**项目报告·数据挖掘作业一：数据探索性分析与数据预处理**

梁天行 5720182040

本次实验借助SPSS软件完成。

**数据源**：oakland-crime-statistics-2011-to-2016 records-for-2016

**数据可视化和摘要**

1. 对标称属性，给出每个可能取值的频数

Location：数据量过大见输出1.pdf

Area Id：

| **AreaId** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  |  | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 |  | 1 | .0 | .0 | .0 |
| 39 | 1 | .0 | .0 | .0 |
| BE | 2 | .0 | .0 | .0 |
| CA | 1 | .0 | .0 | .0 |
| SA | 2 | .0 | .0 | .0 |
| VA | 1 | .0 | .0 | .0 |
| WA | 1 | .0 | .0 | .0 |
| JLS | 1 | .0 | .0 | .0 |
| P1 | 41416 | 37.4 | 37.4 | 37.4 |
| P2 | 19610 | 17.7 | 17.7 | 55.1 |
| P3 | 47421 | 42.8 | 42.8 | 97.9 |
| PCW | 193 | .2 | .2 | 98.0 |
| POU | 2173 | 2.0 | 2.0 | 100.0 |
| TEC | 4 | .0 | .0 | 100.0 |
| WAG | 1 | .0 | .0 | 100.0 |
| 合计 | 110828 | 100.0 | 100.0 |  |

Beat：

| **Beat** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  |  | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 |  | 573 | .5 | .5 | .5 |
| CA | 2 | .0 | .0 | .5 |
| 01X | 1304 | 1.2 | 1.2 | 1.7 |
| 02X | 1746 | 1.6 | 1.6 | 3.3 |
| 02Y | 1659 | 1.5 | 1.5 | 4.8 |
| 03X | 3195 | 2.9 | 2.9 | 7.7 |
| 03Y | 2562 | 2.3 | 2.3 | 10.0 |
| 04X | 4515 | 4.1 | 4.1 | 14.0 |
| 05X | 1342 | 1.2 | 1.2 | 15.2 |
| 05Y | 408 | .4 | .4 | 15.6 |
| 06X | 2580 | 2.3 | 2.3 | 17.9 |
| 07X | 2831 | 2.6 | 2.6 | 20.5 |
| 08X | 3931 | 3.5 | 3.5 | 24.0 |
| 09X | 2158 | 1.9 | 1.9 | 26.0 |
| 10X | 1557 | 1.4 | 1.4 | 27.4 |
| 10Y | 1573 | 1.4 | 1.4 | 28.8 |
| 11X | 1208 | 1.1 | 1.1 | 29.9 |
| 12X | 1299 | 1.2 | 1.2 | 31.1 |
| 12Y | 1987 | 1.8 | 1.8 | 32.9 |
| 13X | 630 | .6 | .6 | 33.4 |
| 13Y | 952 | .9 | .9 | 34.3 |
| 13Z | 1397 | 1.3 | 1.3 | 35.6 |
| 14X | 1832 | 1.7 | 1.7 | 37.2 |
| 14Y | 1027 | .9 | .9 | 38.1 |
| 15X | 1393 | 1.3 | 1.3 | 39.4 |
| 16X | 708 | .6 | .6 | 40.0 |
| 16Y | 907 | .8 | .8 | 40.9 |
| 17X | 969 | .9 | .9 | 41.7 |
| 17Y | 1482 | 1.3 | 1.3 | 43.1 |
| 18X | 721 | .7 | .7 | 43.7 |
| 18Y | 1102 | 1.0 | 1.0 | 44.7 |
| 19X | 3455 | 3.1 | 3.1 | 47.8 |
| 20X | 2702 | 2.4 | 2.4 | 50.3 |
| 21X | 1479 | 1.3 | 1.3 | 51.6 |
| 21Y | 2100 | 1.9 | 1.9 | 53.5 |
| 22X | 1541 | 1.4 | 1.4 | 54.9 |
| 22Y | 1420 | 1.3 | 1.3 | 56.2 |
| 23X | 3076 | 2.8 | 2.8 | 58.9 |
| 24X | 1704 | 1.5 | 1.5 | 60.5 |
| 24Y | 1454 | 1.3 | 1.3 | 61.8 |
| 25X | 2467 | 2.2 | 2.2 | 64.0 |
| 25Y | 739 | .7 | .7 | 64.7 |
| 26X | 1766 | 1.6 | 1.6 | 66.3 |
| 26Y | 3511 | 3.2 | 3.2 | 69.4 |
| 27X | 2333 | 2.1 | 2.1 | 71.5 |
| 27Y | 2517 | 2.3 | 2.3 | 73.8 |
| 28X | 1261 | 1.1 | 1.1 | 75.0 |
| 29X | 2646 | 2.4 | 2.4 | 77.3 |
| 30X | 3416 | 3.1 | 3.1 | 80.4 |
| 30Y | 3473 | 3.1 | 3.1 | 83.6 |
| 31X | 1439 | 1.3 | 1.3 | 84.9 |
| 31Y | 2460 | 2.2 | 2.2 | 87.1 |
| 31Z | 1268 | 1.1 | 1.1 | 88.2 |
| 32X | 2316 | 2.1 | 2.1 | 90.3 |
| 32Y | 2093 | 1.9 | 1.9 | 92.2 |
| 33X | 2276 | 2.1 | 2.1 | 94.3 |
| 34X | 2857 | 2.6 | 2.6 | 96.8 |
| 35X | 2328 | 2.1 | 2.1 | 98.9 |
| 35Y | 1159 | 1.0 | 1.0 | 100.0 |
| P1 | 2 | .0 | .0 | 100.0 |
| P3 | 3 | .0 | .0 | 100.0 |
| PCW | 1 | .0 | .0 | 100.0 |
| PDT | 16 | .0 | .0 | 100.0 |
| 合计 | 110828 | 100.0 | 100.0 |  |

Priority：

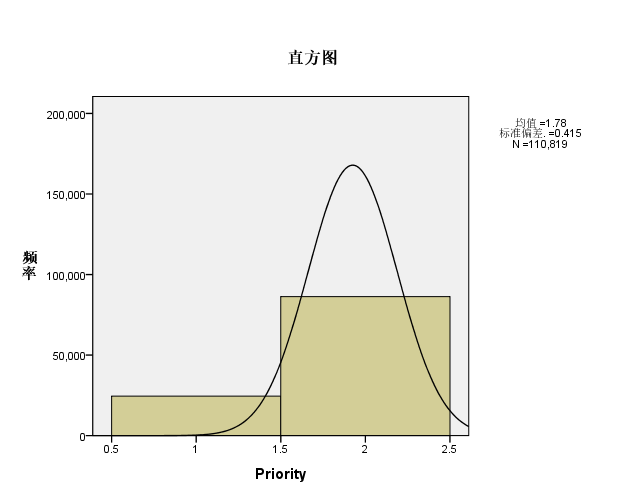
| **Priority** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  |  | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 | 1 | 24553 | 22.2 | 22.2 | 22.2 |
| 2 | 86266 | 77.8 | 77.8 | 100.0 |
| 合计 | 110819 | 100.0 | 100.0 |  |
| 缺失 | 系统 | 9 | .0 |  |  |
| 合计 | | 110828 | 100.0 |  |  |

Incident Type Id：数据量过大见输出2.pdf

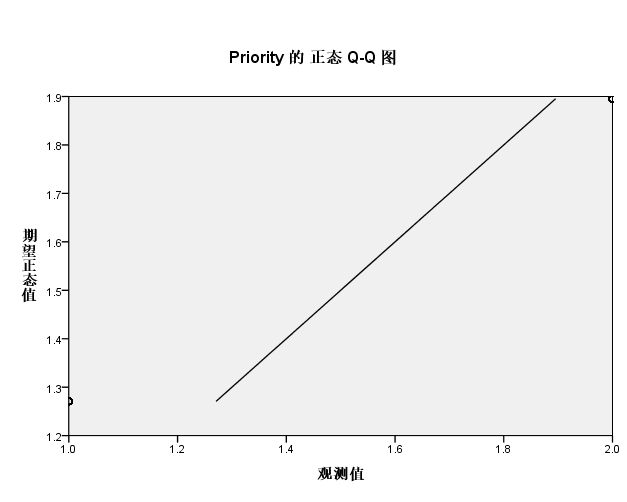
1. 数值属性，给出最大、最小、均值、中位数、四分位数及缺失值的个数

| **统计量** | | |
| --- | --- | --- |
| Priority | | |
| N | 有效 | 110819 |
| 缺失 | 9 |
| 均值 | | 1.78 |
| 中值 | | 2.00 |
| 极小值 | | 1 |
| 极大值 | | 2 |
| 百分位数 | 25 | 2.00 |
| 50 | 2.00 |
| 75 | 2.00 |

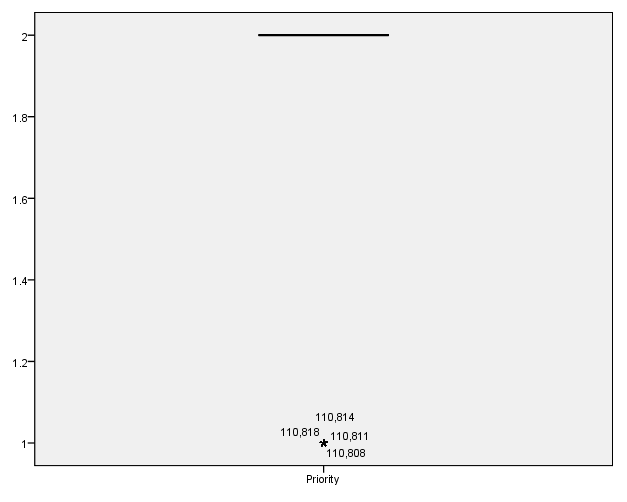
1. 针对数值属性，绘制直方图，用qq图检验其分布是否为正态分布。



| **估计的分布参数** | | |
| --- | --- | --- |
|  |  | Priority |
| 正态分布 | 位置 | 1.78 |
| 标度 | .415 |
| 个案未进行加权。 | | |



1. 绘制盒图，对离群值进行识别



**数据缺失的处理**

1. 观察数据集中缺失数据，分析其缺失的原因。

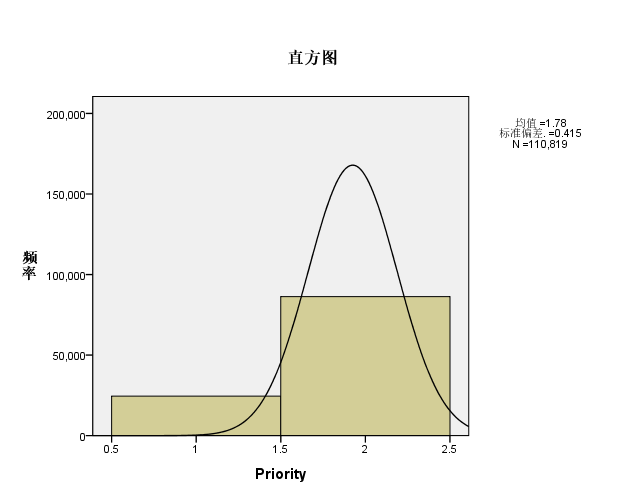
| **统计量** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Location | AreaId | Beat | Priority | IncidentTypeDescription |
| N | 有效 | 110828 | 110828 | 110828 | 110819 | 110828 |
| 缺失 | 0 | 0 | 0 | 9 | 0 |



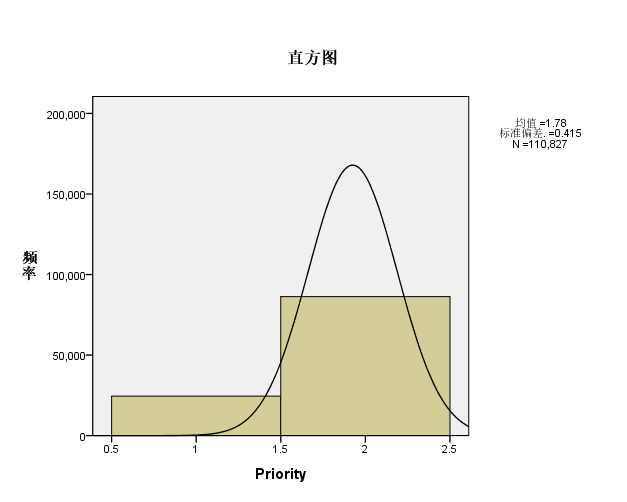
前9行Priority数据缺失，有多余的一行，其他原因未知

1. 分别使用下列四种策略对缺失值进行处理:

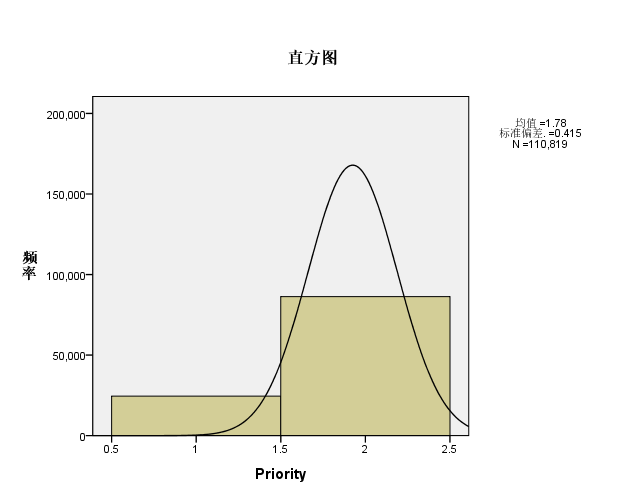
将缺失部分剔除



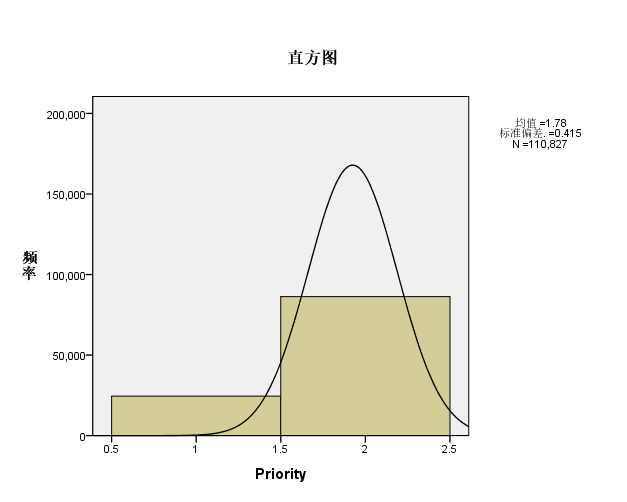
用最高频率值来填补缺失值



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



**数据源**：Wine Reviews winemag-data\_first150k

**数据可视化和摘要**

1. 对标称属性，给出每个可能取值的频数

Country:

| **country** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  |  | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 |  | 5 | .0 | .0 | .0 |
| and the not | 1 | .0 | .0 | .0 |
| fine and ex | 1 | .0 | .0 | .0 |
| lending its | 1 | .0 | .0 | .0 |
| marked by t | 1 | .0 | .0 | .0 |
| which is ri | 1 | .0 | .0 | .0 |
| Albania | 2 | .0 | .0 | .0 |
| Argentina | 5631 | 3.7 | 3.7 | 3.7 |
| Australia | 4957 | 3.3 | 3.3 | 7.0 |
| Austria | 3057 | 2.0 | 2.0 | 9.0 |
| Bosnia and H | 4 | .0 | .0 | 9.1 |
| Brazil | 25 | .0 | .0 | 9.1 |
| Bulgaria | 77 | .1 | .1 | 9.1 |
| Canada | 196 | .1 | .1 | 9.2 |
| Chile | 5816 | 3.9 | 3.9 | 13.1 |
| China | 3 | .0 | .0 | 13.1 |
| Croatia | 89 | .1 | .1 | 13.2 |
| Cyprus | 31 | .0 | .0 | 13.2 |
| Czech Republ | 6 | .0 | .0 | 13.2 |
| Egypt | 3 | .0 | .0 | 13.2 |
| England | 9 | .0 | .0 | 13.2 |
| France | 21098 | 14.0 | 14.0 | 27.2 |
| Georgia | 43 | .0 | .0 | 27.2 |
| Germany | 2452 | 1.6 | 1.6 | 28.8 |
| Greece | 884 | .6 | .6 | 29.4 |
| Hungary | 231 | .2 | .2 | 29.6 |
| India | 8 | .0 | .0 | 29.6 |
| Israel | 630 | .4 | .4 | 30.0 |
| Italy | 23478 | 15.6 | 15.6 | 45.5 |
| Japan | 2 | .0 | .0 | 45.5 |
| Lebanon | 37 | .0 | .0 | 45.6 |
| Lithuania | 8 | .0 | .0 | 45.6 |
| Luxembourg | 9 | .0 | .0 | 45.6 |
| Macedonia | 16 | .0 | .0 | 45.6 |
| Mexico | 63 | .0 | .0 | 45.6 |
| Moldova | 71 | .0 | .0 | 45.7 |
| Montenegro | 2 | .0 | .0 | 45.7 |
| Morocco | 12 | .0 | .0 | 45.7 |
| New Zealand | 3320 | 2.2 | 2.2 | 47.9 |
| Portugal | 5322 | 3.5 | 3.5 | 51.4 |
| Romania | 139 | .1 | .1 | 51.5 |
| Serbia | 14 | .0 | .0 | 51.5 |
| Slovakia | 3 | .0 | .0 | 51.5 |
| Slovenia | 94 | .1 | .1 | 51.6 |
| South Africa | 2258 | 1.5 | 1.5 | 53.1 |
| South Korea | 4 | .0 | .0 | 53.1 |
| Spain | 8268 | 5.5 | 5.5 | 58.6 |
| Switzerland | 4 | .0 | .0 | 58.6 |
| Tunisia | 2 | .0 | .0 | 58.6 |
| Turkey | 52 | .0 | .0 | 58.6 |
| Ukraine | 5 | .0 | .0 | 58.6 |
| Uruguay | 92 | .1 | .1 | 58.7 |
| US | 62397 | 41.3 | 41.3 | 100.0 |
| US-France | 1 | .0 | .0 | 100.0 |
| 合计 | 150935 | 100.0 | 100.0 |  |

Designation: 数据量过大见输出3.pdf

Points:

| **points** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  |  | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 | 8 | 1 | .0 | .0 | .0 |
| 10 | 1 | .0 | .0 | .0 |
| 25 | 1 | .0 | .0 | .0 |
| 35 | 1 | .0 | .0 | .0 |
| 71 | 1 | .0 | .0 | .0 |
| 80 | 898 | .6 | .6 | .6 |
| 81 | 1501 | 1.0 | 1.0 | 1.6 |
| 82 | 4040 | 2.7 | 2.7 | 4.3 |
| 83 | 6046 | 4.0 | 4.0 | 8.3 |
| 84 | 10706 | 7.1 | 7.1 | 15.4 |
| 85 | 12409 | 8.2 | 8.2 | 23.6 |
| 86 | 15569 | 10.3 | 10.3 | 33.9 |
| 87 | 20742 | 13.7 | 13.7 | 47.7 |
| 88 | 17869 | 11.8 | 11.8 | 59.5 |
| 89 | 12918 | 8.6 | 8.6 | 68.1 |
| 90 | 15972 | 10.6 | 10.6 | 78.6 |
| 91 | 10535 | 7.0 | 7.0 | 85.6 |
| 92 | 9240 | 6.1 | 6.1 | 91.7 |
| 93 | 6017 | 4.0 | 4.0 | 95.7 |
| 94 | 3460 | 2.3 | 2.3 | 98.0 |
| 95 | 1716 | 1.1 | 1.1 | 99.2 |
| 96 | 695 | .5 | .5 | 99.6 |
| 97 | 365 | .2 | .2 | 99.9 |
| 98 | 131 | .1 | .1 | 100.0 |
| 99 | 50 | .0 | .0 | 100.0 |
| 100 | 22 | .0 | .0 | 100.0 |
| 合计 | 150906 | 100.0 | 100.0 |  |
| 缺失 | 系统 | 29 | .0 |  |  |
| 合计 | | 150935 | 100.0 |  |  |

Price: 数据量过大见输出4.pdf

Province: 数据量过大见输出5.pdf

Region\_1: 数据量过大见输出6.pdf

Region\_2:

| **region\_2** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  |  | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 |  | 89995 | 59.6 | 59.6 | 59.6 |
| California Other | 3514 | 2.3 | 2.3 | 62.0 |
| Central Coast | 13053 | 8.6 | 8.6 | 70.6 |
| Central Valley | 1115 | .7 | .7 | 71.3 |
| Chardonnay | 2 | .0 | .0 | 71.3 |
| Columbia Valley | 9154 | 6.1 | 6.1 | 77.4 |
| Finger Lakes | 1510 | 1.0 | 1.0 | 78.4 |
| Long Island | 770 | .5 | .5 | 78.9 |
| Mendocino/Lake Counties | 2389 | 1.6 | 1.6 | 80.5 |
| Napa | 8798 | 5.8 | 5.8 | 86.3 |
| Napa-Sonoma | 1645 | 1.1 | 1.1 | 87.4 |
| Nebbiolo | 1 | .0 | .0 | 87.4 |
| New York Other | 147 | .1 | .1 | 87.5 |
| North Coast | 632 | .4 | .4 | 87.9 |
| Oregon Other | 659 | .4 | .4 | 88.4 |
| Pinot Grigio | 1 | .0 | .0 | 88.4 |
| Sierra Foothills | 1660 | 1.1 | 1.1 | 89.5 |
| Sonoma | 11255 | 7.5 | 7.5 | 96.9 |
| Sonoma County | 1 | .0 | .0 | 96.9 |
| South Coast | 198 | .1 | .1 | 97.1 |
| Southern Oregon | 662 | .4 | .4 | 97.5 |
| Washington Other | 593 | .4 | .4 | 97.9 |
| Willamette Valley | 3180 | 2.1 | 2.1 | 100.0 |
| Zinfandel | 1 | .0 | .0 | 100.0 |
| 合计 | 150935 | 100.0 | 100.0 |  |

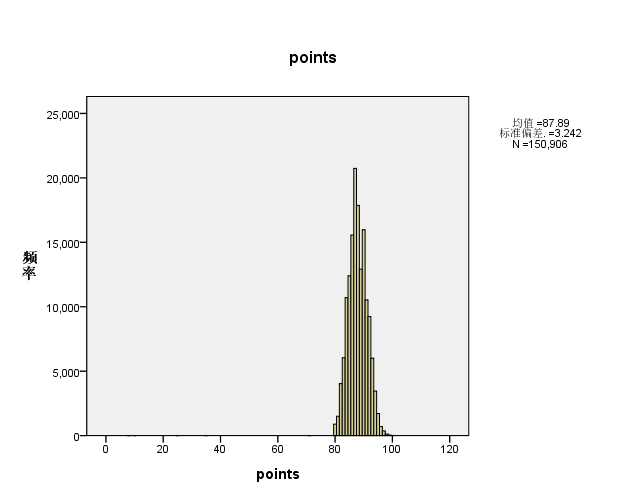
Variety: 数据量过大见输出7.pdf

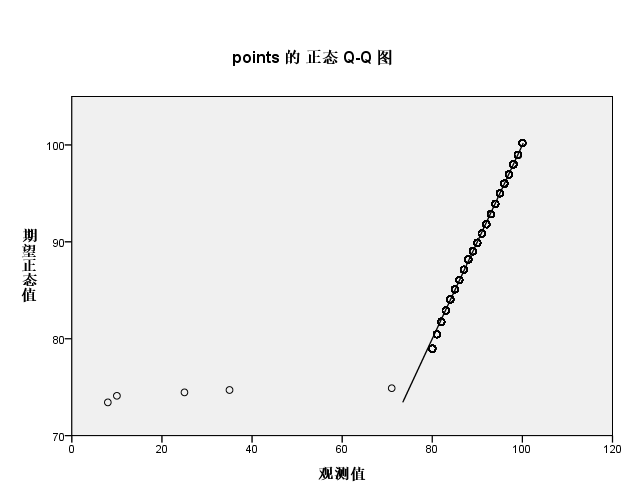
Winery: 数据量过大见输出8.pdf

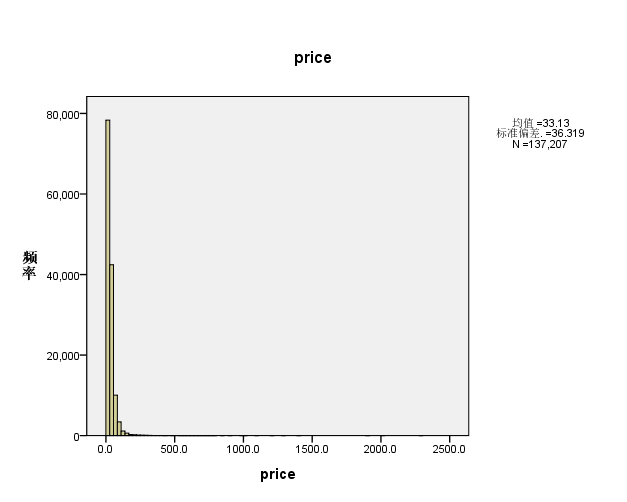
1. 数值属性，给出最大、最小、均值、中位数、四分位数及缺失值的个数

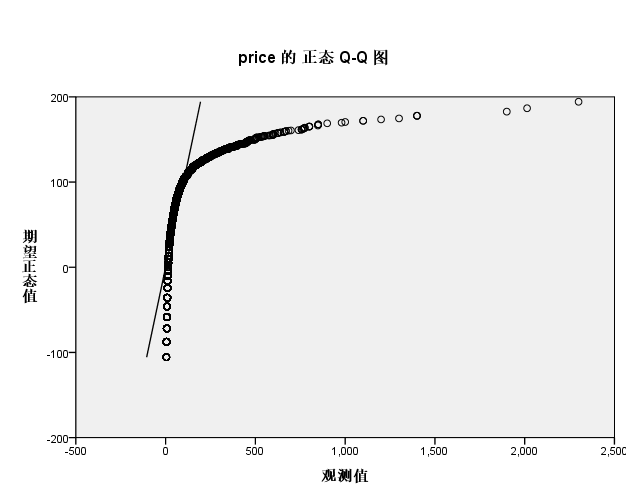
| **统计量** | | | |
| --- | --- | --- | --- |
|  |  | points | price |
| N | 有效 | 150906 | 137207 |
| 缺失 | 29 | 13728 |
| 均值 | | 87.89 | 33.130 |
| 中值 | | 88.00 | 24.000 |
| 极小值 | | 8 | 4.0 |
| 极大值 | | 100 | 2300.0 |
| 百分位数 | 25 | 86.00 | 16.000 |
| 50 | 88.00 | 24.000 |
| 75 | 90.00 | 40.000 |

1. 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。

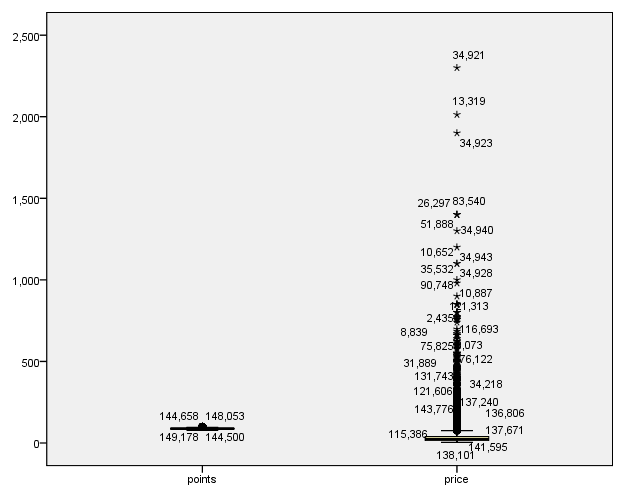








1. 绘制盒图，对离群值进行识别



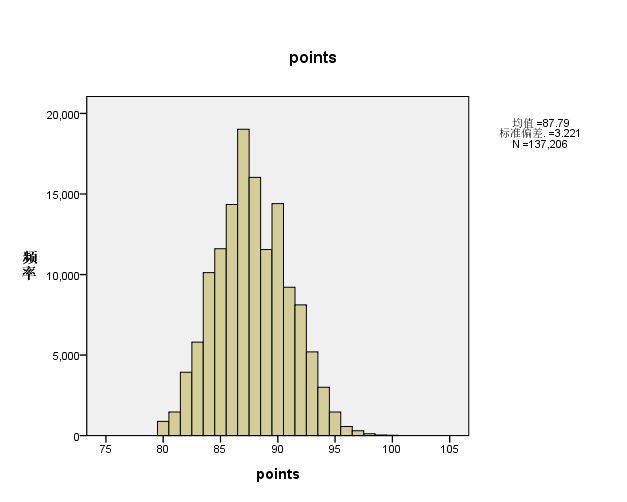
**数据缺失的处理**

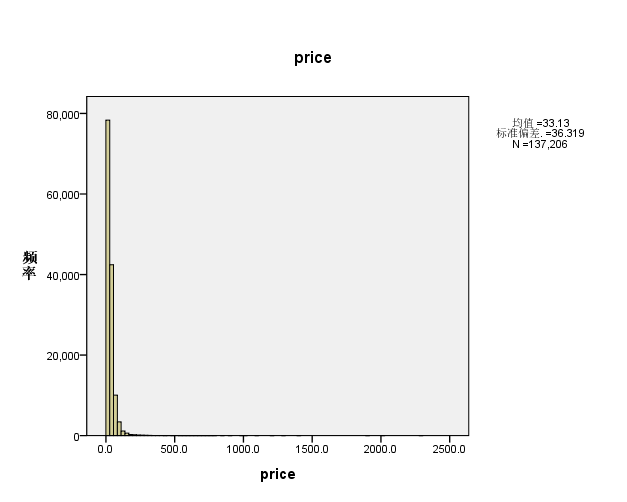
1. 观察数据集中缺失数据，分析其缺失的原因。

| **统计量** | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | points | price | country | designation | province | region\_1 | region\_2 | variety | winery |
| N | 有效 | 150906 | 137207 | 150935 | 150935 | 150935 | 150935 | 150935 | 150935 | 150935 |
| 缺失 | 29 | 13728 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

1. 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

