### 数据挖掘课程作业

### 杨雅婷 2120161070

### 作业一：马的疝病分析

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。

1. **数据说明**

共368个样本，27个特征。

<http://archive.ics.uci.edu/ml/datasets/Horse+Colic>

1. **数据分析过程**

* 程序的编写语言是代码分析，使用python3，pandas，numpy，matplotlib库（使用数据分析库pandas进行数据分析，使用matplotlib进行可视化处理）
  1. **数据摘要**

数值属性为：

rectal temperature(4); pulse(5); respiratory rate(6); nasogastric reflux PH(16);

packed cell volume(19);total protein(20);abdomcentesis total protein(22)

其余为对标称属性

* 对标称属性，给出每个可能取值的频数，

surgery 的频数为：

1.0 180

2.0 119

dtype: int64

Age 的频数为：

1 276

9 24

dtype: int64

Hospital Number 的频数为：

529796 2

529424 2

5279822 2

527916 2

530526 2

528996 2

528729 2

528151 2

529461 2

528931 2

528890 2

528469 2

530693 2

532349 2

528904 2

527544 2

533697 1

533696 1

533736 1

534719 1

530101 1

529399 1

533692 1

528570 1

527957 1

534157 1

521399 1

529045 1

530612 1

528047 1

..

533887 1

535031 1

530301 1

5294369 1

530297 1

529272 1

535415 1

530294 1

534899 1

529777 1

535407 1

522979 1

528743 1

534885 1

534857 1

530276 1

535392 1

535338 1

527709 1

527706 1

533847 1

528214 1

535381 1

533750 1

527698 1

530255 1

530254 1

533836 1

530251 1

535043 1

dtype: int64

temperature of extremities 的频数为：

3.0 109

1.0 78

2.0 30

4.0 27

dtype: int64

peripheral pulse 的频数为：

1.0 115

3.0 103

4.0 8

2.0 5

dtype: int64

mucous membranes 的频数为：

1.0 79

3.0 58

4.0 41

2.0 30

5.0 25

6.0 20

dtype: int64

capillary refill time 的频数为：

1.0 188

2.0 78

3.0 2

dtype: int64

pain 的频数为：

3.0 67

2.0 59

5.0 42

4.0 39

1.0 38

dtype: int64

peristalsis 的频数为：

3.0 128

4.0 73

1.0 39

2.0 16

dtype: int64

abdominal distension 的频数为：

1.0 76

3.0 65

2.0 65

4.0 38

dtype: int64

nasogastric tube 的频数为：

2.0 102

1.0 71

3.0 23

dtype: int64

nasogastric reflux 的频数为：

1.0 120

3.0 39

2.0 35

dtype: int64

rectal examination 的频数为：

4.0 79

1.0 57

3.0 49

2.0 13

dtype: int64

abdomen 的频数为：

5.0 79

4.0 43

1.0 28

2.0 19

3.0 13

dtype: int64

abdominocentesis appearance 的频数为：

2.0 48

3.0 46

1.0 41

dtype: int64

outcome 的频数为：

1.0 178

2.0 77

3.0 44

dtype: int64

surgical lesion 的频数为：

1 191

2 109

dtype: int64

lesion 1 的频数为：

0 56

3111 33

3205 29

2208 20

2205 13

4205 11

2209 11

2124 9

1400 8

31110 7

7111 7

2113 6

2112 5

400 5

3209 4

4300 4

2206 4

5400 4

3112 3

4124 3

2111 3

2207 3

7209 3

5206 2

5124 2

3124 2

5111 2

9400 2

6111 2

2322 2

..

3025 2

8400 2

6112 2

11300 1

4122 1

7113 1

6209 1

3115 1

5000 1

3133 1

4111 1

3400 1

300 1

12208 1

9000 1

5205 1

1111 1

1124 1

8300 1

2305 1

4206 1

4207 1

21110 1

2300 1

3207 1

11400 1

7400 1

3113 1

3300 1

41110 1

dtype: int64

lesion 2 的频数为：

0 293

3111 3

6112 1

7111 1

1400 1

3112 1

dtype: int64

lesion 3 的频数为：

0 299

2209 1

dtype: int64

cp\_data 的频数为：

2 201

1 99

dtype: int64

* 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。

max min mean median quartile \

rectal temperature 40.8 35.4 38.167917 38.20 37.8

pulse 184.0 30.0 71.913043 64.00 48.0

respiratory rate 96.0 8.0 30.417355 24.50 18.5

nasogastric reflux PH 7.5 1.0 4.707547 5.00 3.0

packed cell volume 75.0 23.0 46.295203 45.00 38.0

total protein 89.0 3.3 24.456929 7.50 6.5

abdomcentesis total protein 10.1 0.1 3.019608 2.25 2.0

missing

rectal temperature -40.0

pulse -76.0

respiratory rate -42.0

nasogastric reflux PH 147.0

packed cell volume -71.0

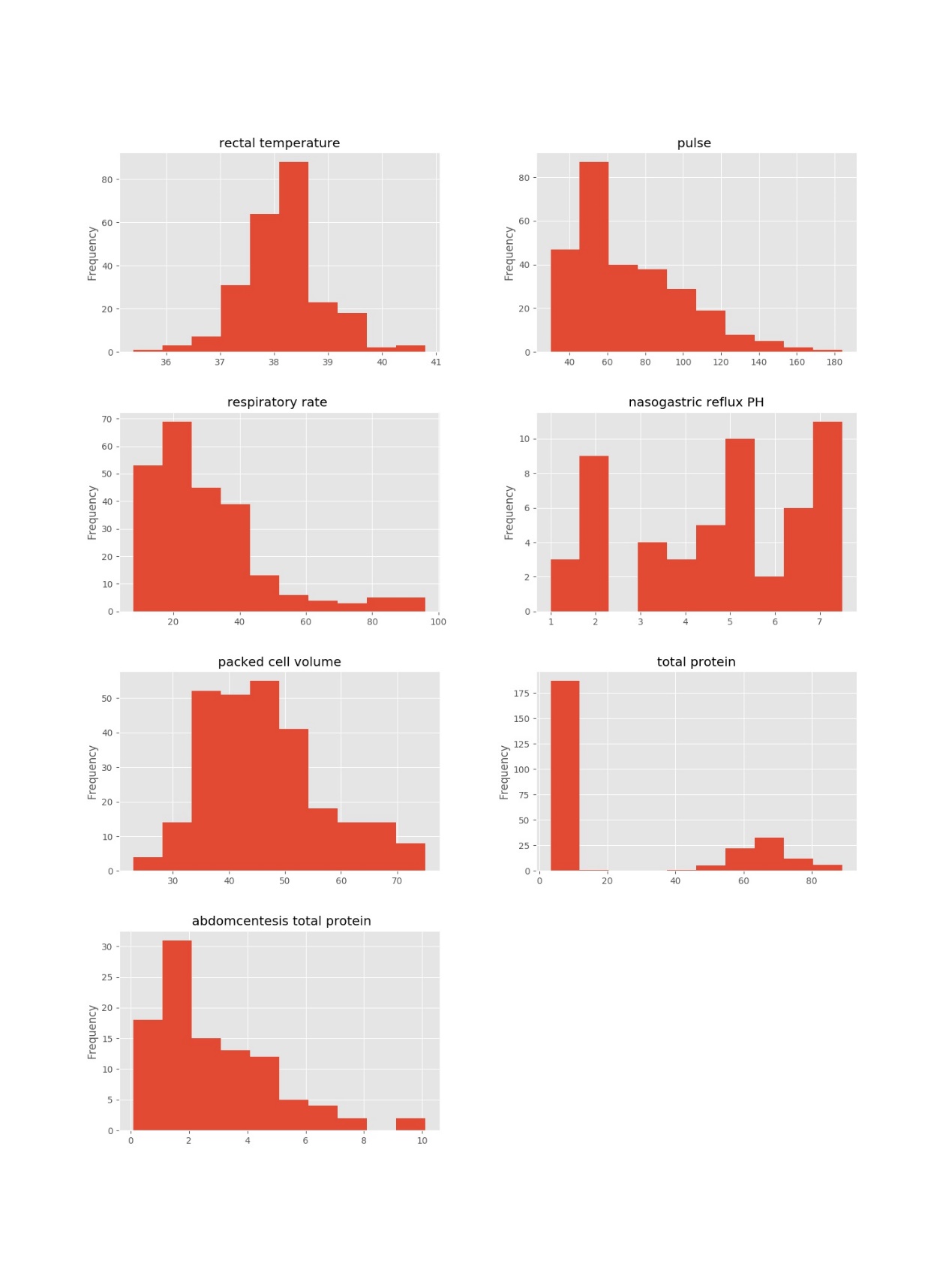
total protein -67.0

abdomcentesis total protein 98.0

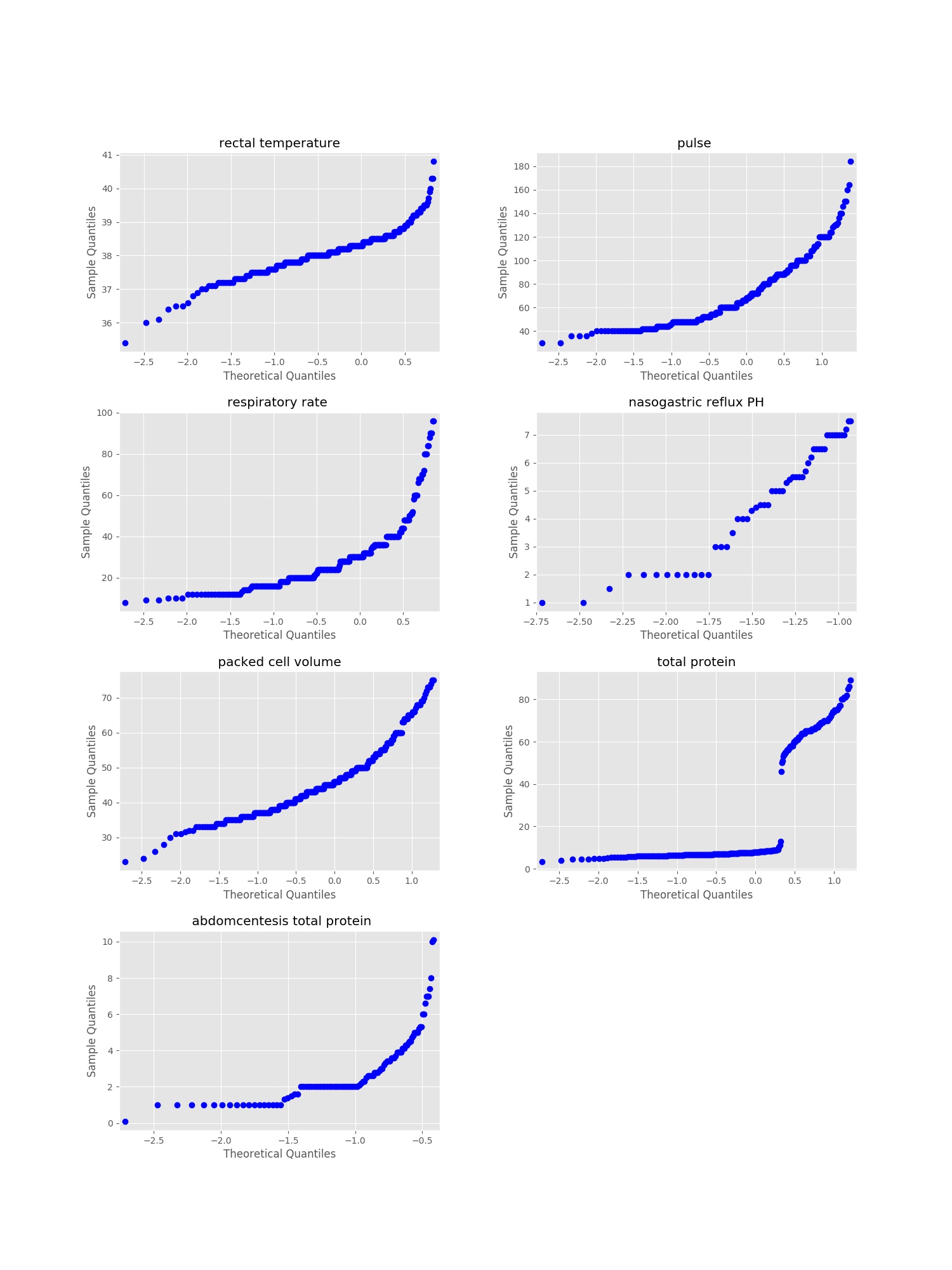
**3.2 数据的可视化**

针对数值属性，绘制

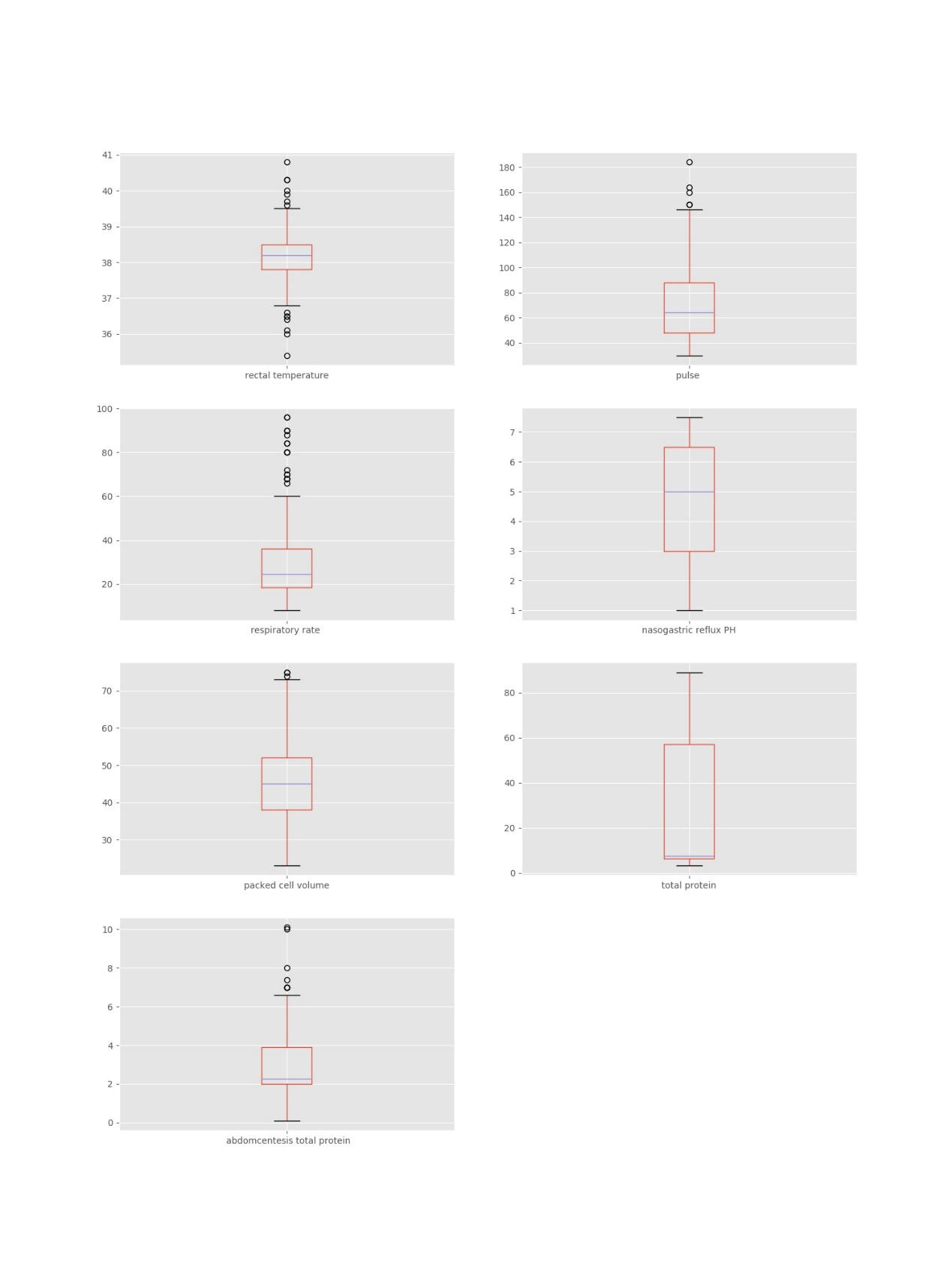
* 直方图



* 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:

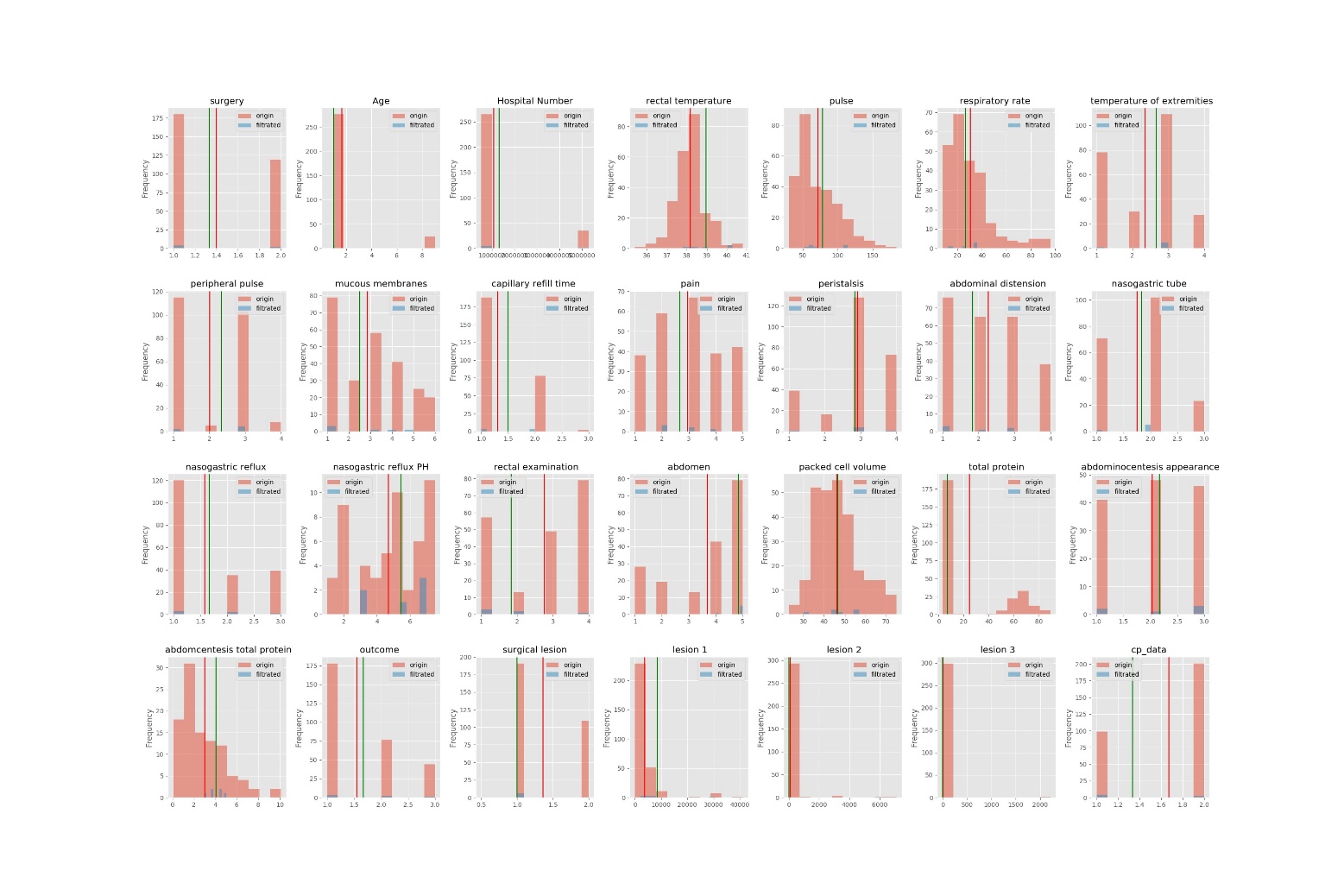
* 将缺失部分剔除
* 用最高频率值来填补缺失值
* 通过属性的相关关系来填补缺失值
* 通过数据对象之间的相似性来填补缺失值

处理后，可视化地对比新旧数据集。

1. **数据缺失的处理：**将缺失部分剔除。找出含有缺失值的数据条目索引值，使用dropna()函数操作删除缺失值。将缺失值对应的数据整条剔除，生成新数据集

nan\_list = pd.isnull(data\_horse).any(1).nonzero()[0]

data\_filtrated = data\_horse.dropna()



1. **数据缺失的处理：** 用最高频率值来填补缺失值。使用value\_counts()函数统计原始数据中，出现频率最高的值，再用fillna()函数将缺失值替换为最高频率值。

# 建立原始数据的拷贝

data\_filtrated = data\_horse.copy()

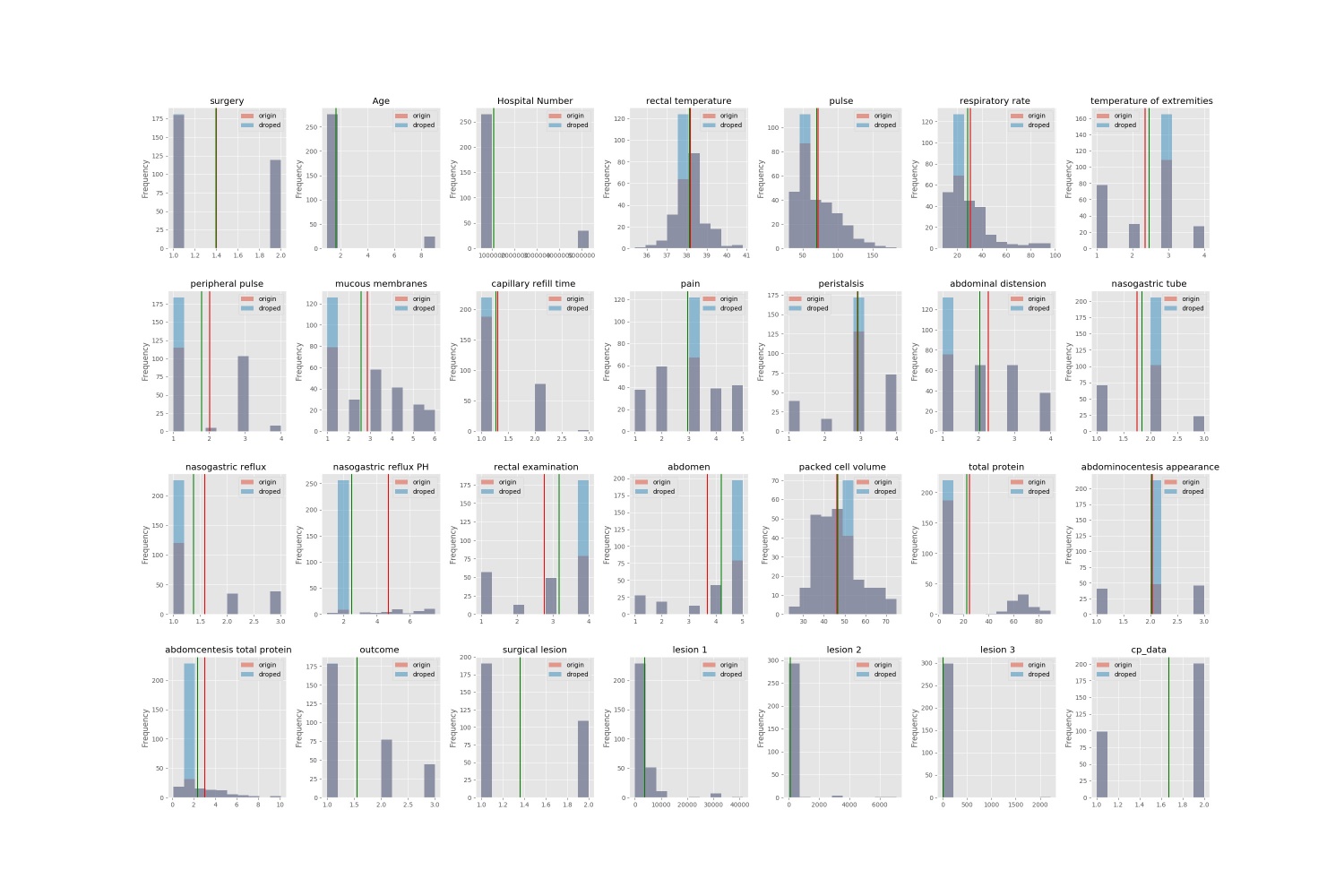
for item in attribute:

# 计算最高频率的值

most\_frequent\_value = data\_filtrated[item].value\_counts().idxmax()

# 替换缺失值

data\_filtrated[item].fillna(value = most\_frequent\_value, inplace = True)



1. **数据缺失的处理：**通过属性的相关关系来填补缺失值。使用pandas中Series的interpolate()函数，对数值属性进行插值计算，并替换缺失值。

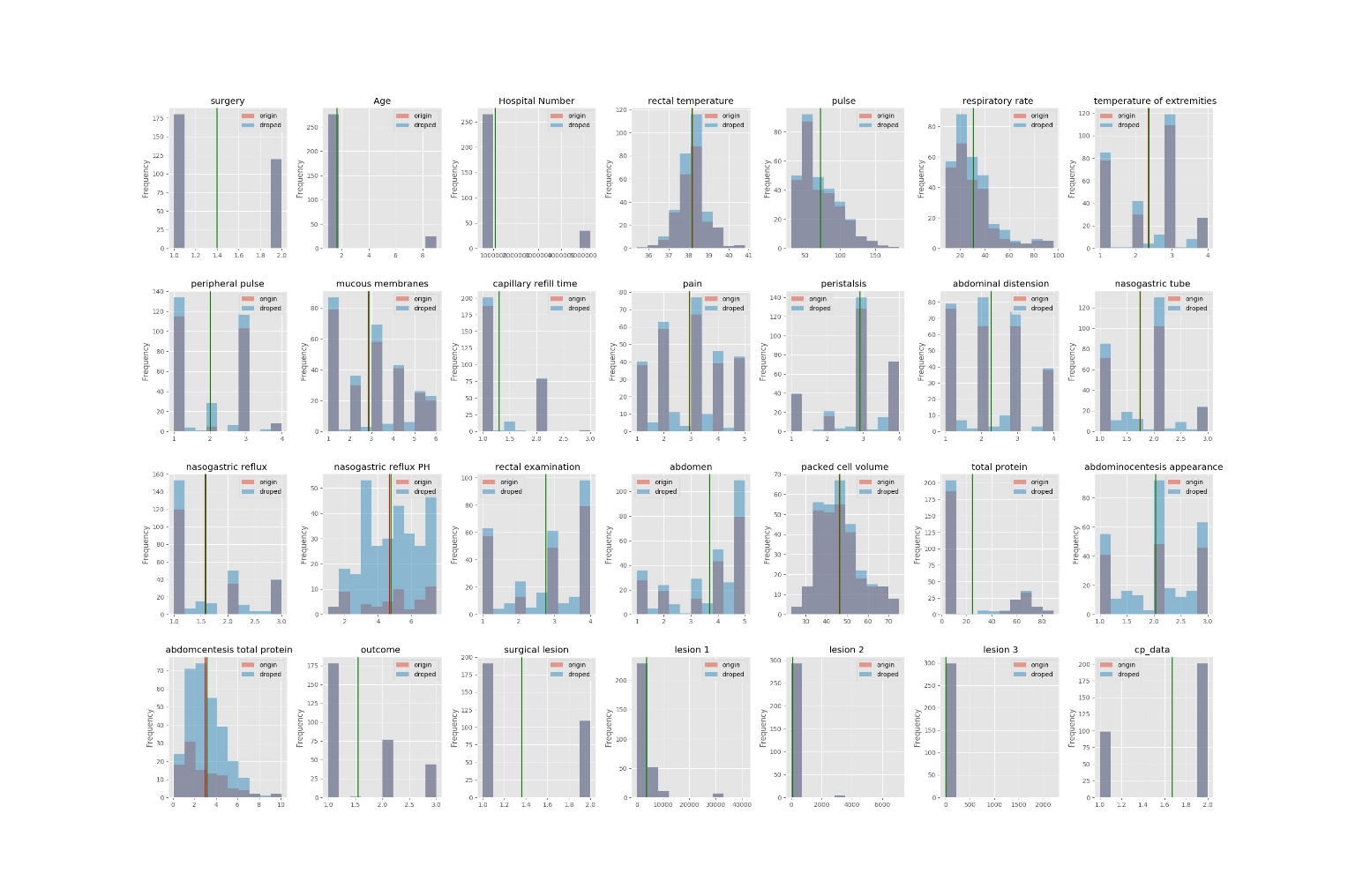
# 建立原始数据的拷贝

data\_filtrated = data\_horse.copy()

#进行插值运算

for item in attribute:

data\_filtrated[item].interpolate(inplace = True)



1. **数据缺失的处理：**通过数据对象之间的相似性来填补缺失值。首先将缺失值设为0，对数据集进行正则化。然后对每两条数据进行差异性计算（分值越高差异性越大）。计算标准为：标称数据不相同记为1分，数值数据差异性分数为数据之间的差值。在处理缺失值时，找到和该条数据对象差异性最小（分数最低）的对象，将最相似的数据条目中对应属性的值替换缺失值。

# 建立原始数据的拷贝，用于正则化处理

data\_norm = data\_horse.copy()

data\_norm[attribute] = data\_norm[attribute].fillna(0)

data\_norm[attribute] = data\_norm[attribute].apply(lambda x : (x - np.mean(x)) / (np.max(x) - np.min(x)))

# 构造分数表

score = {}

range\_length = len(data\_horse)

for i in range(0, range\_length):

score[i] = {}

for j in range(0, range\_length):

score[i][j] = 0

# 对每两条数据条目计算差异性得分，分值越高差异性越大

for i in range(0, range\_length):

for j in range(i, range\_length):

for item in attribute:

temp = abs(data\_norm.iloc[i][item] - data\_norm.iloc[j][item])

score[i][j] += temp

score[j][i] = score[i][j]

data\_filtrated = data\_horse.copy()

# 对有缺失值的条目，用和它相似度最高的数据条目中对应属性的值替换

for index in nan\_list:

best\_friend = sorted(score[index].items(), key=operator.itemgetter(1), reverse = False)[1][0]

for item in attribute:

if pd.isnull(data\_filtrated.iloc[index][item]):

if pd.isnull(data\_horse.iloc[best\_friend][item]):

data\_filtrated.ix[index, item] = data\_horse[item].value\_counts().idxmax()

else:

data\_filtrated.ix[index, item] = data\_horse.iloc[best\_friend][item]

