作业 1 马的疝病分析

数据分析

1 读取数据并将‘？’用‘-100’代替

```
def string2num():
    f1 = open('./data/horse-colic.data.txt', 'r')
    f2 = open('./data/train.txt', 'w')
    s = f1.read()
    s = s.replace('?', '-100')
    f2.write(s)
```

```
npdata = np.loadtxt('./data/train.txt')
df = pd.DataFrame(npdata)
```

2 分析标签属性出现的频数

```
labelattr = [1, 2, 3, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 21, 23, 24, 25, 26, 27, 28]
valueattr = [4, 5, 6, 16, 19, 20, 22]
# print(df)
for attrid in labelattr:
    print(df[attrid-1].value_counts())
    print()
```

```
1.0      180
 2.0     119
-100.0     1
Name: 0, dtype: int64

1.0     276
9.0      24
Name: 1, dtype: int64

532349.0    2
529424.0    2
528729.0    2
527544.0    2
528996.0    2
528931.0    2
528469.0    2
527916.0    2
529461.0    2
528151.0    2
5279822.0   2
528904.0    2
530693.0    2
```

530526.0	2
529796.0	2
528890.0	2
533692.0	1
528742.0	1
528570.0	1
5289419.0	1
534719.0	1
533738.0	1
530101.0	1
530612.0	1
530561.0	1
5305629.0	1
528047.0	1
535208.0	1
534183.0	1
528548.0	1
	..
534403.0	1
5282839.0	1
533887.0	1
533886.0	1
530301.0	1
5275212.0	1
530297.0	1
529272.0	1
535415.0	1
530294.0	1
534899.0	1
529777.0	1
535407.0	1
5290482.0	1
533836.0	1
528743.0	1
529766.0	1
534885.0	1
5290759.0	1
535338.0	1
527709.0	1
527706.0	1
533847.0	1
528214.0	1
535381.0	1
533750.0	1

```
527698.0      1
530255.0      1
530254.0      1
530478.0      1
Name: 2, dtype: int64

 3.0        109
 1.0         78
-100.0       56
 2.0         30
 4.0         27
Name: 6, dtype: int64

 1.0        115
 3.0        103
-100.0       69
 4.0          8
 2.0          5
Name: 7, dtype: int64

 1.0        79
 3.0        58
-100.0      47
 4.0        41
 2.0        30
 5.0        25
 6.0        20
Name: 8, dtype: int64

 1.0        188
 2.0         78
-100.0       32
 3.0          2
Name: 9, dtype: int64

 3.0        67
 2.0        59
-100.0      55
 5.0        42
 4.0        39
 1.0        38
Name: 10, dtype: int64

 3.0        128
```

```
 4.0      73
-100.0    44
 1.0      39
 2.0      16
Name: 11, dtype: int64

 1.0      76
 3.0      65
 2.0      65
-100.0    56
 4.0      38
Name: 12, dtype: int64

-100.0    104
 2.0      102
 1.0       71
 3.0       23
Name: 13, dtype: int64

 1.0      120
-100.0    106
 3.0       39
 2.0       35
Name: 14, dtype: int64

-100.0    102
 4.0       79
 1.0       57
 3.0       49
 2.0       13
Name: 16, dtype: int64

-100.0    118
 5.0       79
 4.0       43
 1.0       28
 2.0       19
 3.0       13
Name: 17, dtype: int64

-100.0    165
 2.0       48
 3.0       46
 1.0       41
```

Name: 20, dtype: int64

1.0	178
2.0	77
3.0	44
-100.0	1

Name: 22, dtype: int64

1.0	191
2.0	109

Name: 23, dtype: int64

0.0	56
3111.0	33
3205.0	29
2208.0	20
2205.0	13
2209.0	11
4205.0	11
2124.0	9
1400.0	8
31110.0	7
7111.0	7
2113.0	6
2112.0	5
400.0	5
2206.0	4
4300.0	4
5400.0	4
3209.0	4
7209.0	3
3112.0	3
2111.0	3
2207.0	3
4124.0	3
5124.0	2
3124.0	2
5206.0	2
5111.0	2
2322.0	2
11124.0	2
6112.0	2
..	
9400.0	2

3025.0 2
8400.0 2
3113.0 1
4111.0 1
21110.0 1
300.0 1
3300.0 1
5205.0 1
6209.0 1
4206.0 1
4122.0 1
8300.0 1
2300.0 1
5000.0 1
2305.0 1
4207.0 1
41110.0 1
9000.0 1
3133.0 1
12208.0 1
3115.0 1
3400.0 1
1111.0 1
3207.0 1
1124.0 1
11300.0 1
7113.0 1
11400.0 1
7400.0 1
Name: 24, dtype: int64

0.0 293
3111.0 3
7111.0 1
3112.0 1
6112.0 1
1400.0 1
Name: 25, dtype: int64

0.0 299
2209.0 1
Name: 26, dtype: int64

2.0 201

1.0 99
Name: 27, dtype: int64
3 数值属性的最小值，1/4 分位数，中位数，均值，3/4 分位数，最大值

```
for attrid in valueattr:
    print(attrname[attrid - 1])
    series = df[attrid - 1].apply(pd.to_numeric, errors='coerce')
    series = series[series.notnull()]
    print('min:', series.min())
    print('1/4 quantile:', series.quantile(0.25))
    print('mean:', series.mean())
    print('median:', series.median())
    print('3/4 quantile:', series.quantile(0.75))
    print('max:', series.max())
    print()
```

rectal_temperature
min: -100.0
1/4 quantile: 37.2
mean: 10.53433333333333
median: 38.0
3/4 quantile: 38.5
max: 40.8

pulse
min: -100.0
1/4 quantile: 48.0
mean: 58.16
median: 60.0
3/4 quantile: 88.0
max: 184.0

respiratory_rate
min: -100.0
1/4 quantile: 12.0
mean: 5.203333333333333
median: 22.0
3/4 quantile: 34.25
max: 96.0

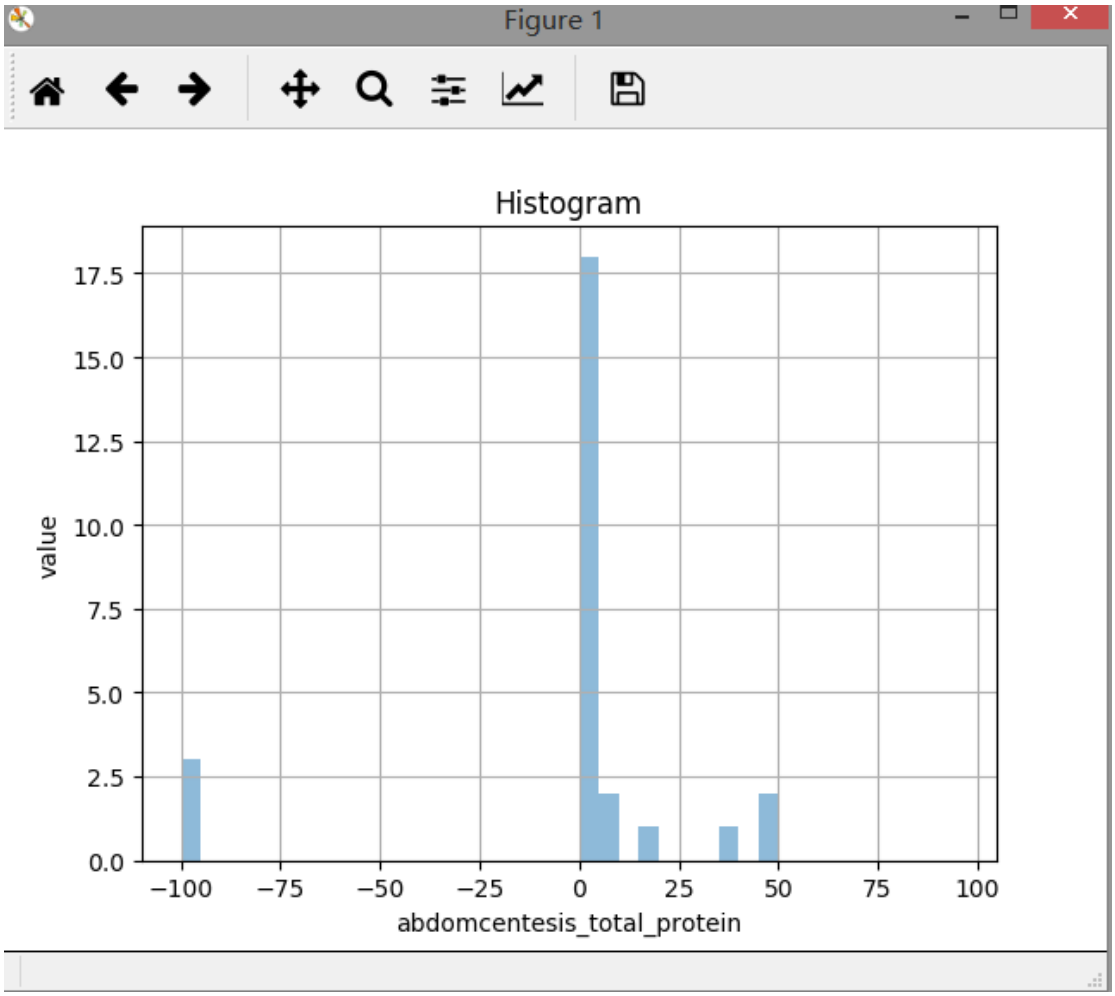
nasogastric_reflux_PH
min: -100.0
1/4 quantile: -100.0
mean: -81.50166666666667
median: -100.0
3/4 quantile: -100.0
max: 7.5

packed_cell_volume
min: -100.0
1/4 quantile: 37.0
mean: 32.153333333333336
median: 43.5
3/4 quantile: 50.0
max: 75.0

total_protein
min: -100.0
1/4 quantile: 6.2
mean: 10.766666666666664
median: 7.2
3/4 quantile: 53.25
max: 89.0

abdomcentesis_total_protein
min: -100.0
1/4 quantile: -100.0
mean: -64.97333333333334
median: -100.0
3/4 quantile: 2.0
max: 10.1
4 绘制直方图

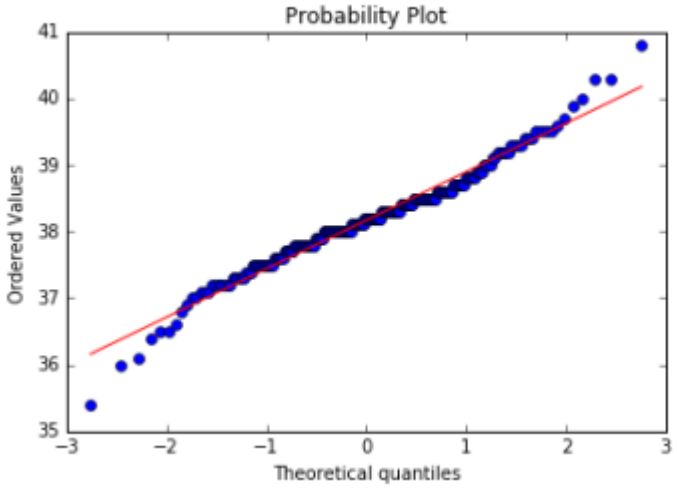
```
for attrid in valueattr:
    print(attrname[attrid - 1])
    x = npdata[attrid - 1]
    print(x)
    bins = np.arange(-100, 100, 5)
    plt.hist(x, bins = bins, alpha=0.5)
    plt.xlabel(attrname[attrid - 1])
    plt.ylabel('value')
    plt.title('Histogram')
    plt.grid(True)
    plt.show()
```



5 qq 图

```
for attrid in valueattr:
    print(attrname[attrid - 1])
    series = df[attrname[attrid - 1]]
    series = series[series != '?'].apply(pd.to_numeric, errors='coerce')
    _ = stats.probplot(series, dist="norm", plot=pylab)
```

```
rectal_temperature
```



6 盒图

```
for attrid in valueattr:
    print(attrname[attrid - 1])
    series = df[attrname[attrid - 1]]
    series = series[series != '?'].apply(pd.to_numeric, errors='coerce')
    _ = pd.DataFrame(series).boxplot()
```