马的疝病分析——数据分析

学院：计算机学院　　姓名：康丽琪　　学号：2120161006

简要说明

对于本作业，我用 python 做了两种方法的处理，其中第二种方法（dataAnalysis_2.py）是在利用 pandas 这个库之后，对第一种方法（dataAnalysis_1.py）进行了改进和完善。这种方法大大简化了处理过程，降低了处理难度。报告中呈现的结果为两种方法的综合处理结果。

一、数据的基本分析

这份数据共 368 个样本，28 个特征，其中：

1）标称属性（特征序号）：

1, 2, 3, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 21, 23, 24, 25, 26, 27, 28

2）数值属性（特征序号）：

4, 5, 6, 16, 19, 20, 22

即共有 21 个标称属性，7 个数值属性。

在方法 1 中，直接对 horse-colic-all.data 进行处理。

在方法 2 中，因为需要使用 read_csv()方法，所以先将数据间的空格全部转换为逗号，为了便于处理，将表格中的缺失值全部换成 np.nan。

二、数据处理

1）数据摘要

对于标称属性：

在方法 1 中，直接利用字典进行统计。

在方法 2 中，读入数据为 DataFrame 实例，取出一列为 Series 实例。将 Series 对象转换成 list 对象，使用 count()方法即可。

对于数值属性：

可使用 Series 对象的 max,min,median,quantile 等方法。

对于标称属性，每个可能取值的频数如下：

surgery：

{'1': 214, '2': 152}

Age：

{'1': 340, '9': 28}

Hospital Number：

{'528298': 1, '528299': 1, '528742': 1, '528743': 1, '527698': 1, '529840': 1, '527940': 1, '530170': 1, '529849': 1, '5279821': 1, '534156': 1, '528268': 1,

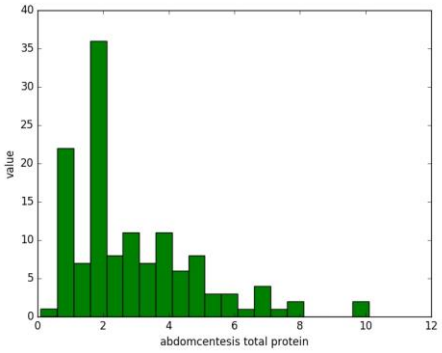
'535364': 1, '534885': 1, '534163': 1, '528904': 2, '528047': 1, '530612': 1, '528461': 1, '533983': 1, '528452': 1, '528469': 2, '5279441': 1, '5279442': 1, '528630': 1, '530310': 1, '529483': 1, '528638': 1, '530251': 1, '530319': 1, '530255': 1, '530254': 1, '530505': 1, '528999': 1, '534756': 1, '528996': 2, '529340': 1, '528682': 1, '535085': 1, '534644': 1, '534833': 1, '529930': 1, '5281441': 1, '521399': 1, '5290481': 1, '534429': 1, '534938': 1, '5290482': 1, '530381': 1, '534933': 1, '530478': 1, '5290409': 1, '535407': 1, '5278331': 1, '526639': 1, '535338': 1, '535337': 1, '527494': 1, '535330': 1, '530033': 1, '530034': 1, '527365': 1, '535240': 1, '535246': 1, '527563': 1, '534111': 1, '534115': 1, '534053': 1, '5299603': 1, '528310': 1, '527709': 1, '534753': 1, '534618': 1, '528959': 1, '534615': 1, '527702': 1, '528702': 1, '527706': 1, '528179': 1, '528178': 1, '528890': 2, '530624': 1, '533968': 1, '530544': 1, '5299049': 1, '529475': 1, '5287279': 1, '529478': 1, '528006': 1, '533723': 1, '529615': 1, '530431': 1, '5282839': 1, '530439': 1, '530438': 1, '535054': 1, '529685': 1, '535196': 1, '530297': 1, '5301219': 1, '529272': 1, '535176': 1, '529296': 1, '534475': 1, '527524': 1, '527916': 2, '534478': 1, '5294539': 1, '534579': 1, '5299253': 1, '534572': 1, '533793': 1, '528214': 1, '529991': 1, '530002': 1, '530001': 1, '528355': 1, '534719': 1, '534899': 1, '534898': 1, '528919': 1, '528031': 1, '533928': 1, '5274919': 2, '533847': 1, '534626': 1, '534624': 1, '534157': 1, '5288249': 1, '527677': 1, '534092': 1, '530301': 1, '529498': 1, '528626': 1, '529493': 1, '528620': 1, '5291409': 1, '530576': 1, '530670': 2, '527664': 1, '528134': 1, '529729': 1, '529925': 1, '529663': 1, '529888': 1, '529667': 1, '535130': 1, '535137': 1, '530239': 2, '529399': 1, '527957': 1, '527950': 1, '530233': 1, '534925': 1, '530402': 1, '534922': 1, '5292489': 1, '526090': 1, '5294369': 1, '529812': 1, '527829': 1, '534324': 1, '534280': 1, '534998': 1, '534268': 1, '5297379': 1, '527734': 1, '529865': 1, '530051': 1, '529373': 1, '528369': 1, '5291329': 2, '528248': 1, '528247': 1, '534197': 1, '528067': 1, '5275211': 1, '5275212': 1, '529597': 1, '533871': 1, '528964': 1, '529567': 1, '5291719': 1, '533738': 1, '534069': 1, '533736': 1, '5279822': 2, '530276': 1, '529960': 1, '534963': 1, '5290759': 1, '535043': 1, '529695': 1, '5277409': 1, '530526': 2, '528668': 1, '529135': 1, '534491': 1, '535263': 1, '523190': 1, '5292929': 1, '534497': 1, '534790': 1, '534403': 1, '527518': 1, '534817': 1, '529628': 1, '529629': 1, '529183': 1, '535314': 1, '5283431': 1, '5262541': 1, '5262543': 1, '5262542': 1, '529980': 1, '530157': 1, '529821': 1, '529827': 1, '527463': 1, '527465': 1, '535381': 1, '528382': 1, '534135': 1, '528926': 2, '528433': 1, '533836': 1, '529528': 1, '527642': 1, '526802': 1, '527883': 1, '534145': 1, '528729': 2, '528151': 2, '530334': 1, '528653': 1, '530561': 1, '533942': 1, '529796': 2, '532110': 1, '529777': 1, '529893': 1, '528812': 1, '534857': 1, '529388': 1, '529386': 1, '5297159': 1, '534523': 1, '532349': 2, '534917': 1, '529703': 1, '535031': 1, '5305629': 1, '530384': 1, '5287179': 1, '527544': 2, '534293': 1, '530294': 1, '527933': 1, '534597': 1, '529172': 1, '528800': 1, '528804': 1, '5299629': 1, '530401': 1, '534183': 1, '533889': 1, '533887': 1, '533886': 1, '533885': 1, '528977': 1, '534004': 1, '528503': 1, '5278332': 1, '533902': 1, '528872': 1, '528183': 1, '529518': 1, '534073': 1, '528570': 1, '527758': 1, '530366': 1, '530360': 1, '532985': 1, '529736': 1, '530242': 1, '528590': 1, '530693': 2,

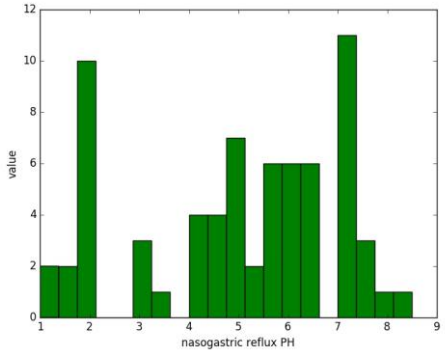
'530695': 1, '5289419': 1, '529428': 1, '529427': 1, '529424': 2, '529642': 1, '529640': 1, '535158': 1, '529126': 1, '534783': 1, '534784': 1, '534787': 1, '534788': 1, '514279': 1, '5281091': 1, '5281092': 1, '534556': 1, '535415': 1, '529045': 1, '530028': 1, '5305129': 1, '527807': 1, '535392': 1, '528931': 2, '534686': 1, '528305': 1, '528523': 1, '528641': 1, '528169': 1, '533954': 1, '529461': 2, '533696': 1, '533697': 1, '533692': 1, '528548': 1, '528019': 1, '533750': 1, '529765': 1, '529764': 1, '529766': 1, '533815': 2, '529607': 1, '530354': 1, '534519': 1, '535029': 1, '529304': 1, '521681': 1, '535166': 1, '518476': 1, '535163': 1, '535208': 1, '530101': 1, '530107': 1, '527929': 1, '527927': 1, '534921': 1, '522979': 1, '529160': 1, '535292': 1}
temperature of extremities：
{'1': 95, '3': 135, '2': 39, '4': 34}
peripheral pulse：
{'1': 151, '3': 116, '2': 6, '4': 12}
mucous membranes：
{'1': 98, '3': 81, '2': 38, '5': 28, '4': 50, '6': 25}
capillary refill time：
{'1': 232, '3': 2, '2': 96}
pain：
{'1': 49, '3': 82, '2': 77, '5': 50, '4': 47}
peristalsis：
{'1': 49, '3': 154, '2': 22, '4': 91}
abdominal distension：
{'1': 101, '3': 85, '2': 75, '4': 42}
nasogastric tube：
{'1': 89, '3': 27, '2': 121}
nasogastric reflux：
{'1': 141, '3': 49, '2': 45}
rectal examination：
{'1': 68, '3': 61, '2': 14, '4': 97}
abdomen：
{'1': 31, '3': 19, '2': 24, '5': 96, '4': 55}
abdominocentesis appearance：
{'1': 52, '3': 60, '2': 62}
outcome：
{'1': 225, '3': 52, '2': 89}
surgical lesion：
{'1': 232, '2': 136}
type of lesion：
{'03400': 1, '05205': 1, '05206': 2, '03133': 1, '05124': 2, '31110': 9, '41110': 1, '09000': 1, '03207': 1, '03205': 35, '02124': 9, '05400': 4, '11300': 1, '03115': 1, '03112': 3, '03113': 2, '03111': 41, '04111': 1, '11400': 1, '06112': 4, '06111': 3, '07400': 1, '04300': 4, '02300': 2, '05111': 3, '05110': 1, '03209': 6, '04122': 1, '01124': 2, '07113': 2, '21110': 1, '07111': 10, '08300': 1, '08400': 2, '04207': 1,

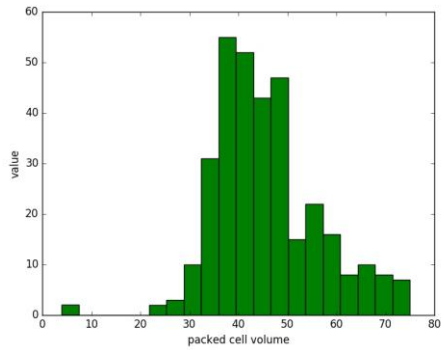
'04206': 3, '04205': 11, '03300': 1, '02322': 2, '00400': 7, '07209': 3, '11124': 2, '08405': 1, '02207': 3, '02206': 5, '02209': 15, '02208': 23, '03025': 2, '02305': 1, '04124': 5, '03124': 4, '01111': 1, '09400': 2, '02111': 4, '02113': 8, '02112': 6, '06209': 1, '00300': 1, '05000': 1, '12208': 1, '00000': 67, '02205': 17, '01400': 10}
type of lesion：
{'07111': 1, '01400': 1, '03205': 2, '00000': 358, '02208': 1, '03111': 3, '03112': 1, '06112': 1}
type of lesion：
{'02209': 1, '00000': 366, '000000': 1}
cp_data：
{'1': 124, '2': 244}
对于数值属性，最大、最小、均值、中位数、四分位数及缺失值的个数如下，其中，括号里的三个数分别为上四分位数，中位数，下四分位数。
rectal temperature：
(40.8, 35.4, 38.13, (39.925, 38.0, 38.2), 38.0, 0)
pulse：
(184.0, 30.0, 70.76, (98.0, 62.0, 64.5), 62.0, 0)
respiratory rate：
(96.0, 8.0, 30.52, (32.0, 24.0, 24.5), 24.0, 0)
nasogastric reflux PH：
(8.5, 1.0, 4.96, (6.49, 4.96, 2.74), 4.96, 0)
packed cell volume：
(75.0, 4.0, 45.66, (53.25, 43.5, 40.25), 43.5, 0)
total protein：
(89.0, 3.3, 29.82, (7.625, 66.5, 57.8), 66.5, 0)
abdomcentesis total protein：
(10.1, 0.1, 2.95, (4.1125, 1.975, 1.4875), 1.975, 0)

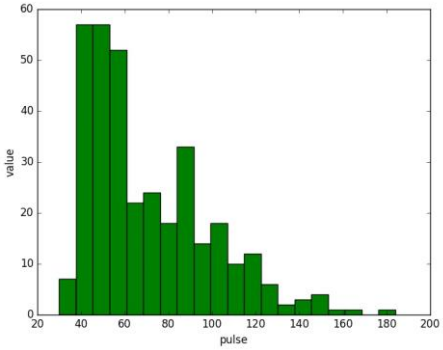
2）数据可视化

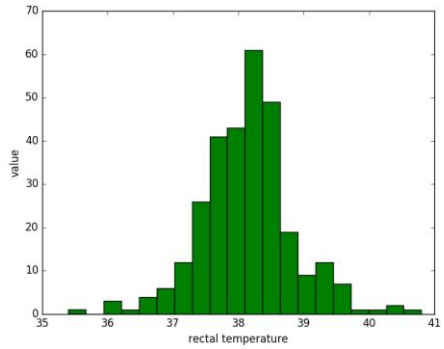
使用 python 的 matplotlib 库作为画图工具。
针对数值属性，绘制直方图，如下：

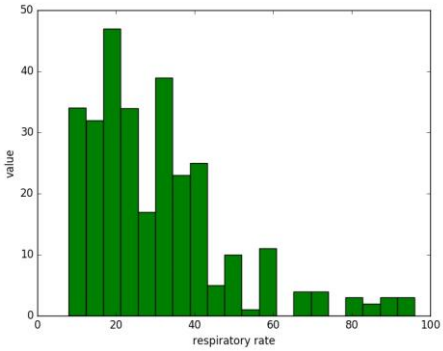


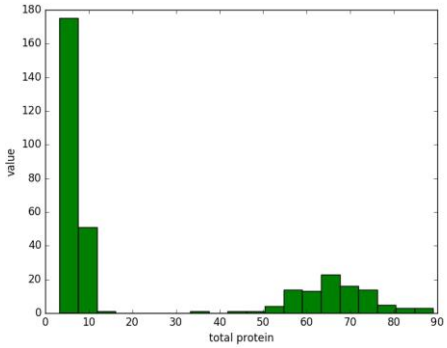




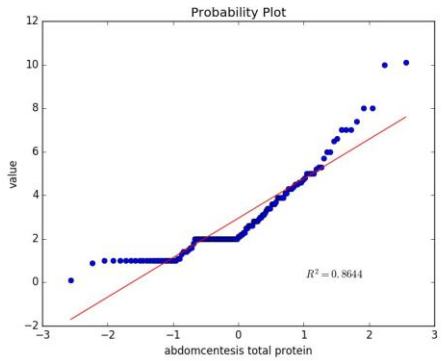


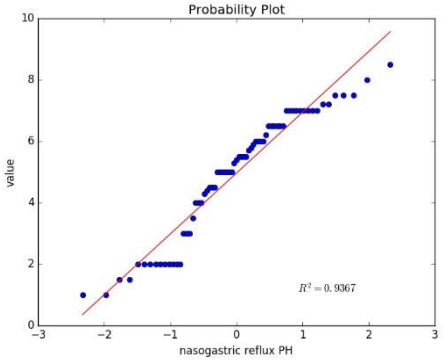


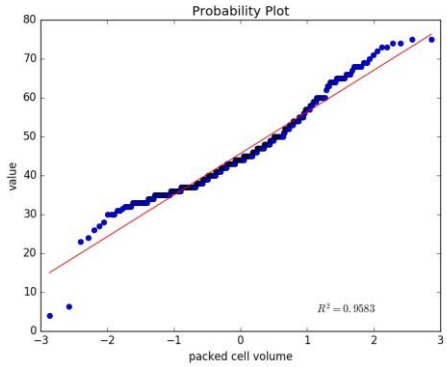


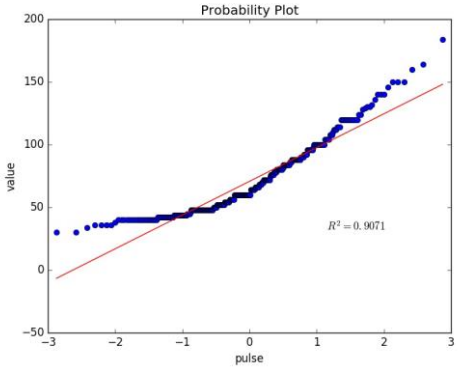


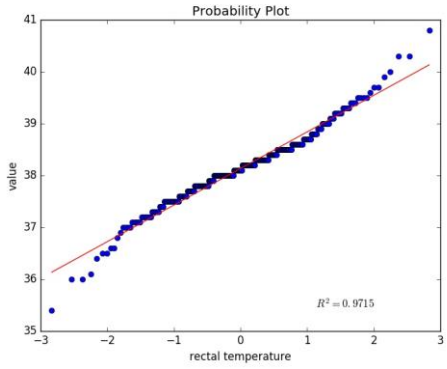
针对数值属性，绘制 qq 图，python 的 SciPy 库中已经实现了画 qq 图的函数，就是 stats.probplot()函数。结果如下：

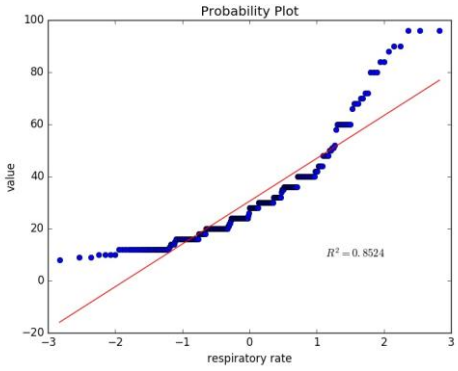


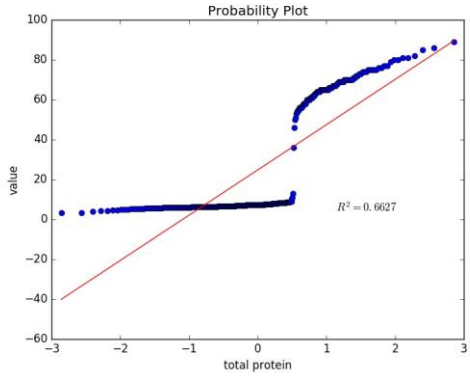




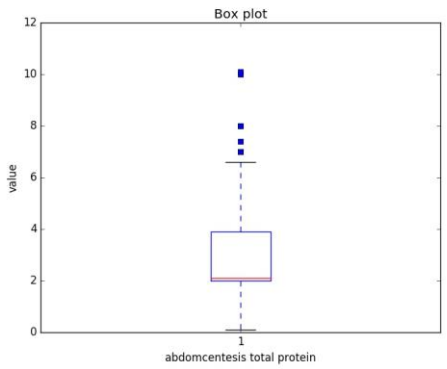


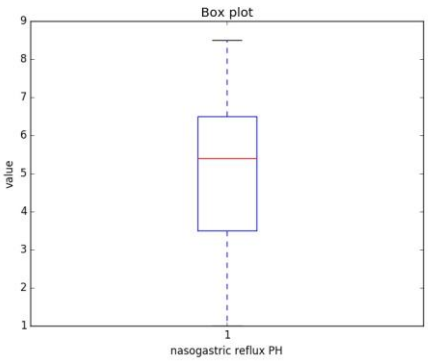


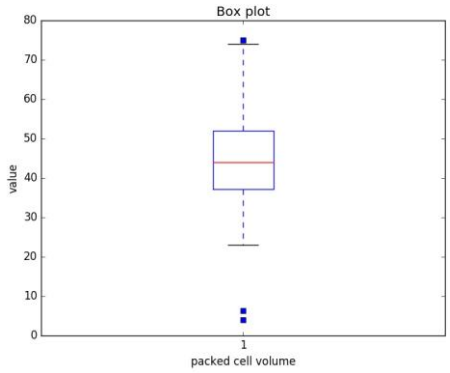


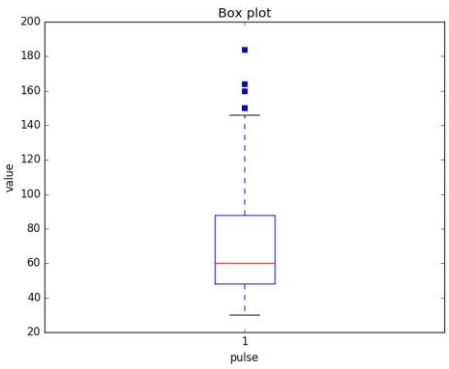


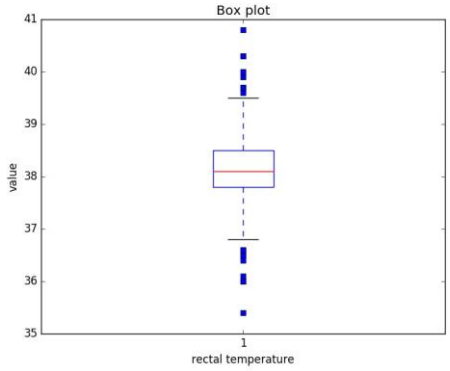
针对数值属性，绘制盒图，使用 matplotlib 库中的 boxplot 函数实现。结果如下：

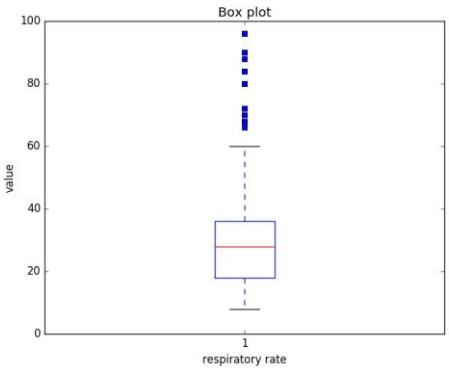


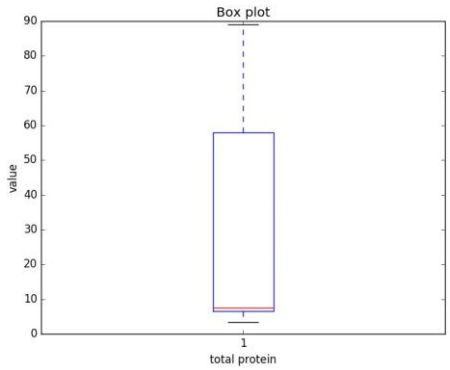












3）数据缺失的处理

受篇幅限制，该部分只展示数据处理之后得到的直方图，另外两种图形可在相应文件下看到。

- 将缺失部分剔除

剔除缺失部分有两种方案，一种是剔除含有缺失数据的样本，另一种是剔除样本中的缺失值。

方法一中使用剔除样本中的缺失值的方法，也就是对每一个属性列分别去掉缺失的部分。上面就是使用这种方法得到的结果。

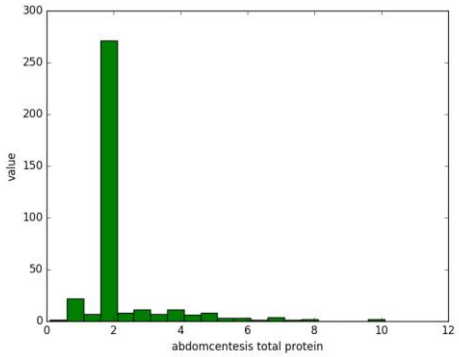
方法二使用剔除含有缺失数据的样本的方法，使用 DataFrame 的 dropna 方法实现。但由于该数据含有的缺失数据较多，剔除之后只剩 7 条记录，效果不理想。

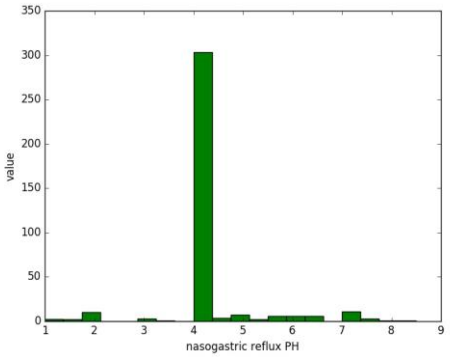
- 用最高频率值来填补缺失值

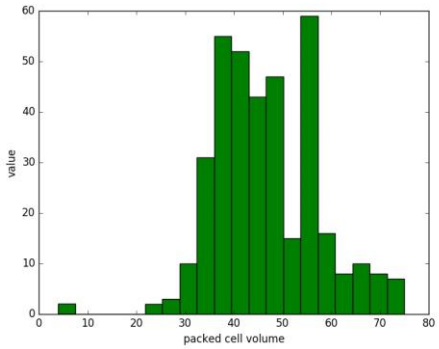
对于标称属性，用出现频率最高的数据来填补。

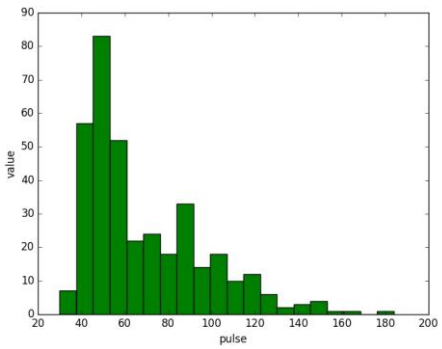
对于数值属性，对于正态分布的数据用均值填补，对偏态分布的数据用中位数填补。

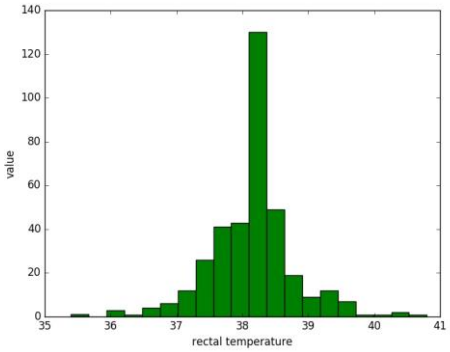
详见方法一中该部分的代码。可视化得到的直方图如下：

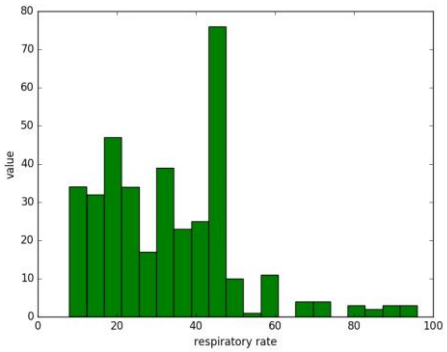


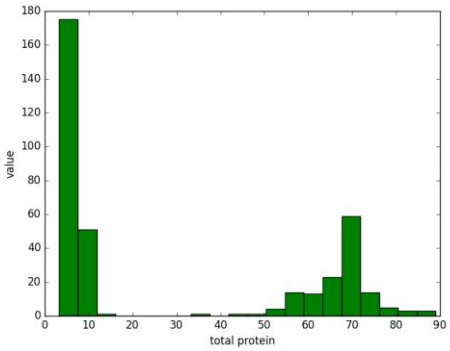










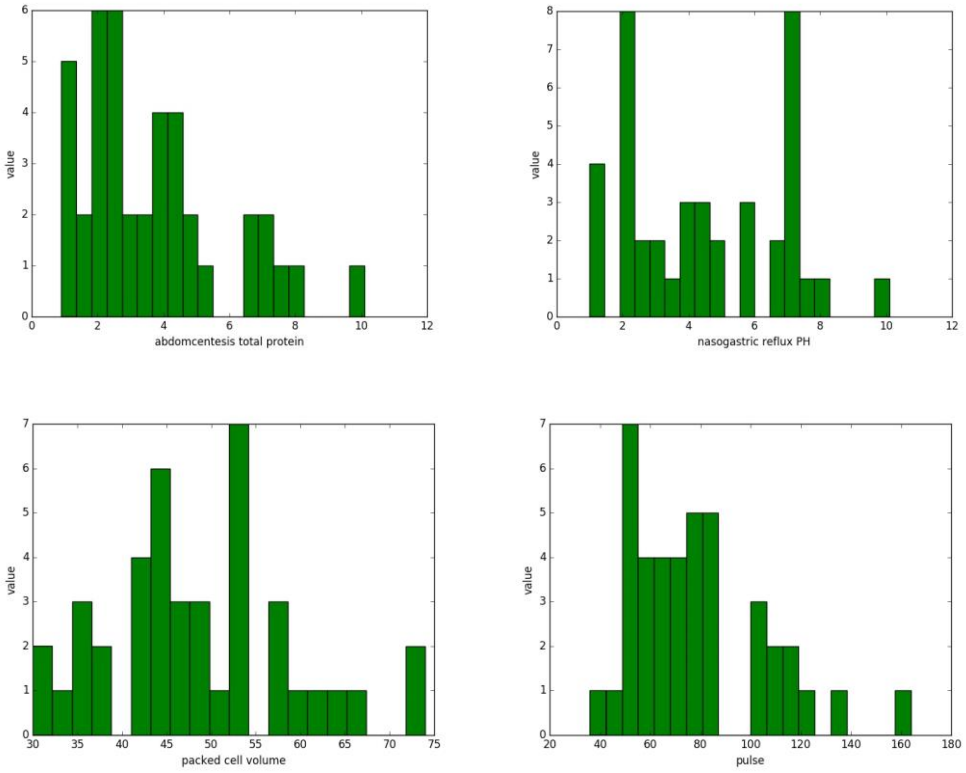


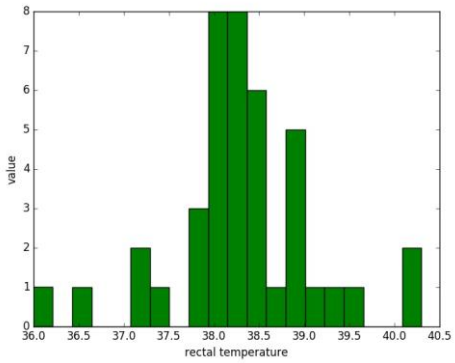
- 通过属性的相关关系来填补缺失值

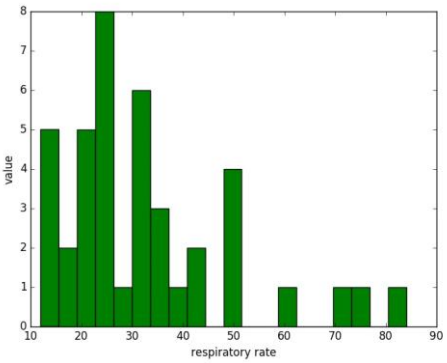
因为缺失数据较多的条目的利用价值不大，所以将缺失数据在两个及以上的行去掉，把其他的缺失值填补好。 具体见方法二中对应部分的代码。其中，相关系数求取可以通过 DataFrame 的 corr 方法取得。然后使用相关系数最大的列来直接填补缺失值。相关系数结果如下：

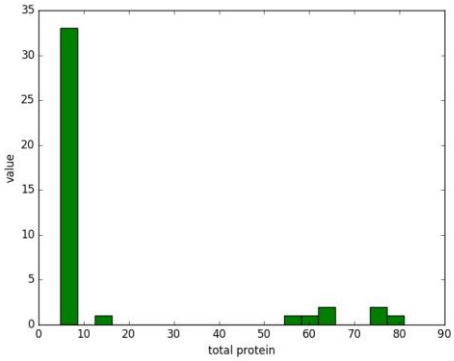
	rectal temperature	pulse	respiratory rate	nasogastric reflux PH	packed cell volume	total protein	abdomcentesis total protein
rectal temperature	1.000000	0.181491	0.373830	0.239061	-0.017837	-0.143299	0.066197
pulse	0.181491	1.000000	0.701787	0.153036	0.272232	-0.306326	0.215016
respiratory rate	0.373830	0.701787	1.000000	0.113745	0.214210	-0.277895	0.055332
nasogastric reflux PH	0.239061	0.153036	0.113745	1.000000	0.248209	-0.614953	0.298710
packed cell volume	-0.017837	0.272232	0.214210	0.248209	1.000000	-0.214838	0.028930
total protein	-0.143299	-0.306326	-0.277895	-0.614953	-0.214838	1.000000	-0.443397
abdomcentesis total protein	0.066197	0.215016	0.055332	0.298710	0.028930	-0.443397	1.000000

填补之后可视化得到的直方图如下：







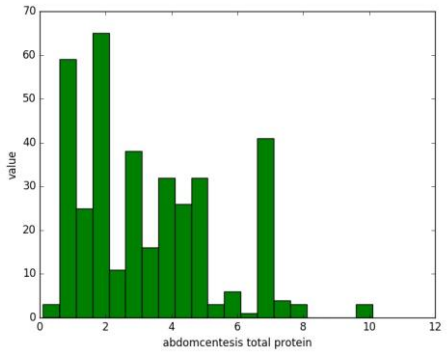


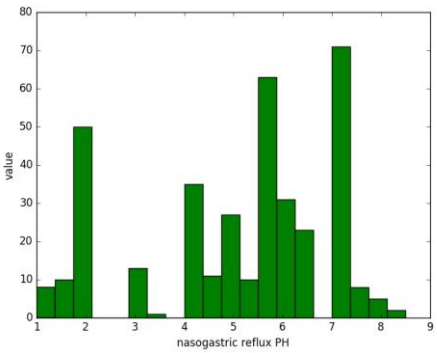
- 通过数据对象之间的相似性来填补缺失值

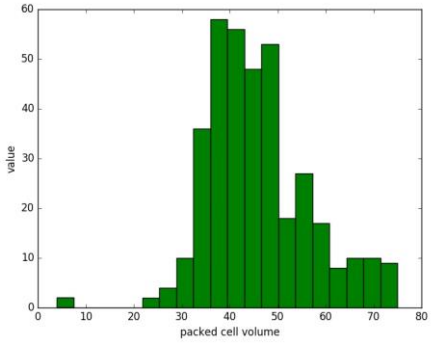
  所谓数据相似性，也就是数据行之间的相似性，考虑使用向量的相似性模型，这里采用欧几里得距离作为度量方法，公式如下：

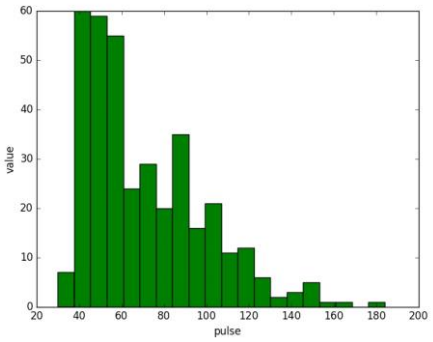
$$dist(a,b)=\sqrt{\sum_{i=1}^{n}(a_i-b_i)^2}$$

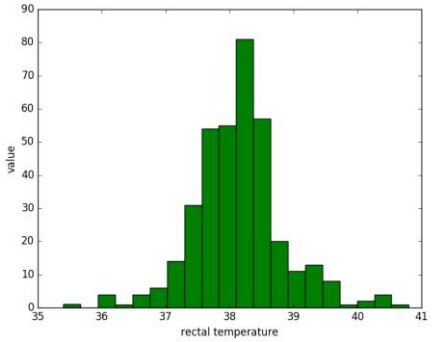
  对于包含 NaN 的向量，不考虑 NaN 。采用直接填充的方法，具体见方法二中对应部分的代码。填补之后可视化得到的直方图如下：

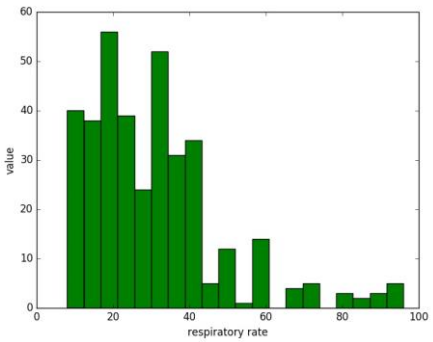


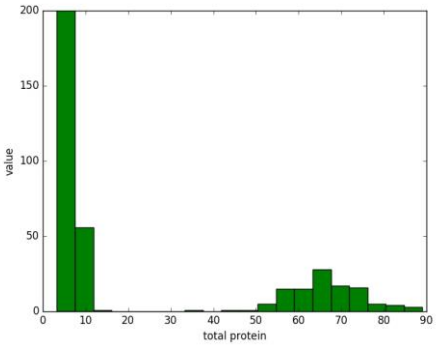



value
packed cell volume


value
pulse


value
rectal temperature


value
respiratory rate


value
total protein