Wine Reviews 数据挖掘报告

姓名：贺鹏飞　　学号：3220180700

数据可视化和摘要

导入数据

使用 MATLAB 对数据进行处理，首先将后缀.csv 的数据文件导入到 MATLAB。

```
%% 导入文本文件中的数据。
%       E:\MyWorkplace\DataMining\wine-reviews\winemag-data_first150k.csv

%% 初始化变量。
filename = 'E:\MyWorkplace\DataMining\wine-reviews\winemag-data_first150k.csv';
delimiter = ',';
startRow = 2;

%% 每个文本行的格式:
%     列 1: 双精度值 (%f)
%     列 2: 分类 (%C)
%     列 3: 文本 (%q)
%     列 4: 文本 (%q)
%     列 5: 双精度值 (%f)
%     列 6: 双精度值 (%f)
%     列 7: 分类 (%C)
%     列 8: 分类 (%C)
%     列 9: 分类 (%C)
%     列 10: 分类 (%C)
%     列 11: 文本 (%q)
formatSpec = '%f%C%q%q%f%f%C%C%C%C%q%[^\n\r]';

%% 打开文本文件。
fileID = fopen(filename,'r');

%% 根据格式读取数据列。
dataArray = textscan(fileID, formatSpec, 'Delimiter', delimiter, 'TextType', 'string', 'EmptyValue',
NaN, 'HeaderLines' ,startRow-1, 'ReturnOnError', false, 'EndOfLine', '\r\n');

%% 关闭文本文件。
fclose(fileID);
```

```
%% 将导入的数组分配给列变量名称
VarName1 = dataArray{:, 1};
country = dataArray{:, 2};
description = cellstr(dataArray{:, 3});
designation = cellstr(dataArray{:, 4});
points = dataArray{:, 5};
price = dataArray{:, 6};
province = dataArray{:, 7};
region_1 = dataArray{:, 8};
region_2 = dataArray{:, 9};
variety = dataArray{:, 10};
winery = cellstr(dataArray{:, 11});

%% 清除临时变量
clearvars filename delimiter startRow formatSpec fileID dataArray ans;
```

数据摘要

对标称属性 country、province、region1、region2、variety，给出每一个可能的频数，并将统计结果保存在 Excel 文件中，每一个属性对应一个活动页。在这里按照从大到小的顺序列出前五项，完整表格参见 Excel 文件“Tabulate_150k.xls”。

```
%% 标称属性的频数
coun=tabulate(country);
prov=tabulate(province);
reg1=tabulate(region_1);
reg2=tabulate(region_2);
vari=tabulate(variety);
xlswrite('frequency_150k.xls',coun,'country');
xlswrite('frequency_150k.xls',prov,'province');
xlswrite('frequency_150k.xls',reg1,'region1');
xlswrite('frequency_150k.xls',reg2,'region2');
xlswrite('frequency_150k.xls',vari,'variety');
```

表 1.属性 Country 的频数和频率

Country	频数	频率
US	62397	41.34%
Italy	23478	15.56%
France	21098	13.98%
Spain	8268	5.48%
Chile	5816	3.85

表 2.属性 Province 的频数和频率

Province	频数	频率
California	44508	29.49%
Washington	9750	6.46%
Tuscany	7281	4.82%
Bordeaux	6111	4.05%
Northern Spain	4892	3.24%

表 3.属性 Region1 的频数和频率

Region1	频数	频率
Napa Valley	6209	4.93%
Columbia Valley (WA)	4975	3.95%
Mendoza	3586	2.84%
Russian River Valley	3571	2.84%
California	3462	2.75%

表 4.属性 Region2 的频数和频率

Region2	频数	频率
Central Coast	13057	21.42%
Sonoma	11258	18.47%
Columbia Valley	9157	15.02%
Napa	8801	14.44%
California Other	3516	5.768%

表 5.属性 Variety 的频数和频率

Variety	频数	频率
Chardonnay	14482	9.60%
Pinot Noir	14291	9.47%
Cabernet Sauvignon	12800	8.48%
Red Blend	10062	6.67%
Bordeaux-style Red Blend	7347	4.87%

对数值属性 price 和 points，给出他们的最大、最小、均值、中位数、四分位数及缺失值的个数。由于数据集中含有缺失项，所以分别使用 MATLAB 函数：nanmax、nanmin、nanmean、nanmedian、quantile、ismissing。

```
%% 数值属性 price
max_price=nanmax(price);
min_price=nanmin(price);
mean_price=nanmean(price);
median_price=nanmedian(price);
```

```
quantile_price1=quantile(price,0.25,1);
quantile_price3=quantile(price,0.75,1);
miss_price=sum(ismissing(price));
%% 数值属性 points
max_points=nanmax(points);
min_points=nanmin(points);
mean_points=nanmean(points);
median_points=nanmedian(points);
quantile_points1=quantile(points,0.25,1);
quantile_points3=quantile(points,0.75,1);
miss_points=sum(ismissing(points));
```

表 6.属性 points

最大值	最小值	均值	中位数	第一四分位数	第三四分位数	缺失值个数
100	80	87.8884	88	86	90	0

表 7.属性 price

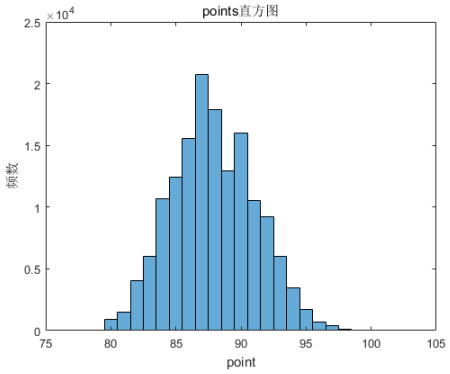
最大值	最小值	均值	中位数	第一四分位数	第三四分位数	缺失值个数
2300	4	33.1315	24	16	40	13695

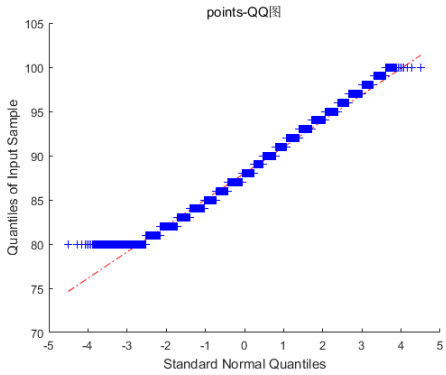
数据可视化

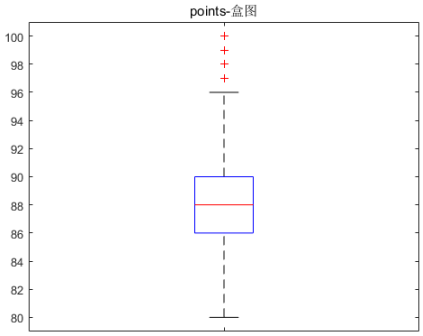
对数值属性 points，绘制直方图，QQ 图以及盒图。Points 的 QQ 图分布趋近于一条直线，因此 points 服从正态分布。

```
%points
figure;
histogram(points);
title('points-直方图');
xlabel('point');
ylabel('频数');
figure;
qqplot(points);
title('points-QQ 图');
figure;
boxplot(points);
title('points-盒图');
```

set(gca,'XTickLabel',{' '});

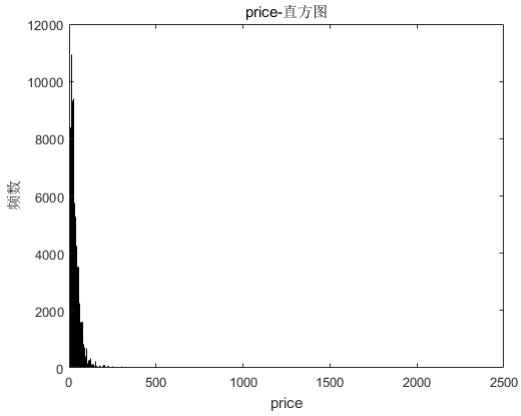


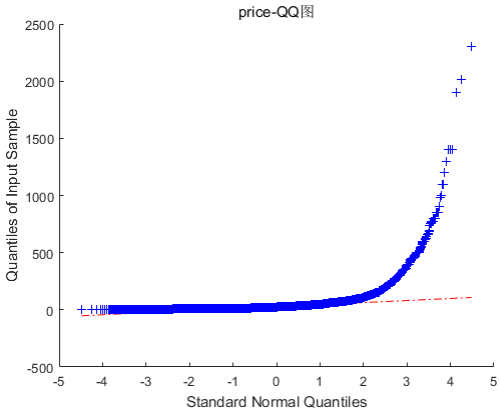


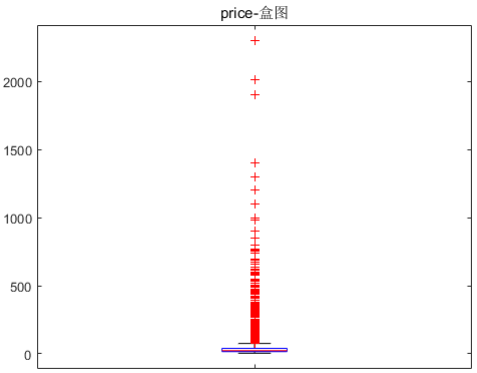


对数值属性 price，绘制直方图，QQ 图以及盒图。price 的 QQ 图分布不是一条直线，因此 price 不服从正态分布

```
%% price
figure;
histogram(price);
title('price-直方图');
xlabel('price');
ylabel('频数');
figure;
qqplot(price);
title('price-QQ 图');
figure;
boxplot(price);
title('price-盒图');
set(gca,'XTickLabel',{' '});
```







数据缺失的处理

经统计，属性 price 缺失 13695 项，分别使用下列四种策略对缺失值进行处理，并可视化地对比新旧数据集。(a)(b)(c)为原始数据集的图像，(d)(e)(f)为处理后的数据集图像。

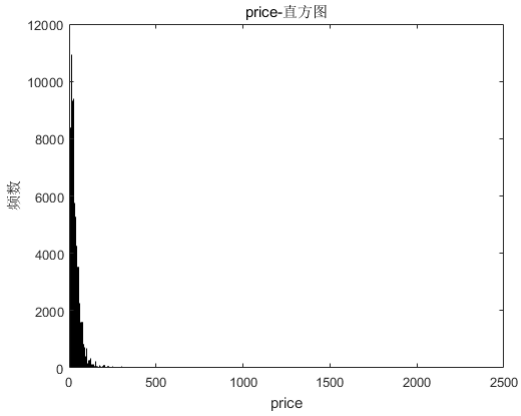
（1） 将缺失部分剔除
（2） 用最高频率值来填补缺失值
（3） 通过属性的相关关系来填补缺失值
（4） 通过数据对象之间的相似性来填补缺失值

1. 将缺失部分剔除

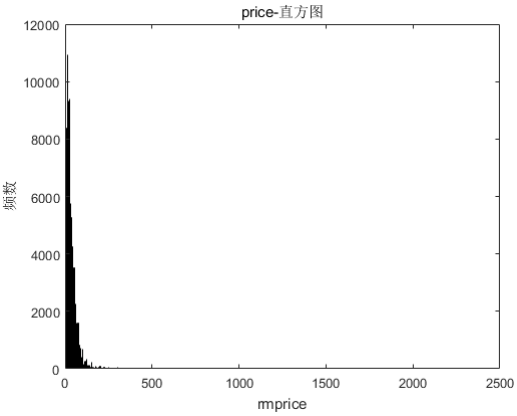
使用 matlab 自带的 rmmissing 函数剔除缺失部分，并可视化。

```
%% 剔除缺失部分,可视化
rmprice=rmmissing(price);
figure;
histogram(rmprice);
title('price-直方图');
xlabel('price');
ylabel('频数');
figure;
qqplot(rmprice);
```

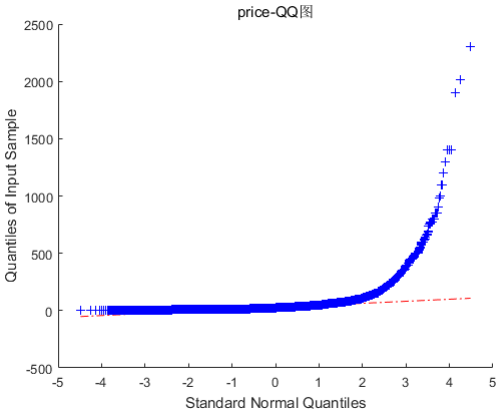
```
title('price-QQ 图');
figure;
boxplot(rmprice);
title('price-盒图');
set(gca,'XTickLabel',{' '});
```



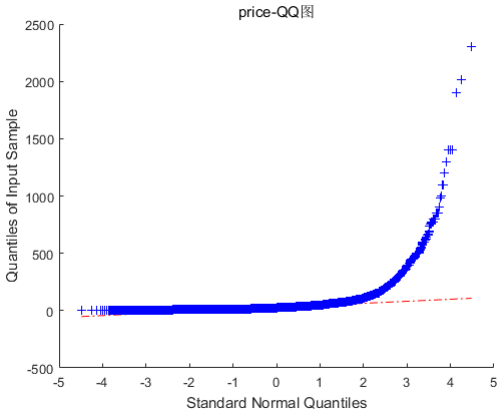
(a)



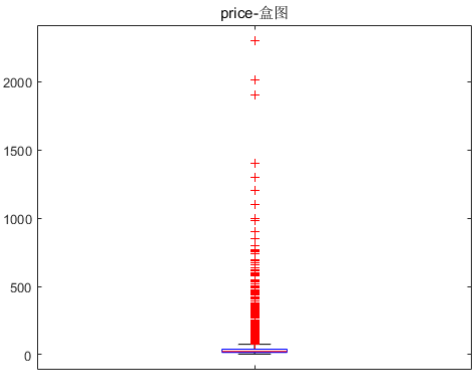
(d)



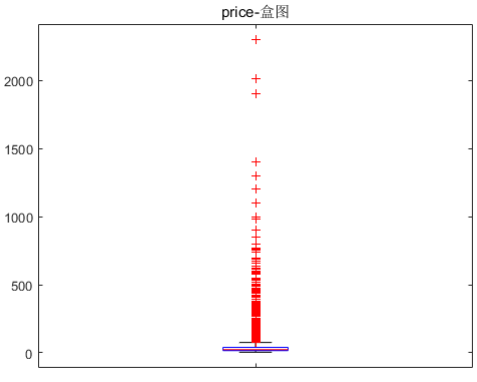
(b)



(e)



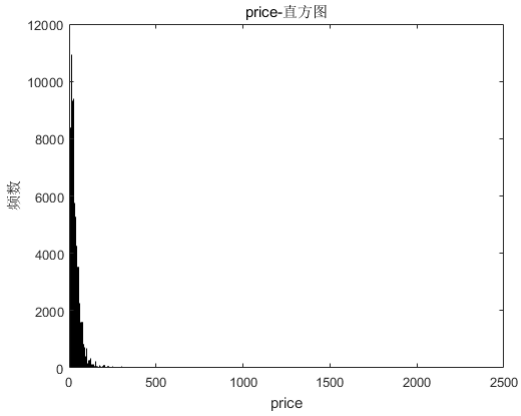
(c)



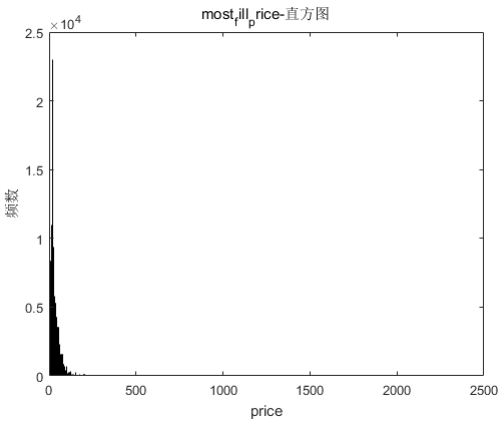
(f)

2.用最高频率值来填补缺失值

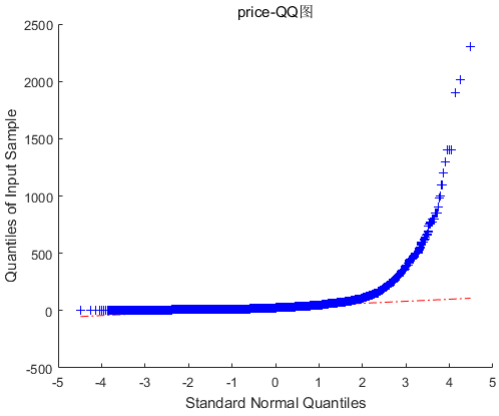
%% 用最高频率值来填补缺失值
most_fill_price=price;
most_fill_price(ismissing(price))=mode(price);



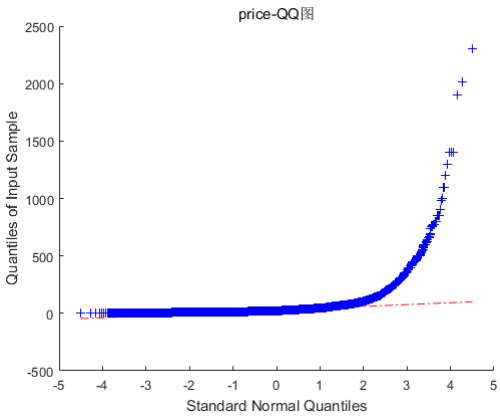
(a)



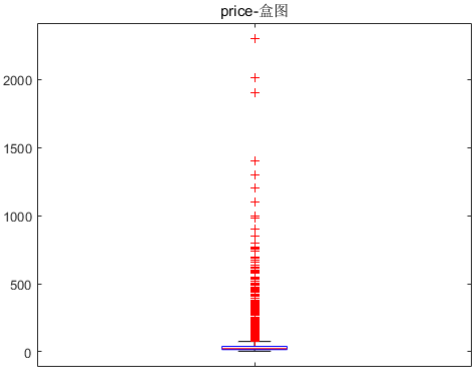
(d)



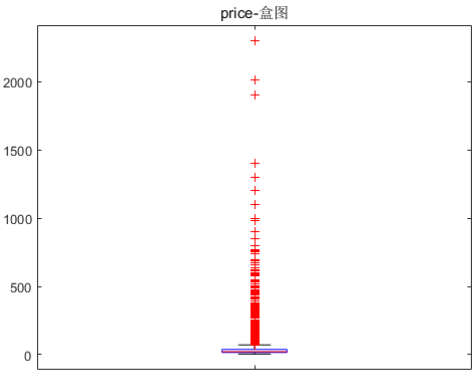
(b)



(e)



(c)

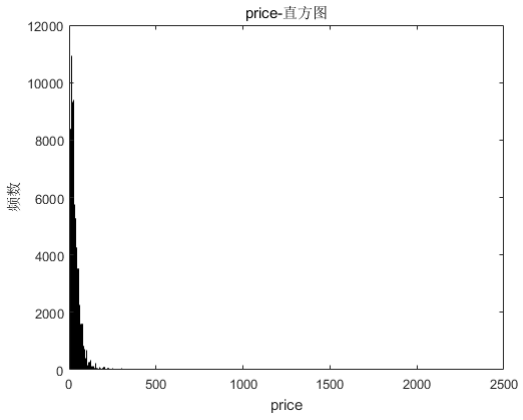


(f)

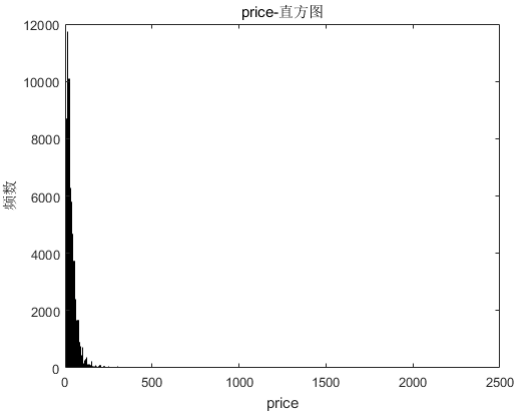
3.通过属性的相关关系来填补缺失值

通过观察，发现 points 属性和 price 的属性具有线性相关，因此使用线性插值法求出缺失的 price 值。

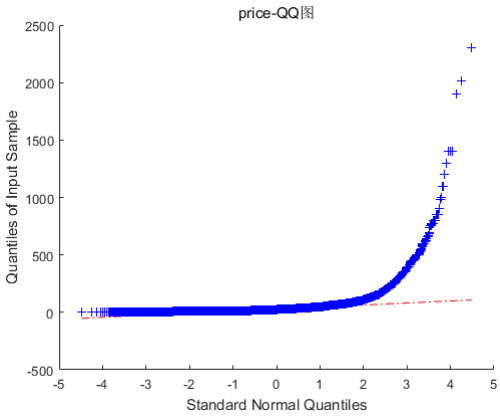
```
%% 通过属性的相关关系来填补缺失值
linear_fill_price=price;
a=polyfit(points,price,1);
linear_fill_price(ismissing(price))=polyval(a,points(ismissing(price)));
```



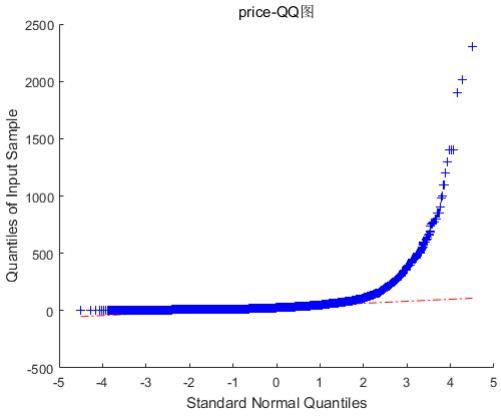
(a)



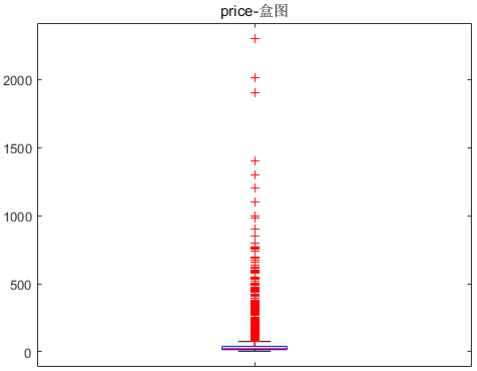
(d)



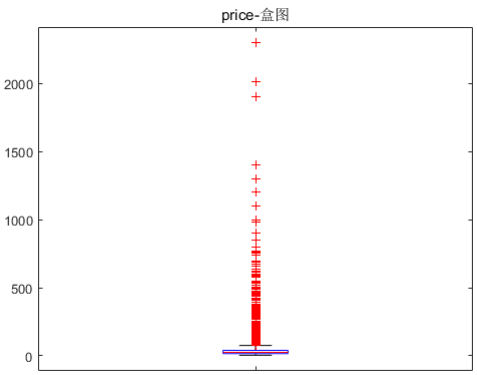
(b)



(e)



(c)

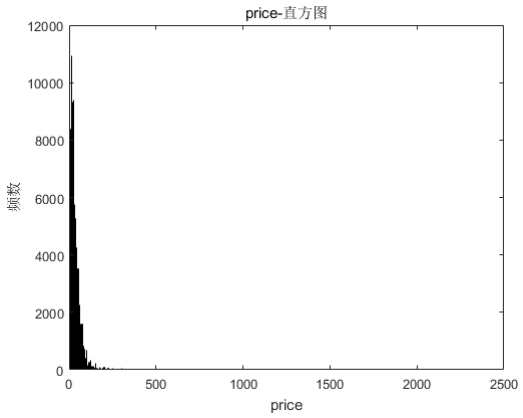


(f)

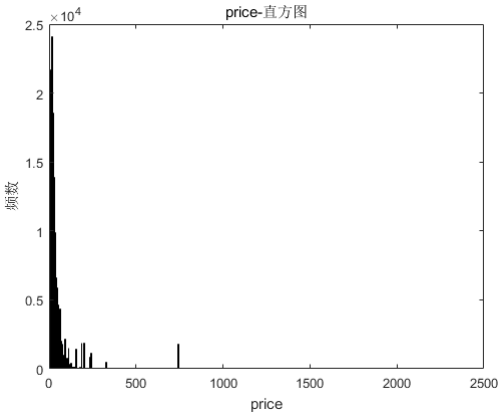
4.通过数据对象之间的相似性来填补缺失值

使用 k-临近算法对数据按照相似性分类，将未缺失的数据项作为训练集，缺失 price 的数据项作为测试集。使用 matlab 的 knnclassify 函数实现。

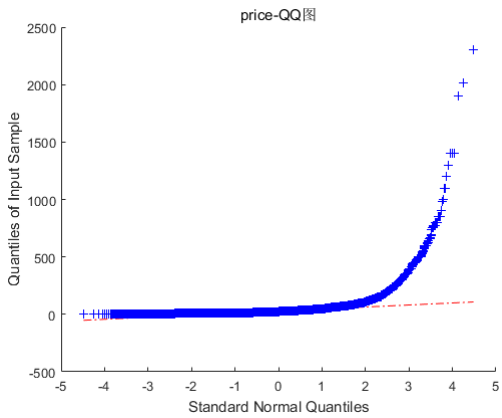
```
%% 通过数据相似性填补缺失值
knnprice=price;
test_data=points(ismissing(price));
train_data=points(rmmissing(price));
train_label=rmmissing(price);
knnprice(ismissing(price))=knnclassify(test_data,train_data,train_label);
```



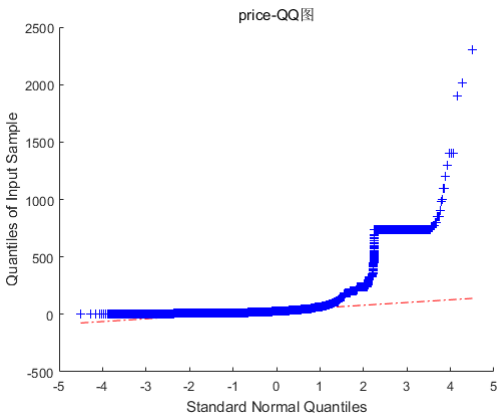
(a)



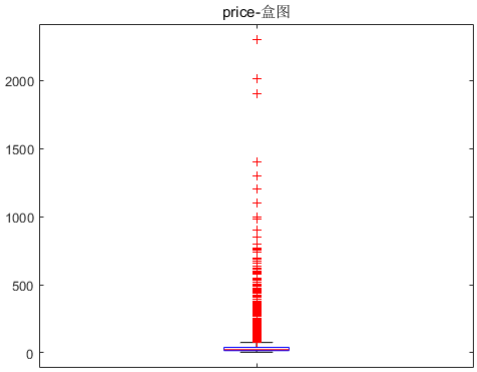
(d)



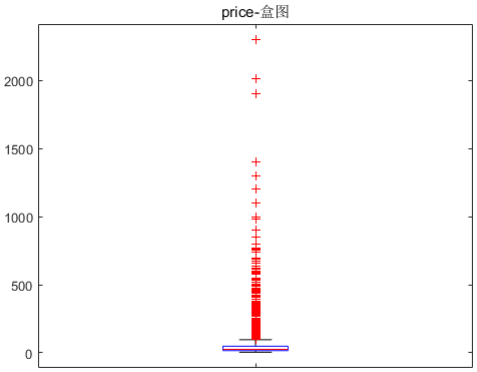
(b)



(e)



(c)



(f)