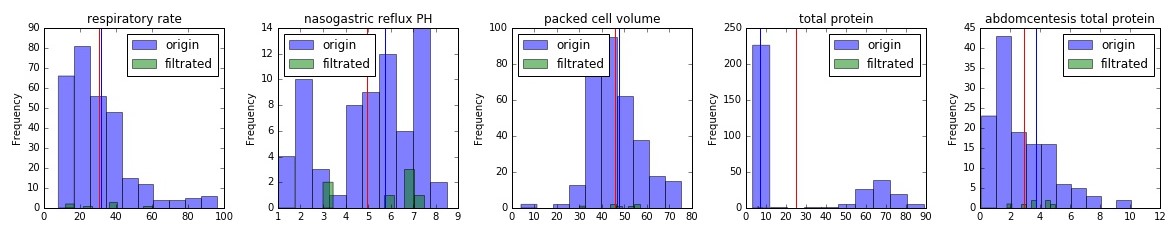


通过直方图图比较新旧数据集的标称属性：



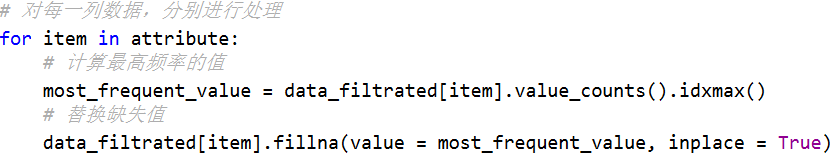
在直方图中，红色的垂线是旧数据集的均值，蓝色的垂线是新数据集的均值，可以看到，虽然新旧数据集在样本数目上差别很大，但是除了total protein属性之外，其余属性的数据分布非常相似，而且均值变化不大





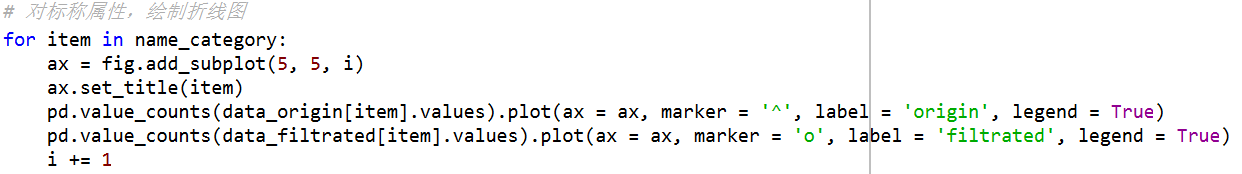
**3.2 用最高频率值来填补缺失值**

分别对每个属性列进行处理，首先找到每个属性中出现次数最多的值，再用这个值填充这个属性中所有的缺失值：

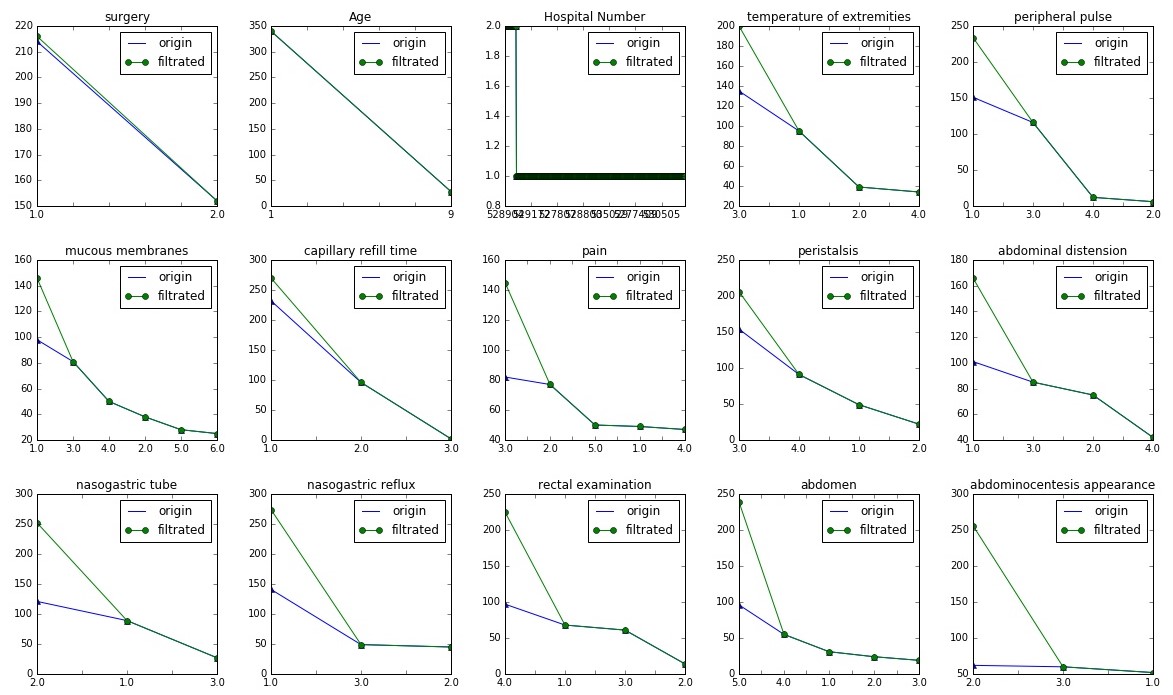


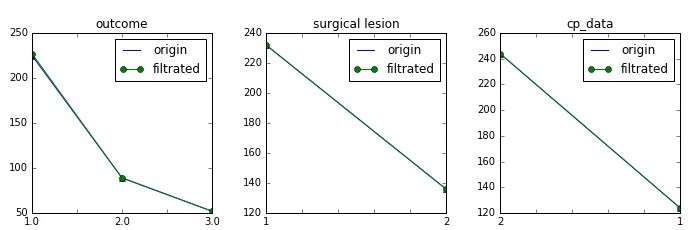
然后可视化的对新旧数据集进行比较。

通过折线图比较新旧数据集的标称属性：

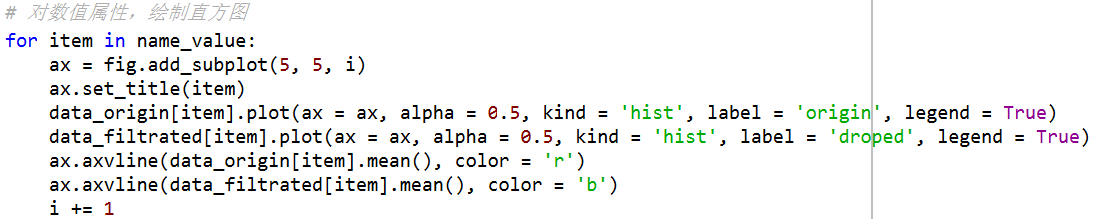


通过折线图可以看到，只有surgery、Age、Hospital Number、outcome、surgical lesion、cp\_data这几个属性在新旧数据集上大概相同，其余的新数据上的数量都明显大于旧数据集，说明这些属性缺失比较严重：

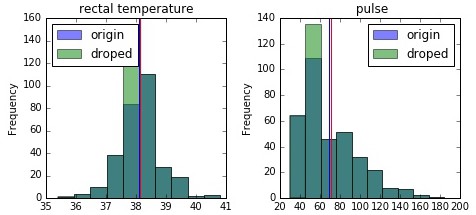


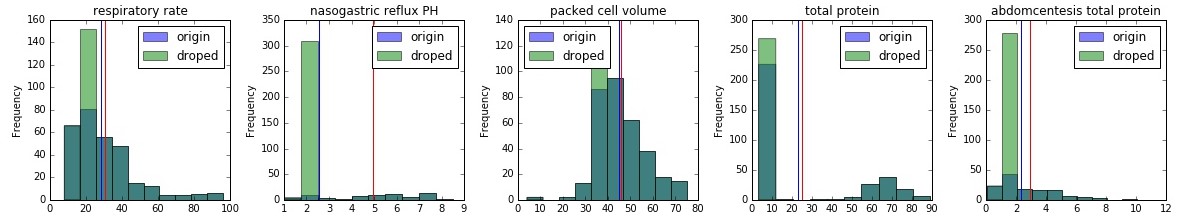


通过直方图图比较新旧数据集的标称属性：



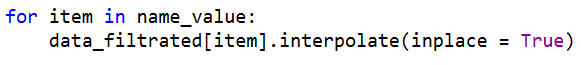
从直方图中可以看出，补全的都是出现最高频率的值，并且除了属性nasogastric reflux PH，其余的属性的均值都几乎保持不变：



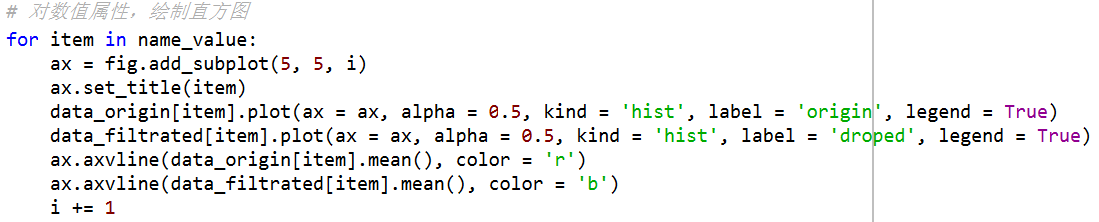


**3.3 通过属性的相关关系来填补缺失值**

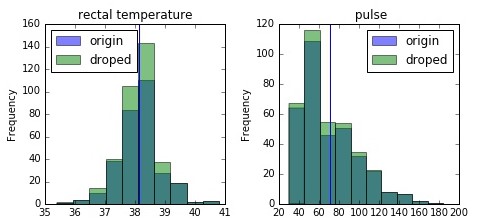
使用pandas中的interpolate()函数，分别对每个数值属性进行插值计算，并用得到的插值填充缺失值：

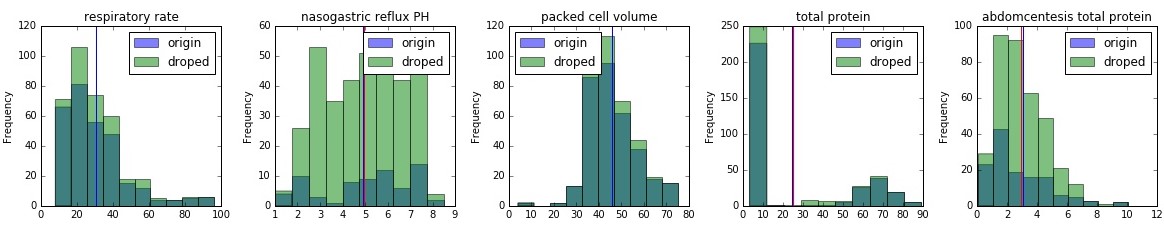


因为只填充了数值属性的缺失值，故我们只通过数值属性的直方图来可视化比较新旧数据集：



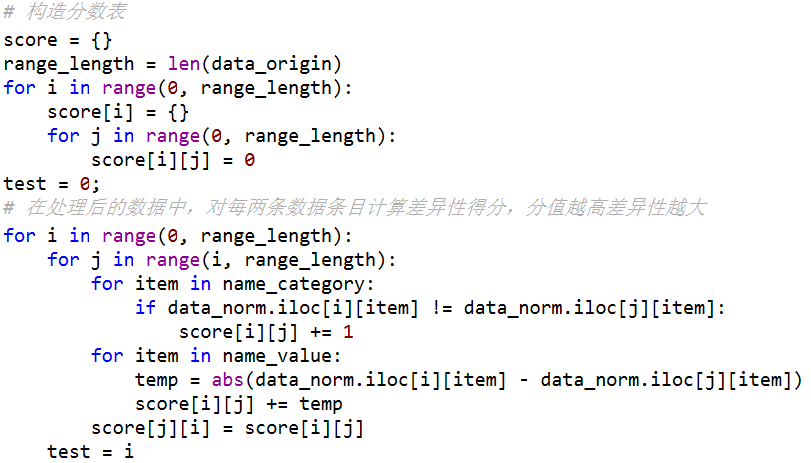
从直方图中可以看出，对于每个数值属性，新的数据集几乎都添加了多个值，数据分布和旧数据集几乎相同，且新旧数据集的均值几乎相等：

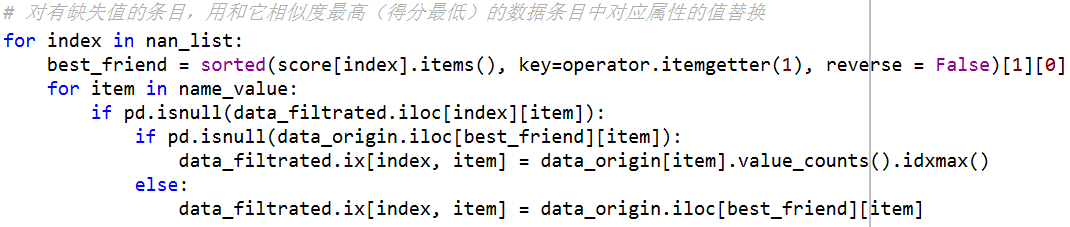




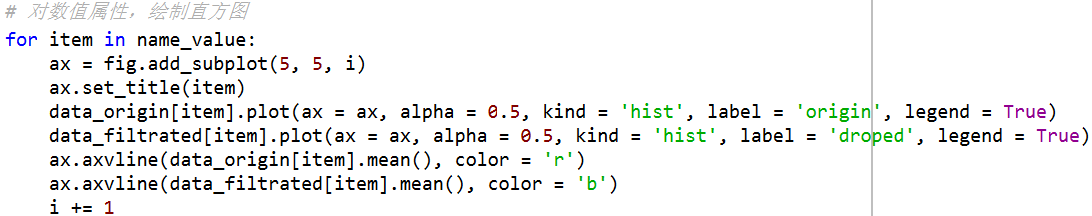
**3.3 通过数据对象之间的相似性来填补缺失值**

计算每两条数据之间的相似度，对每条数据填充缺失值时，选择与其最相似的一条数据的同一属性的值进行填充：





通过数值属性的直方图来可视化比较新旧数据集：



通过直方图可以看出，每个数值属性填充的缺失值大部分都集中在出现频率较高的值上，除了属性nasogastric reflux PH，新旧数据集的均值几乎相同：

