姓名：李欣悦 学号：5720182039

本次作业使用spss软件对数据基本处理，未使用编程，处理过程及结果如下所示。

**第一个数据集数据集：**[**Wine Reviews**](https://www.kaggle.com/zynicide/wine-reviews)

**数据摘要**

* 对标称属性，给出每个可能取值的频数

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **country** | | | | | |
|  | | 频率 | 百分比 | 有效百分比 | 累计百分比 |
| 有效 |  | 63 | .0 | .0 | .0 |
| Argentina | 3798 | 2.9 | 2.9 | 3.0 |
| Armenia | 2 | .0 | .0 | 3.0 |
| Australia | 2329 | 1.8 | 1.8 | 4.8 |
| Austria | 3344 | 2.6 | 2.6 | 7.3 |
| Bosnia and Herzegovina | 2 | .0 | .0 | 7.3 |
| Brazil | 52 | .0 | .0 | 7.4 |
| Bulgaria | 141 | .1 | .1 | 7.5 |
| Canada | 257 | .2 | .2 | 7.7 |
| Chile | 4471 | 3.4 | 3.4 | 11.1 |
| China | 1 | .0 | .0 | 11.1 |
| Croatia | 73 | .1 | .1 | 11.2 |
| Cyprus | 11 | .0 | .0 | 11.2 |
| Czech Republic | 12 | .0 | .0 | 11.2 |
| Egypt | 1 | .0 | .0 | 11.2 |
| England | 74 | .1 | .1 | 11.3 |
| France | 22089 | 17.0 | 17.0 | 28.3 |
| Georgia | 86 | .1 | .1 | 28.3 |
| Germany | 2165 | 1.7 | 1.7 | 30.0 |
| Greece | 466 | .4 | .4 | 30.3 |
| Hungary | 146 | .1 | .1 | 30.5 |
| India | 9 | .0 | .0 | 30.5 |
| Israel | 505 | .4 | .4 | 30.9 |
| Italy | 19537 | 15.0 | 15.0 | 45.9 |
| Lebanon | 35 | .0 | .0 | 45.9 |
| Luxembourg | 6 | .0 | .0 | 45.9 |
| Macedonia | 12 | .0 | .0 | 45.9 |
| Mexico | 70 | .1 | .1 | 46.0 |
| Moldova | 59 | .0 | .0 | 46.0 |
| Morocco | 28 | .0 | .0 | 46.1 |
| New Zealand | 1419 | 1.1 | 1.1 | 47.1 |
| Peru | 16 | .0 | .0 | 47.2 |
| Portugal | 5690 | 4.4 | 4.4 | 51.5 |
| Romania | 120 | .1 | .1 | 51.6 |
| Serbia | 12 | .0 | .0 | 51.6 |
| Slovakia | 1 | .0 | .0 | 51.6 |
| Slovenia | 87 | .1 | .1 | 51.7 |
| South Africa | 1401 | 1.1 | 1.1 | 52.8 |
| Spain | 6645 | 5.1 | 5.1 | 57.9 |
| Switzerland | 7 | .0 | .0 | 57.9 |
| Turkey | 90 | .1 | .1 | 58.0 |
| Ukraine | 14 | .0 | .0 | 58.0 |
| Uruguay | 109 | .1 | .1 | 58.1 |
| US | 54496 | 41.9 | 41.9 | 100.0 |
| 总计 | 129951 | 100.0 | 100.0 |  |

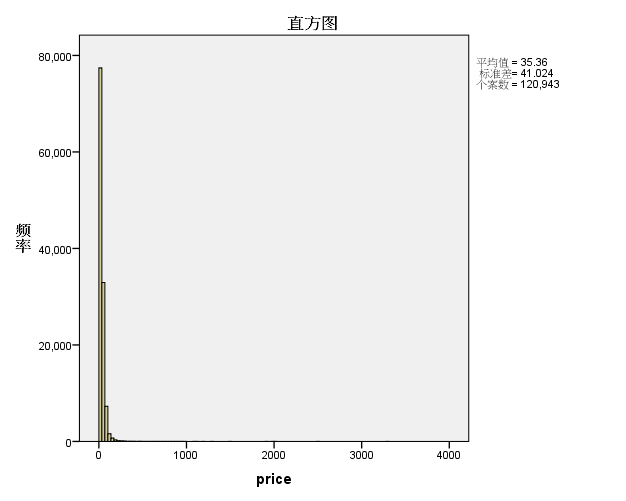
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **points** | | | | | |
|  | | 频率 | 百分比 | 有效百分比 | 累计百分比 |
| 有效 | 80 | 397 | .3 | .3 | .3 |
| 81 | 689 | .5 | .5 | .8 |
| 82 | 1833 | 1.4 | 1.4 | 2.2 |
| 83 | 3021 | 2.3 | 2.3 | 4.6 |
| 84 | 6477 | 5.0 | 5.0 | 9.6 |
| 85 | 9529 | 7.3 | 7.3 | 16.9 |
| 86 | 12600 | 9.7 | 9.7 | 26.6 |
| 87 | 16930 | 13.0 | 13.0 | 39.6 |
| 88 | 17201 | 13.2 | 13.2 | 52.9 |
| 89 | 12224 | 9.4 | 9.4 | 62.3 |
| 90 | 15410 | 11.9 | 11.9 | 74.1 |
| 91 | 11355 | 8.7 | 8.7 | 82.9 |
| 92 | 9611 | 7.4 | 7.4 | 90.3 |
| 93 | 6487 | 5.0 | 5.0 | 95.3 |
| 94 | 3755 | 2.9 | 2.9 | 98.1 |
| 95 | 1535 | 1.2 | 1.2 | 99.3 |
| 96 | 523 | .4 | .4 | 99.7 |
| 97 | 228 | .2 | .2 | 99.9 |
| 98 | 77 | .1 | .1 | 100.0 |
| 99 | 33 | .0 | .0 | 100.0 |
| 100 | 18 | .0 | .0 | 100.0 |
| 总计 | 129933 | 100.0 | 100.0 |  |
| 缺失 | 系统 | 18 | .0 |  |  |
| 总计 | | 129951 | 100.0 |  |  |

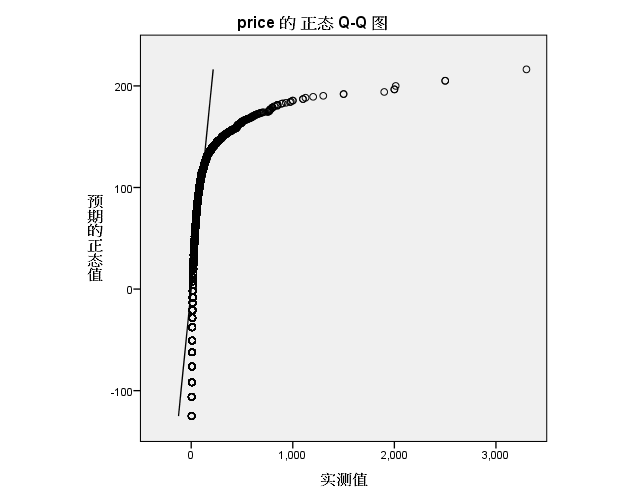
* 数值属性，给出最大、最小、均值、中位数、四分位数及缺失值的个数。

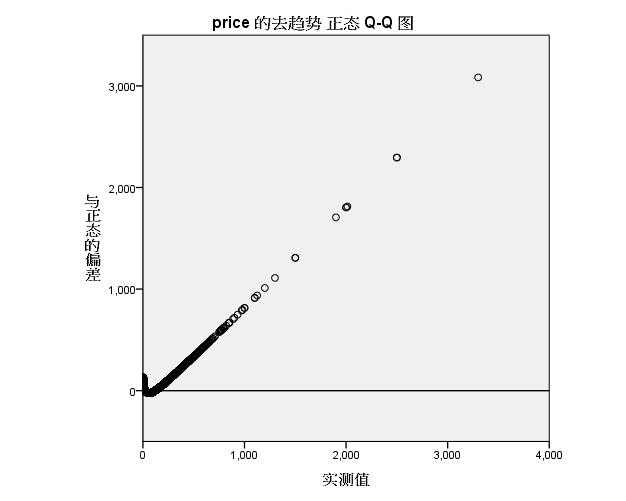
|  |  |  |
| --- | --- | --- |
| **统计** | | |
| price | | |
| 个案数 | 有效 | 120943 |
| 缺失 | 9008 |
| 平均值 | | 35.36 |
| 中位数 | | 25.00 |
| 最小值 | | 4 |
| 最大值 | | 3300 |
| 百分位数 | 25 | 17.00 |
| 50 | 25.00 |
| 75 | 42.00 |

**数据的可视化**

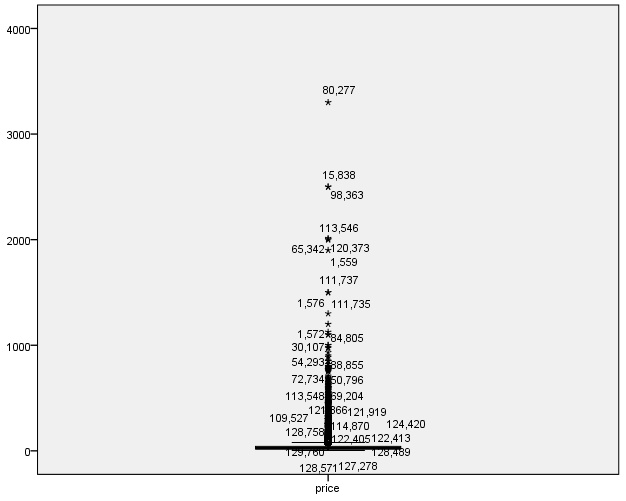
针对数值属性，

* 绘制直方图，用qq图检验其分布是否为正态分布。
* 





* 绘制盒图，对离群值进行识别



**3.2 数据缺失的处理**

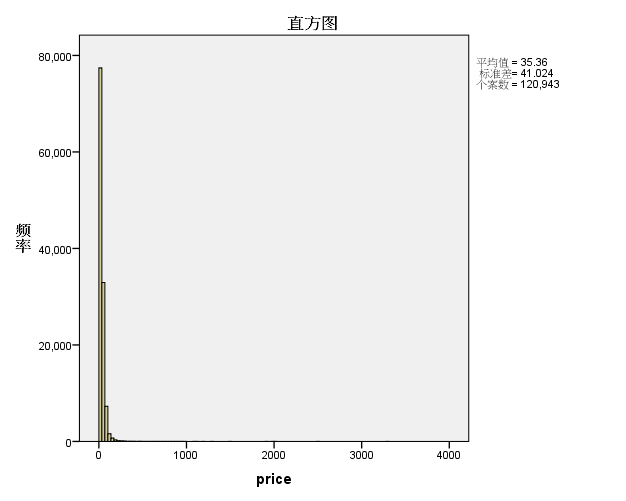
观察数据集中缺失数据，分析其缺失的原因。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **单变量统计** | | | | | | | |
|  | 个案数 | 平均值 | 标准差 | 缺失 | | 极值数a | |
| 计数 | 百分比 | 低 | 高 |
| price | 120943 | 35.36 | 41.024 | 9008 | 6.9 | 0 | 2675 |

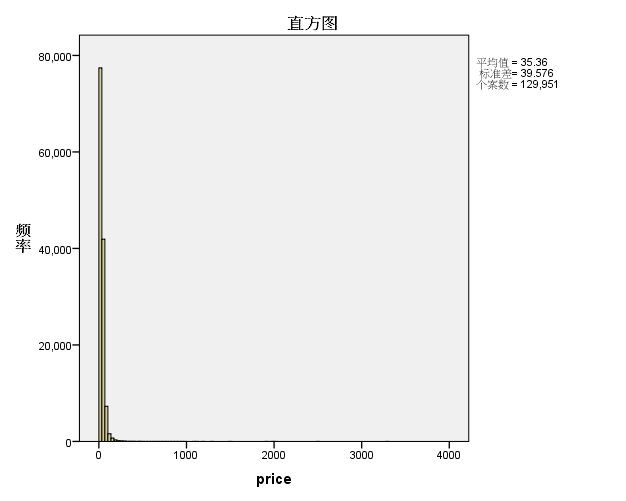
|  |  |  |  |
| --- | --- | --- | --- |
| **估算平均值摘要** | | | |
|  | | price | |
| 所有值 | | 35.36 | |
| EM | | 35.36 | |
| 回归 | | 35.41 | |
| **估算标准差摘要** | | | |
|  | | price | |
| 所有值 | | 41.024 | |
| EM | | 41.024 | |
| 回归 | | 41.612 | |
| **EM 平均值a** | | | | |
| **EM 协方差a** | | |
|  | price | |
| price | 1682.954 | |

分别使用下列四种策略对缺失值进行处理:

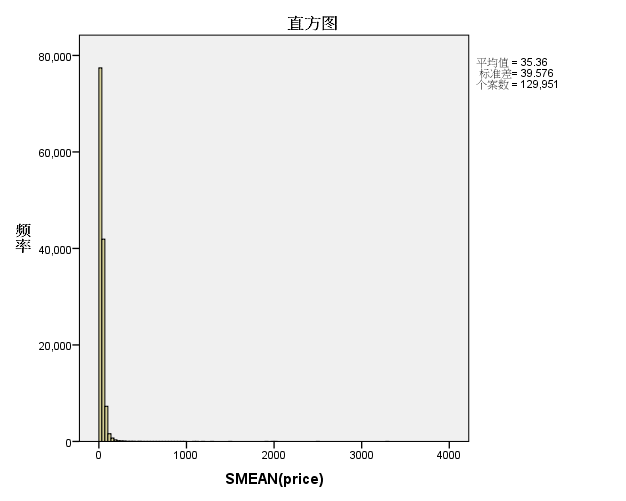
* 将缺失部分剔除后数据集直方图如下：



* 通过属性的相关关系来填补缺失值，数据集直方图如下：



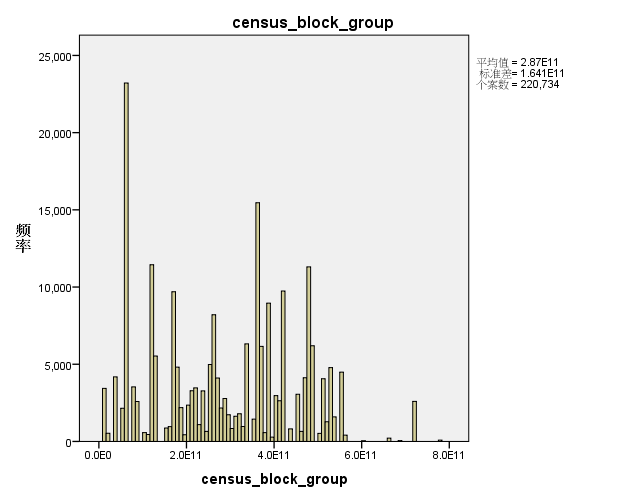
* 通过数据对象之间的相似性来填补缺失值，数据集直方图如下：



**第二个数据集：**[Consumer & Visitor Insights For Neighborhoods](https://www.kaggle.com/safegraph/visit-patterns-by-census-block-group)

**数据摘要**

* 对标称属性，给出每个可能取值的频数



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **date\_range\_start** | | | | | |
|  | | 频率 | 百分比 | 有效百分比 | 累计百分比 |
| 有效 | 1538352000 | 220735 | 100.0 | 100.0 | 100.0 |

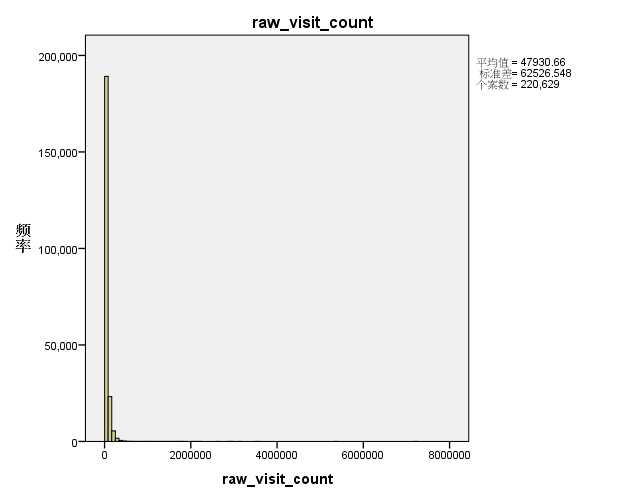
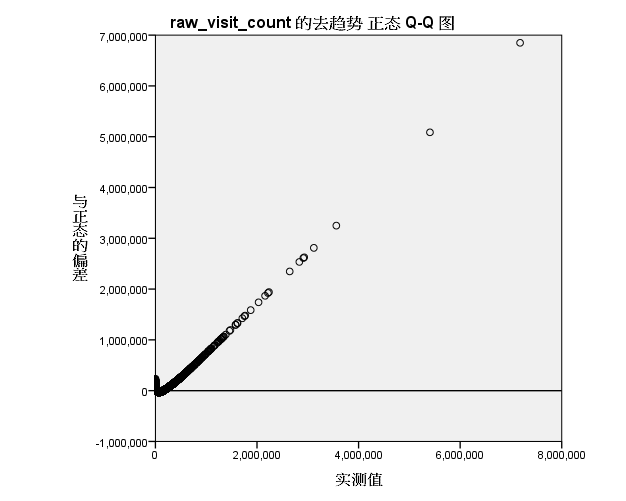
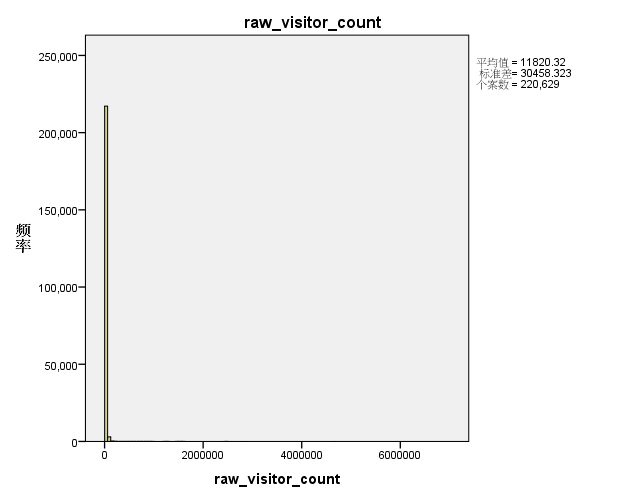
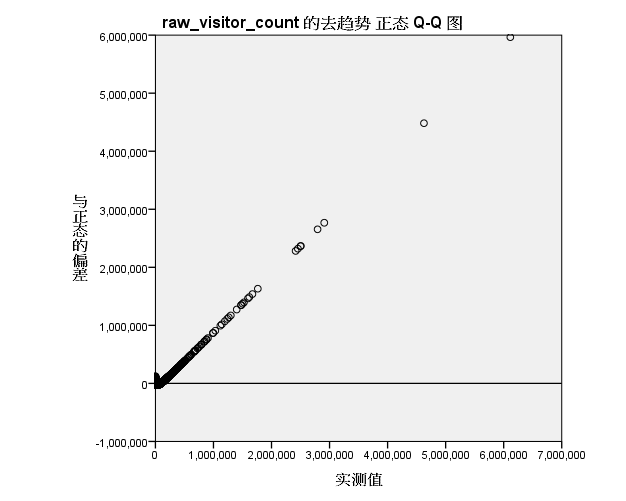
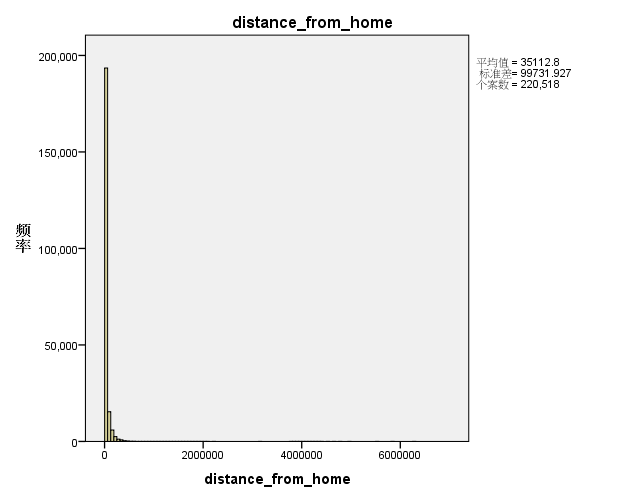
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **date\_range\_end** | | | | | |
|  | | 频率 | 百分比 | 有效百分比 | 累计百分比 |
| 有效 | 1541030400 | 220735 | 100.0 | 100.0 | 100.0 |

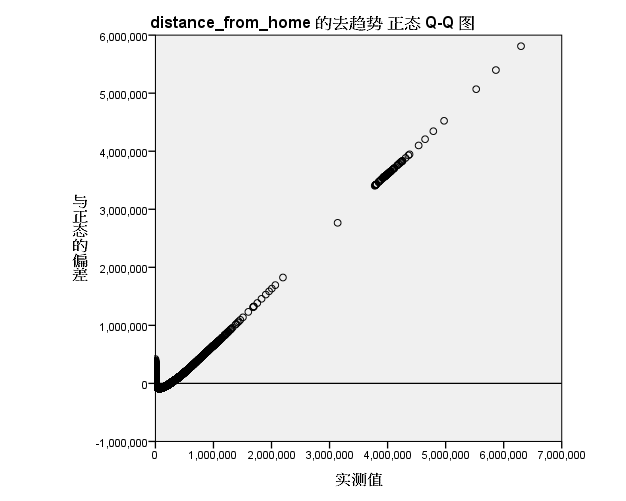
* 数值属性，给出最大、最小、均值、中位数、四分位数及缺失值的个数。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **统计** | | | | |
|  | | raw\_visit\_count | raw\_visitor\_count | distance\_from\_home |
| 个案数 | 有效 | 220629 | 220629 | 220518 |
| 缺失 | 106 | 106 | 217 |
| 平均值 | | 47930.66 | 11820.32 | 35112.80 |
| 中位数 | | 30640.00 | 6541.00 | 14614.00 |
| 最小值 | | 60 | 50 | 706 |
| 最大值 | | 7179900 | 6113949 | 6297845 |
| 百分位数 | 25 | 17042.00 | 3430.00 | 8584.00 |
| 50 | 30640.00 | 6541.00 | 14614.00 |
| 75 | 56678.00 | 13099.00 | 31398.00 |

**数据的可视化**

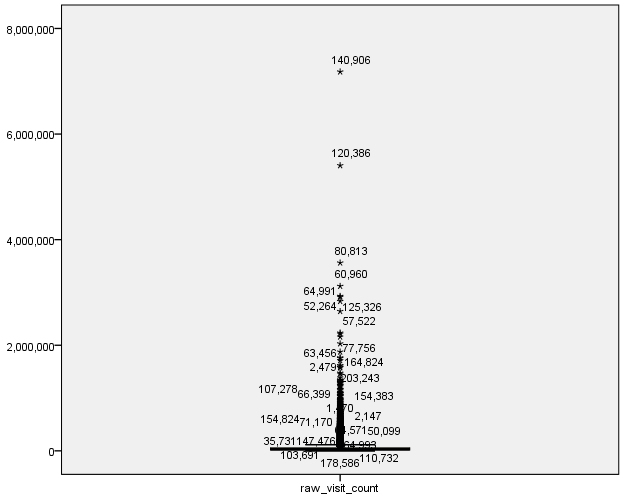
针对数值属性，

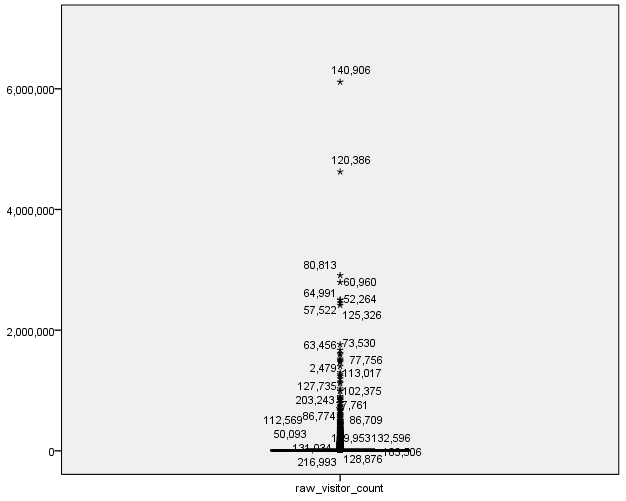
* 绘制直方图，用qq图检验其分布是否为正态分布。
* 
* 
* 
* 
* 



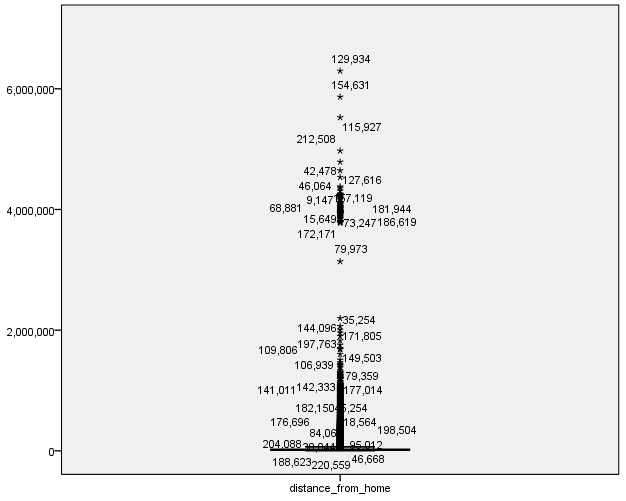
* 绘制盒图，对离群值进行识别

Raw\_visit\_count盒图：

Raw\_visitor\_count盒图：



Distance\_from\_home盒图：

* 

**3.2 数据缺失的处理**

观察数据集中缺失数据，分析其缺失的原因。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **单变量统计** | | | | | | | |
|  | 个案数 | 平均值 | 标准差 | 缺失 | | 极值数a | |
| 计数 | 百分比 | 低 | 高 |
| raw\_visit\_count | 220629 | 47930.66 | 62526.548 | 106 | .0 | 0 | 7569 |
| raw\_visitor\_count | 220629 | 11820.32 | 30458.323 | 106 | .0 | 0 | 2157 |
| distance\_from\_home | 220518 | 35112.80 | 99731.927 | 217 | .1 | 0 | 3657 |
| |  |  |  |  | | --- | --- | --- | --- | | **估算平均值摘要** | | | | |  | raw\_visit\_count | raw\_visitor\_count | distance\_from\_home | | 所有值 | 47930.66 | 11820.32 | 35112.80 | | EM | 47930.66 | 11820.32 | 35112.52 | | 回归 | 47924.54 | 11818.93 | 35108.42 |  |  |  |  |  | | --- | --- | --- | --- | | **估算标准差摘要** | | | | |  | raw\_visit\_count | raw\_visitor\_count | distance\_from\_home | | 所有值 | 62526.548 | 30458.323 | 99731.927 | | EM | 62526.548 | 30458.323 | 99731.898 | | 回归 | 62516.450 | 30451.949 | 99692.224 | | | | | | | | |

|  |  |  |
| --- | --- | --- |
| **EM 平均值a** | | |
| raw\_visit\_count | raw\_visitor\_count | distance\_from\_home |
| 47930.66 | 11820.32 | 35112.52 |

|  |  |  |  |
| --- | --- | --- | --- |
| **EM 协方差a** | | | |
|  | raw\_visit\_count | raw\_visitor\_count | distance\_from\_home |
| raw\_visit\_count | 3909569211.000 |  |  |
| raw\_visitor\_count | 1553749457.000 | 927709420.000 |  |
| distance\_from\_home | 135407311.500 | 99849563.160 | 9946451549.000 |
| |  |  |  |  | | --- | --- | --- | --- | | **EM 相关性a** | | | | |  | raw\_visit\_count | raw\_visitor\_count | distance\_from\_home | | raw\_visit\_count | 1 |  |  | | raw\_visitor\_count | .816 | 1 |  | | distance\_from\_home | .022 | .033 | 1 | | a. 利特尔 MCAR 检验：卡方 = 82.788，自由度 = 2，重要性 = .000 | | | | | | | |

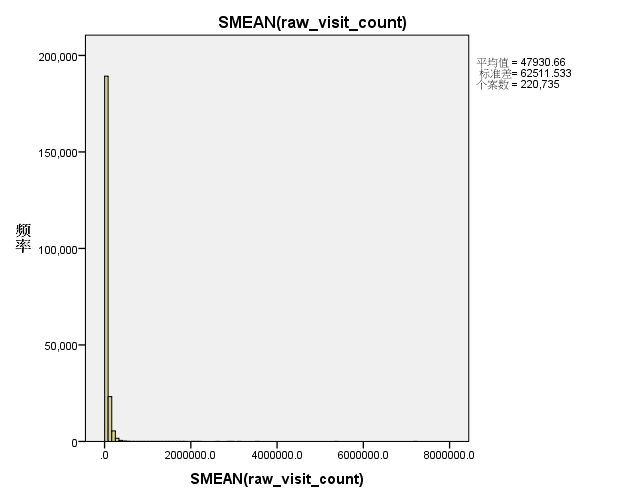
分别使用下列四种策略对缺失值进行处理:

* 将raw\_visitor\_count、raw\_visit\_count、distance\_from\_home缺失部分剔除后数据集统计结果如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **统计** | | | | |
|  | | raw\_visit\_count | raw\_visitor\_count | distance\_from\_home |
| 个案数 | 有效 | 220518 | 220518 | 220518 |
| 缺失 | 0 | 0 | 0 |
| 平均值 | | 47954.56 | 11826.18 | 35112.80 |

* 通过数据对象之间的相似性来填补缺失值

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **结果变量** | | | | | | |
|  | 结果变量 | 替换的缺失值数 | 非缺失值的个案编号 | | 有效个案数 | 创建函数 |
| 第一个 | 最后一个 |
| 1 | raw\_visit\_count\_1 | 106 | 1 | 220735 | 220735 | SMEAN(raw\_visit\_count) |
| 2 | raw\_visitor\_count\_1 | 106 | 1 | 220735 | 220735 | SMEAN(raw\_visitor\_count) |
| 3 | distance\_from\_home\_1 | 217 | 1 | 220735 | 220735 | SMEAN(distance\_from\_home) |

* 使用直方图表示替换后的数据集：
* 
* 