**数据挖掘分析报告**

1. Wine Reviews

* 数据可视化与摘要：  
  1.对标称属性，给出每个可能取值的频数

|  |  |
| --- | --- |
| **country** | |
|  |  |
| **designation** | **province** |
|  |  |
| region\_1 | region\_2 |
|  |  |
| variety | winery |
|  |  |

2.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。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 属性名称 | 最大值 | 最小值 | 中位数 | 平均值 | 四分位数 | 缺失值 |
| price | 2300.0 | 4.0 | 24.0 | 33.1314 | （48，16） | 13695 |
| 属性名称 | 最大值 | 最小值 | 中位数 | 平均值 | 四分位数 | 缺失值 |
| points | 100 | 80 | 26.0 | 38.1914 | （90，86） | 0 |

1. 绘制直方图，用qq图检验其分布是否为正态分布。

（1）：points

|  |  |
| --- | --- |
| 直方图 | qq图 |
|  |  |
| 盒图 | 以上为points属性的三种图像。根据qq图得知points属于正态分布，通过盒图识别离群值。 |
|  |

1. 、price：

|  |  |
| --- | --- |
| 直方图 | qq图 |
|  |  |
| 盒图 | 以上为price属性的三种图像。根据qq图得知price不属于正态分布，通过盒图识别离群值。 |
|  |

* 数据缺失的处理：

1. 将缺失部分剔除：

|  |  |
| --- | --- |
| 直方图 | qq图 |
|  |  |
| 盒图 | 以上为删除缺失行后price属性的三种图像。 |
|  |

1. 用最高频率值来填补缺失值：

|  |  |
| --- | --- |
| 直方图 | qq图 |
|  |  |
| 盒图 | 以上为用最高频率值来填补缺失值后price属性的三种图像。 |
|  |

1. 通过属性的相关关系填补缺失值：

|  |  |
| --- | --- |
| 直方图 | qq图 |
|  |  |
| 盒图 | 以上为通过属性的相关关系填补缺失值后price属性的三种图像。 |
|  |

1. 通过数据对象之间的相似性填补缺失值：

|  |  |
| --- | --- |
| 直方图 | qq图 |
|  |  |
| 盒图 | 以上为通过数据对象之间的相似性来填补缺失值后price属性的三种图像。 |
|  |

1. Youtube videos

* 数据可视化与摘要：  
  1.对标称属性，给出每个可能取值的频数

|  |  |
| --- | --- |
| **thumbnail\_link** | |
|  |  |
| **comments\_disabled** | **ratings\_disabled** |
|  |  |
| **video\_error\_or\_removed** |  |
|  |  |

2.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。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 属性名称 | 最大值 | 最小值 | 中位数 | 平均值 | 四分位数 | 缺失值 |
| views | 424538912 | 117 | 177370.5 | 1326568.26 | (647685.0, 46978.0) | 0 |
| 属性名称 | 最大值 | 最小值 | 中位数 | 平均值 | 四分位数 | 缺失值 |
| likes | 5613827 | 0 | 3446.0 | 37884.31 | (17477.0, 669.0) | 0 |
| 属性名称 | 最大值 | 最小值 | 中位数 | 平均值 | 四分位数 | 缺失值 |
| dislikes | 1944971 | 0 | 179.0 | 2126.11 | (749.0, 41.0) | 0 |
| 属性名称 | 最大值 | 最小值 | 中位数 | 平均值 | 四分位数 | 缺失值 |
| dislikes | 1626501 | 0 | 511.0 | 4253.78 | (2011.0, 109.0) | 0 |

1. 绘制直方图，用qq图检验其分布是否为正态分布。

（1）：views

|  |  |
| --- | --- |
| 直方图 | qq图 |
|  |  |
| 盒图 | 以上为viewts属性的三种图像。根据qq图得知views不属于正态分布，通过盒图识别离群值。 |
|  |

(2)、likes：

|  |  |
| --- | --- |
| 直方图 | qq图 |
|  |  |
| 盒图 | 以上为likes属性的三种图像。根据qq图得知likes不属于正态分布，通过盒图识别离群值。 |
|  |

(3)、dislikes：

|  |  |
| --- | --- |
| 直方图 | qq图 |
|  |  |
| 盒图 | 以上为dislikes属性的三种图像。根据qq图得知dislikes不属于正态分布，通过盒图识别离群值。 |

(4)、comment\_count：

|  |  |
| --- | --- |
| 直方图 | qq图 |
|  |  |
| 盒图 | 以上为comment\_count属性的三种图像。根据qq图得知comment\_count不属于正态分布，通过盒图识别离群值。 |

* 数据缺失的处理：无数据缺失