Oakland Crime Statistics 2011 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

```
%% 导入文本文件中的数据。
%
%E:\MyWorkplace\DataMining\oakland-crime-statistics-2011-to-2016\records-for-2011.csv
%

%% 初始化变量。
filename = 'E:\MyWorkplace\DataMining\oakland-crime-statistics-2011-to-2016\records-for-2011.csv';
delimiter = ',';
startRow = 2;
endRow = 180016;

%% 每个文本行的格式:
%   列 1: 分类 (%C)
%   列 2: 文本 (%s)
%   列 3: 分类 (%C)
%   列 4: 双精度值 (%f)
%   列 5: 分类 (%C)
%   列 6: 双精度值 (%f)
%   列 7: 分类 (%C)
%   列 8: 分类 (%C)
%   列 9: 分类 (%C)
%   列 10: 文本 (%s)
formatSpec = '%C%s%C%f%C%f%C%C%C%s%[^\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, endRow-startRow+1, '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
Agency = dataArray{:, 1};
CreateTime = cellstr(dataArray{:, 2});
Location = dataArray{:, 3};
AreaId = dataArray{:, 4};
Beat = dataArray{:, 5};
Priority = dataArray{:, 6};
IncidentTypeId = dataArray{:, 7};
IncidentTypeDescription = dataArray{:, 8};
EventNumber = dataArray{:, 9};
ClosedTime = cellstr(dataArray{:, 10});

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

数据摘要

对标称属性 Agency、Location、Beat、IncidentTypeId、IncidentTypeDescription，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“CrimeTabulate_2011.xls”。

```
%% 标称属性的频数
Agency_t=tabulate(Agency);
Location_t=tabulate(Location);
Beat_t=tabulate(Beat);
InctyTpeId_t=tabulate(IncidentTypeId);
InctyTypeDes_t=tabulate(IncidentTypeDescription);
xlswrite('CrimeTabulate_2011.xls',Agency_t,'Agency');
xlswrite('CrimeTabulate_2011.xls',Location_t,'Location');
xlswrite('CrimeTabulate_2011.xls',Beat_t,'Beat');
xlswrite('CrimeTabulate_2011.xls',InctyTpeId_t,'IncidentTypeId');
xlswrite('CrimeTabulate_2011.xls',InctyTypeDes_t,'IncidentTypeDescription');
```

表 1.属性 Agency 的频数和频率

Agency	频数	频率
OP	180015	100%

表 2.属性 Location 的频数和频率

Location	频数	频率
INTERNATIONAL BLVD	3866	2.14%
MACARTHUR BLVD	3129	1.74%
AV&INTERNATIONAL BLVD	3067	1.70%
BROADWAY	2132	1.18%
FOOTHILL BLVD	1791	1.00%

表 3.属性 Beat 的频数和频率

Beat	频数	频率
04X	7410	4.13%
08X	6885	3.84%
26Y	5478	3.05%
30Y	5295	2.95%
06X	5119	2.85%

表 4.属性 IncidentTypeId 的频数和频率

IncidentTypeId	频数	频率
933R	17348	9.64%
911H	12817	7.12%
SECCK	11393	6.33%
415	10752	5.97%
10851	7180	3.99%

表 5. 属性 IncidentTypeDescription 的频数和频率

IncidentTypeDescription	频数	频率
ALARM-RINGER	17348	9.64%
911 HANG-UP	12817	7.12%
SECURITY CHECK	11393	6.33%
STOLEN VEHICLE	7180	3.99%
415 UNKNOWN	6624	3.68%

对数值属性 Areald 和 priority，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 Areald
max_Areald=nanmax(Areald);
min_Areald=min(Areald);
mean_Areald=mean(Areald);
median_Areald=median(Areald);
quantile_Areald_1=quantile(Areald,0.25,1);
quantile_Areald3=quantile(Areald,0.75,1);
```

```
miss_AreaId=sum(ismissing(AreaId));
%% 数值属性 Priority
max_Priority=nanmax(Priority);
min_Priority=nanmin(Priority);
mean_Priority=mean(Priority);
median_Priority=median(Priority);
quantile_Priority_1=quantile(Priority,0.25,1);
quantile_Priority_=quantile(Priority,0.75,1);
miss_Priority=sum(ismissing(Priority));
```

表 1.属性 Areald

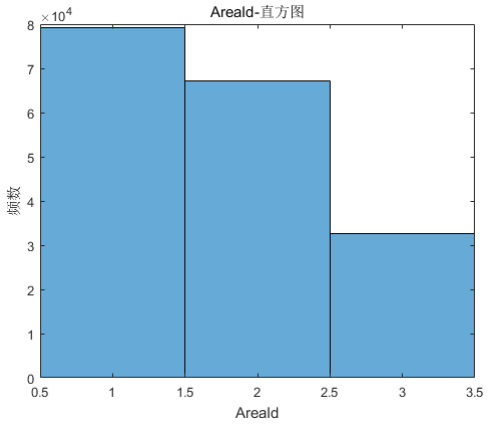
最大值	最小值	均值	第一四分位	中位数	第三四分位	缺失值个数
3	1	1.74	1	2	2	903

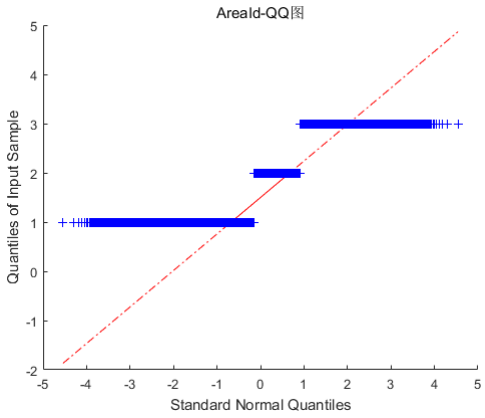
表 2.属性 priority

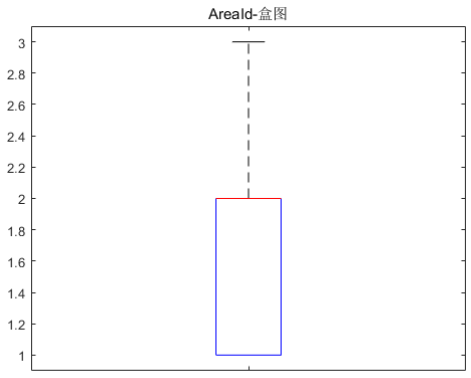
最大值	最小值	均值	第一四分位	中位数	第三四分位	缺失值个数
2	0	1.79	2	2	2	0

数据可视化

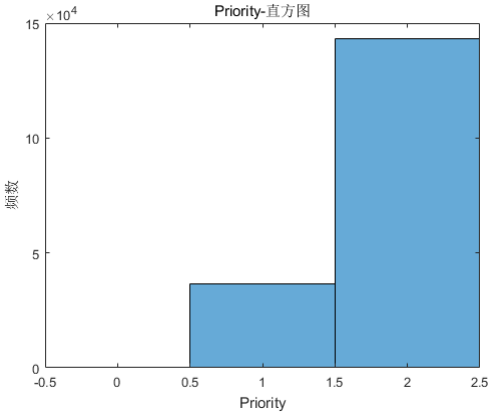
对数值属性 Areald，绘制直方图、QQ 图以及盒图，很显然，Areald 不服从正态分布。

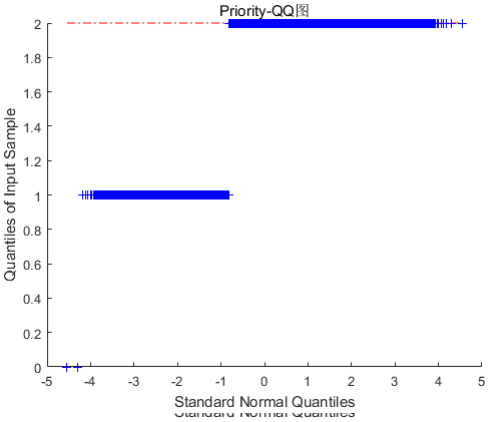


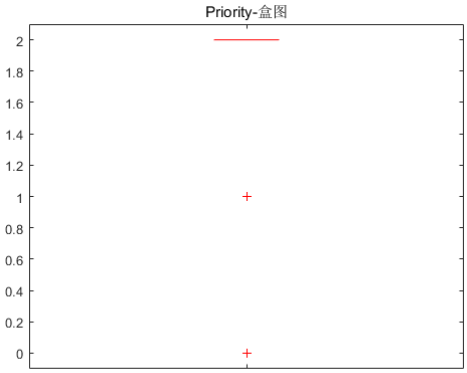




数值属性 priority 的直方图、QQ 图和盒图，由 QQ 图看出，priority 不服从正态分布。

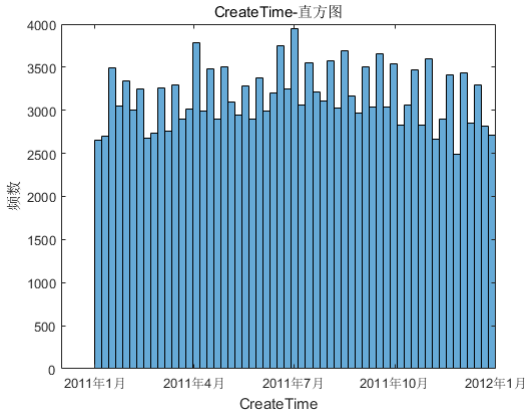


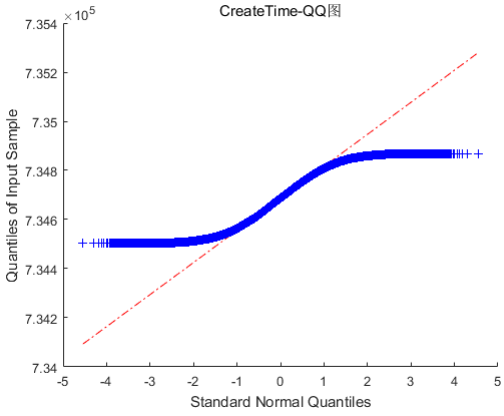


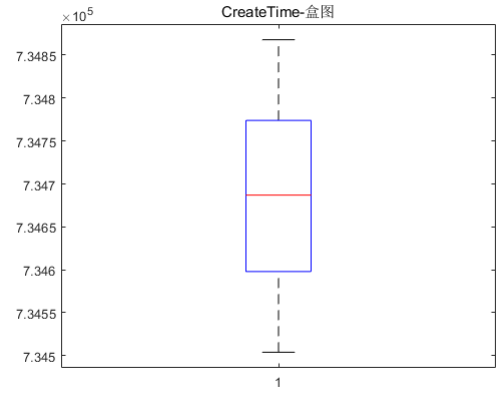


在绘制时间属性 CreateTime 的直方图、QQ 图和盒图时，首先使用 datenum 函数将日期数据转换为数值型数据，然后进行绘制。CreateTime 近似正态分布。

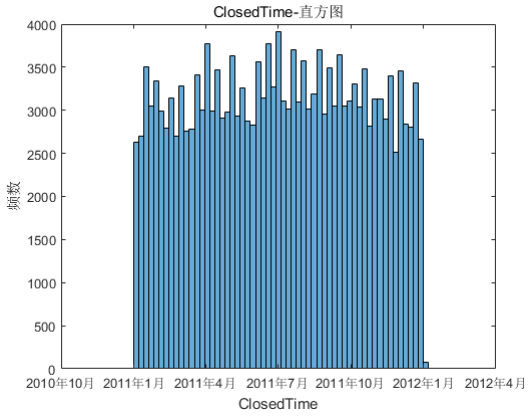
```
createnum=datenum(CreateDate)；
figure；
histogram（createnum）；
title（‘CreateTime-直方图’）；
```

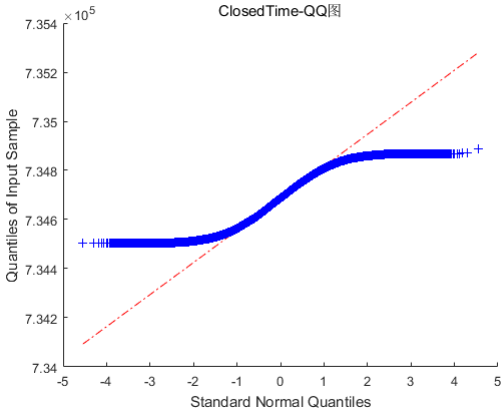


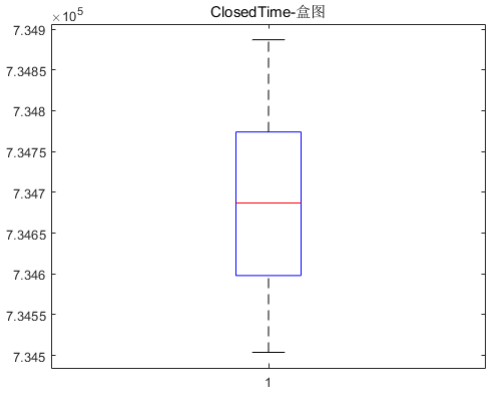




同上所述，绘制时间属性 ClosedTime 的直方图、QQ 图和盒图，ClosedTime 同样近似服从正态分布。







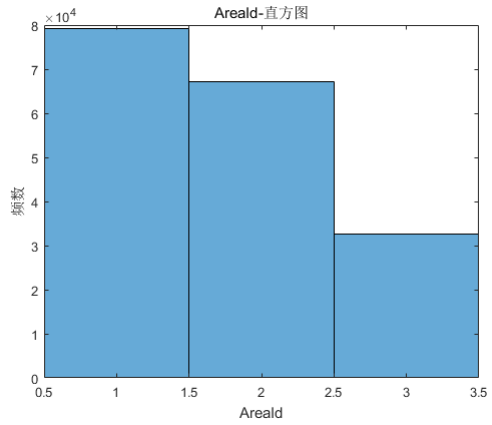
数据缺失的处理

经统计，属性 Areald 缺失 903 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。

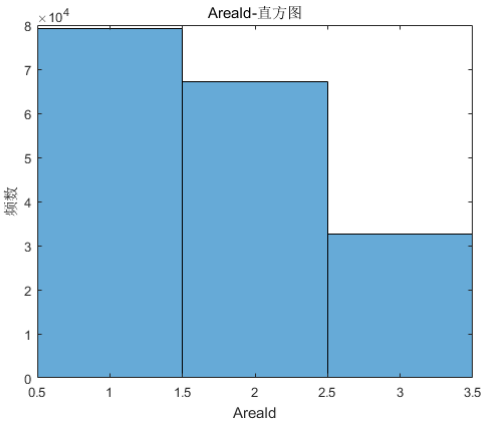
（1） 将缺失部分剔除
（2） 用最高频率值来填补缺失值
（3） 通过属性的相关关系来填补缺失值
（4） 通过数据对象之间的相似性来填补缺失值

1. 将缺失部分剔除

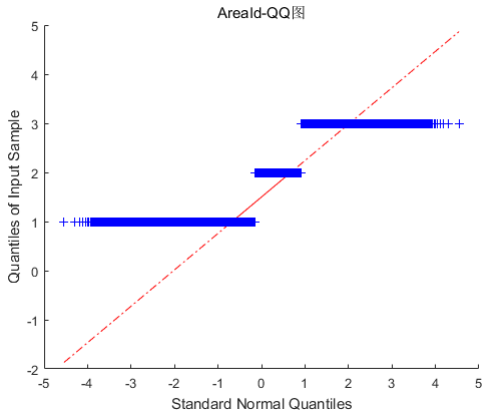
```
%% 剔除缺失部分,可视化
rmAreald=rmmissing(Areald);
figure;
histogram(rmAreald);
title('price-直方图');
xlabel('Areald');
ylabel('频数');
figure;
qqplot(rmAreald);
title('Areald-QQ 图');
figure;
boxplot(rmAreald);
title('Areald-盒图');
set(gca,'XTickLabel',{' '});
```



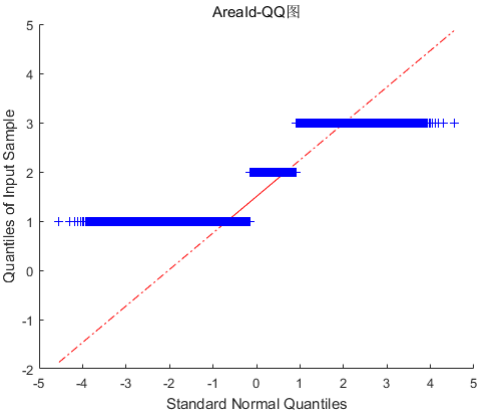
(a)



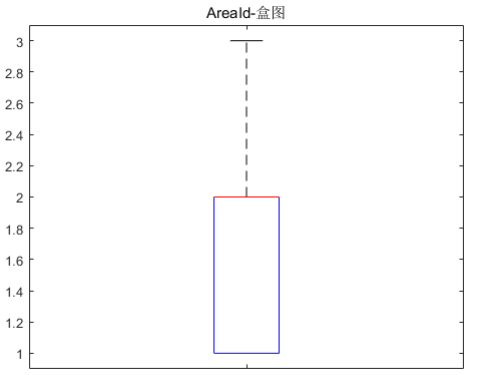
(d)



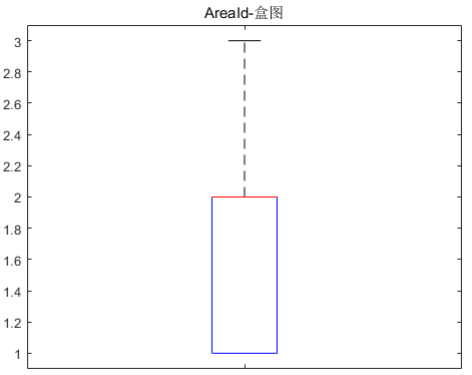
(b)



(e)



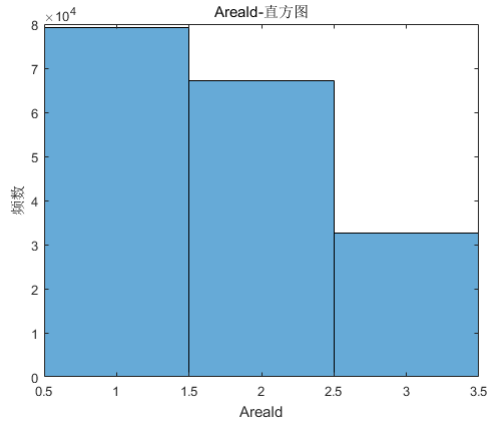
(c)



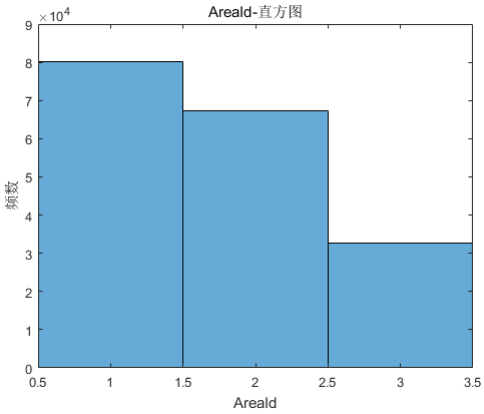
(f)

2.用最高频率值来填补缺失值

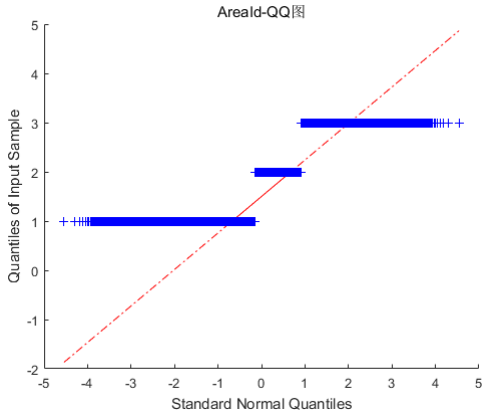
```
%% 用最高频率值来填补缺失值
most_fill_Areald=Areald;
most_fill_Areald(ismissing(Areald))=mode(Areald);
```



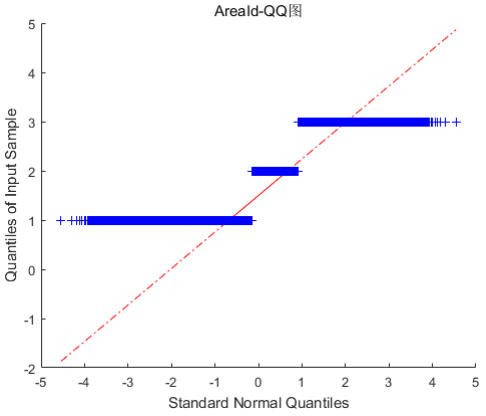
(a)



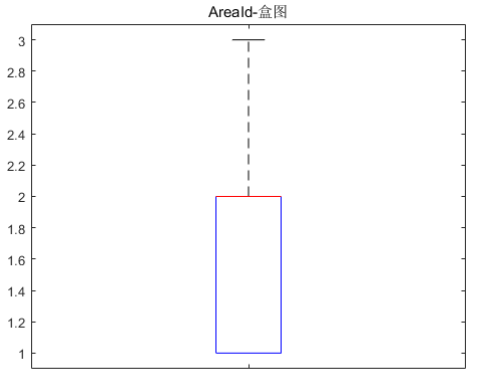
(d)



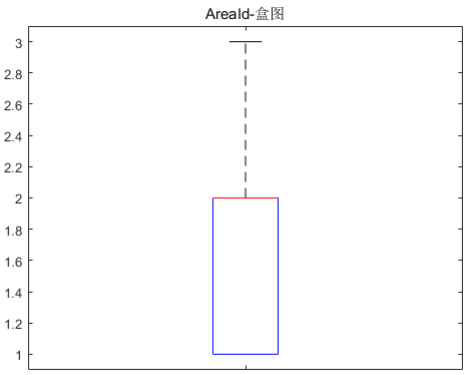
(b)



(e)



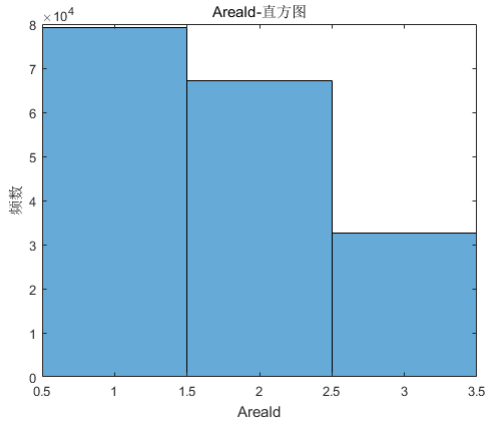
(c)



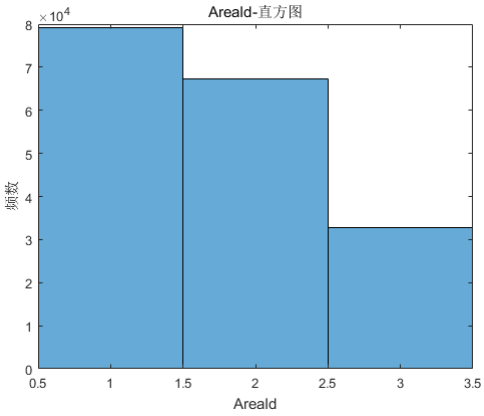
(f)

3.通过属性的相关关系来填补缺失值

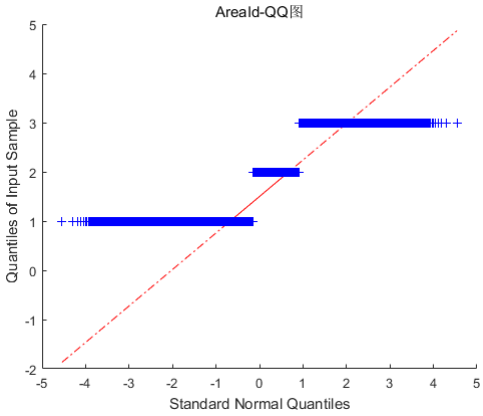
```
%% 通过属性的相关关系来填补缺失值
linear_fill_Areald=Areald;
a=polyfit(Priority,Areald,1);
linear_fill_Areald(ismissing(Areald))=polyval(a,Priority(ismissing(Areald)));
```



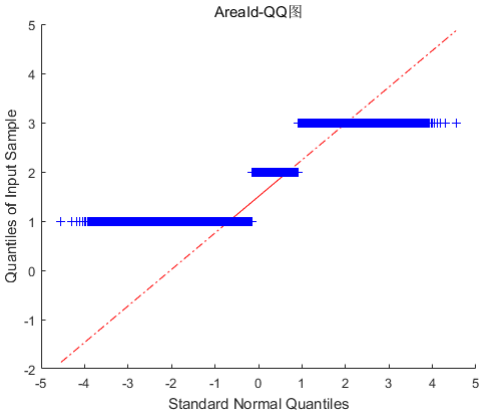
(a)



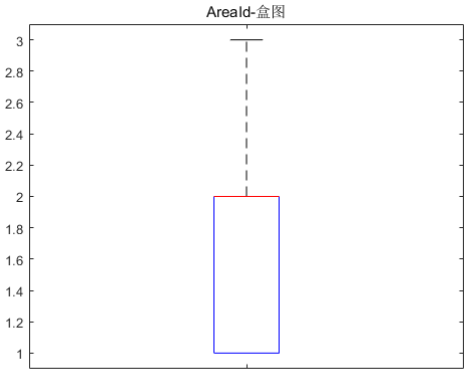
(d)



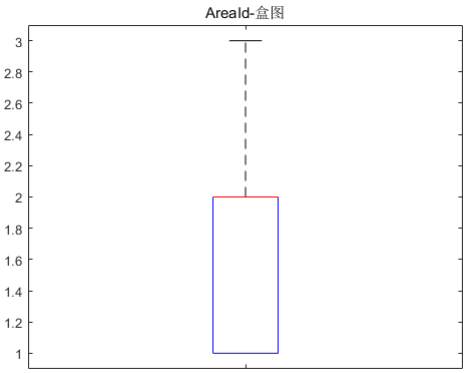
(b)



(e)



(c)

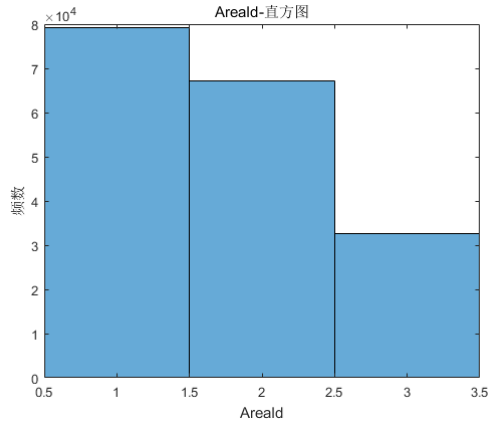


(f)

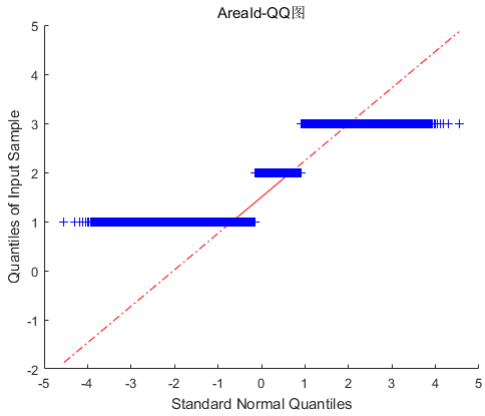
4. 通过数据对象之间的相似性来填补缺失值

```
knnAreald=Areald;
test_data=Priority(ismissing(Areald));
train_data=Priority(rmmissing(Areald));
train_label=rmmissing(Areald);
```

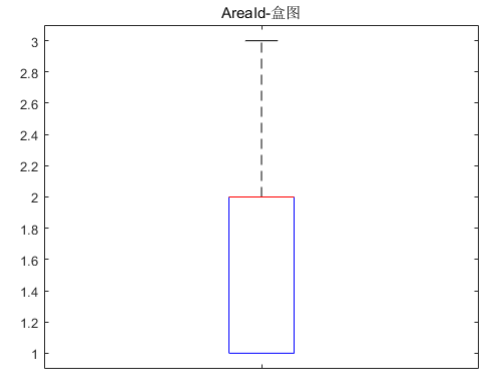
knnAreaId(ismissing(AreaId))=knnclassify(test_data,train_data,train_label);



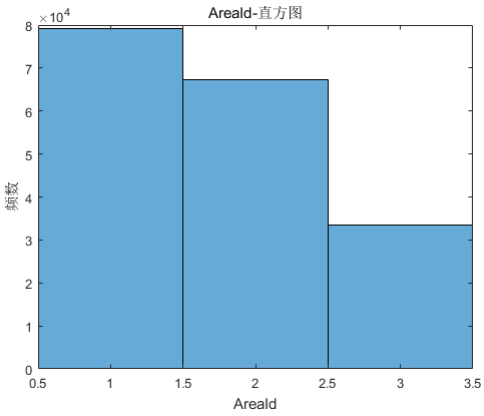
(a)



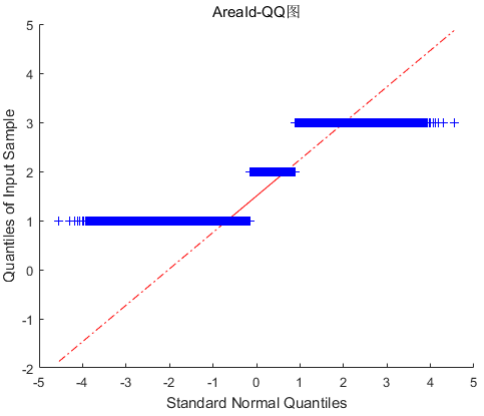
(b)



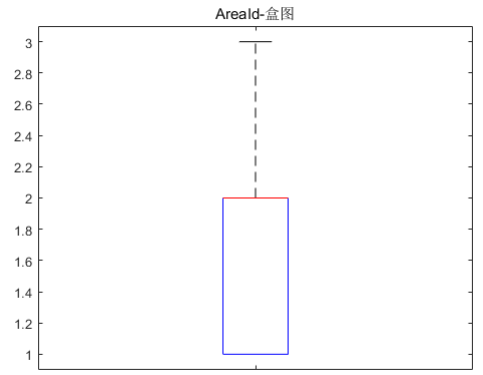
(c)



(d)



(e)



(f)