数据挖掘互评作业1：数据探索性分析与数据预处理

姓名：曹健 学号：3120190978

1. 数据可视化和摘要
2. 数据集选择：Wine Reviews 和 Consumer & Visitor Insights For Neighborhoods
3. 数据集描述：

**数据集 Wine Reviews：**

winemag-data-130k-v2.csv 包含13列和130k行葡萄酒评论

|  |  |  |
| --- | --- | --- |
| 序号 | 属性 | 含义 |
| 1 | Country | 葡萄酒产地 |
| 2 | Description | 介绍 |
| 3 | Designation | 所属葡萄园 |
| 4 | Points | WineEnthusiast对葡萄酒的评分 |
| 5 | Price | 单瓶葡萄酒成本 |
| 6 | Province | 葡萄酒来源的省或州 |
| 7 | Region\_1 | 一个省或州的葡萄酒种植区 |
| 8 | Region\_2 | 有时在葡萄酒种植区域中指定了更多特定区域，但此值有时可能为空 |
| 9 | Taster\_name |  |
| 10 | Taster\_twitter\_handle |  |
| 11 | Title | 评论的标题 |
| 12 | Variety | 葡萄种类 |
| 13 | Winery | 所属酿酒厂 |

标称属性：country，designation，province，region\_1，region\_2，taster\_name，taster\_twitter\_handle，variety，winery；

数值属性：points，price

winemag-data\_first150k.csv包含10列和150k行葡萄酒评论

与上一个数据集相比缺少taster\_name，taster\_twitter\_handle，title三个属性。

标称属性：country，designation，province，region\_1，region\_2，variety，winery；

数值属性：points，price

**数据集 Consumer & Visitor Insights For Neighborhoods：**

cbg\_patterns.csv文件，包含13列22万条数据

|  |  |  |
| --- | --- | --- |
| 序号 | 属性 | 含义 |
| 1 | census\_block\_group | 人口普查区组唯一的12位FIPS码 |
| 2 | date\_range\_start | 测量周期的开始时间作为时间戳(UTC秒) |
| 3 | date\_range\_end | 测量周期的时间结束作为时间戳(UTC秒) |
| 4 | raw\_visit\_count | 在此日期范围内，我们的专家组对该CBG的访问次数 |
| 5 | raw\_visitor\_count | 在日期范围内，我们的小组访问POI的独立访问人数 |
| 6 | visitor\_home\_cbgs | 此列列出了在census\_block\_group列（目标CBG）中列出的CBG中访问过目的地的设备的所有原始归属CBG |
| 7 | visitor\_work\_cbgs | 此列列出了在census\_block\_group列（目标CBG）中列出的CBG中访问过目标的设备的所有工作位置CBG。 |
| 8 | distance\_from\_home | 访客（我们确定住所的访客）到CBG的平均距离（以米为单位） |
| 9 | related\_same\_day\_brand | 这些CBG的访问者在访问CBG的同一天访问的品牌 |
| 10 | related\_same\_month\_brand | 访问CBG的客户在访问CBG的同一个月访问的品牌 |
| 11 | top\_brands | 在此期间，访问过CBG的顶级品牌列表。 仅限十大品牌。 |
| 12 | popularity\_by\_hour | 在当地时间范围内，每天的小时数与每小时的访问量之间的映射 |
| 13 | popularity\_by\_day | 在日期范围内，将星期几映射为每天的访问量(当地时间)。 |

标称属性：census\_block\_group，date\_range\_start，date\_range\_end，visitor\_home\_cbgs，visitor\_work\_cbgs，related\_same\_day\_brand，related\_same\_month\_brand，top\_brands，popularity\_by\_hour，popularity\_by\_day；

数值属性：raw\_visit\_count，raw\_visitor\_count，distance\_from\_home

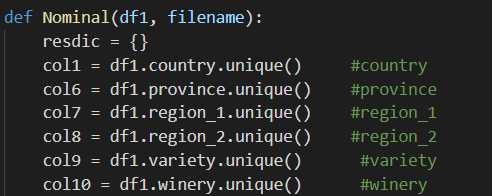
1. 数据集摘要（**以数据集 Wine Reviews为例**）

**针对标称属性**，给出每个可能取值的频数

使用pandas包的read\_csv()方法将csv文件输入并存为dataframe格式。



按列分别取出各标称属性：



然后通过list类型统计频数：

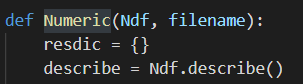


统计结果的部分截图：



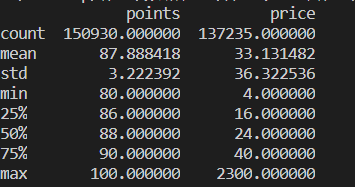
**针对数值属性**，给出5数概括及缺失值的个数（最大值、最小值、均值、中位数、四分位数）

选择需要处理的数值属性的两列构建新的dataframe，调用describe()方法得到对数值属性列的重要指标描述。

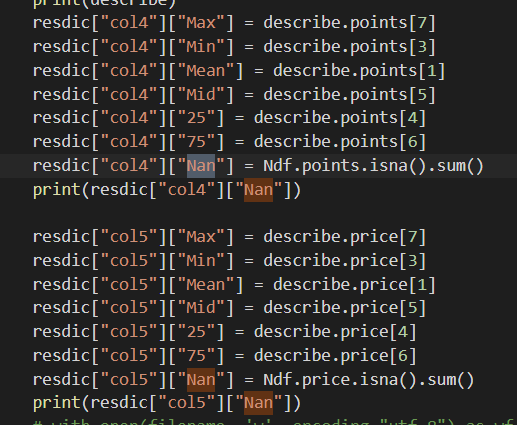




得到的指标值如下所示：

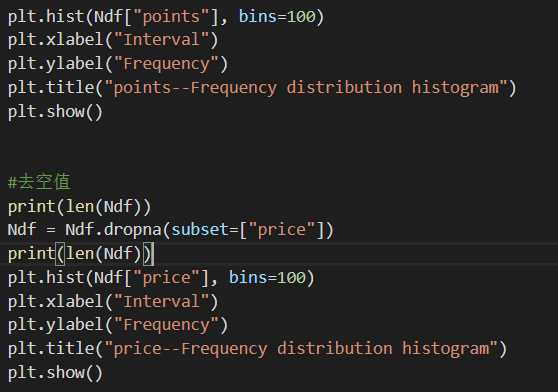


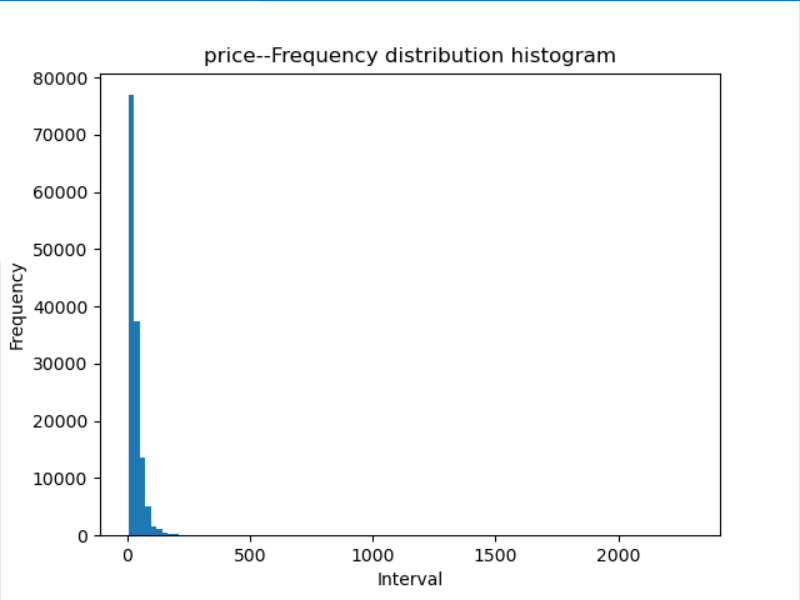
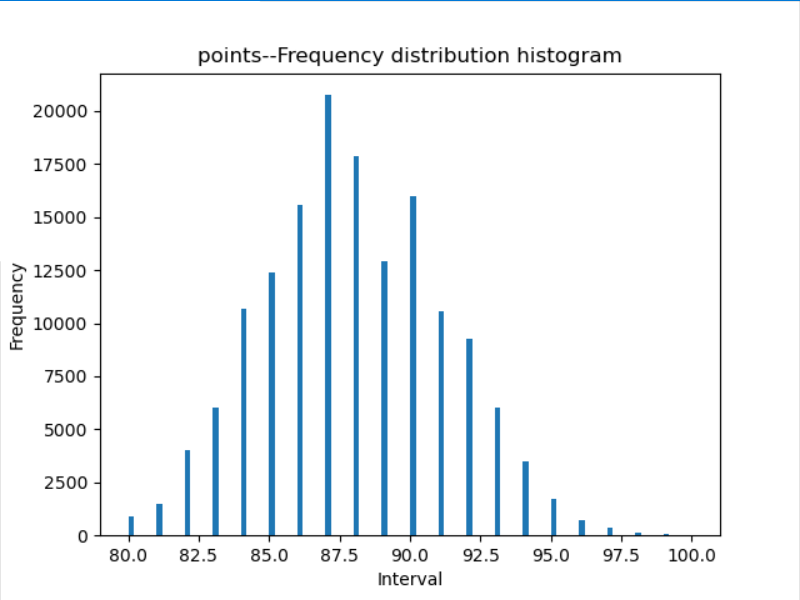
同时通过对每一个数值属性列判断空值并统计个数，得到缺失值的个数：



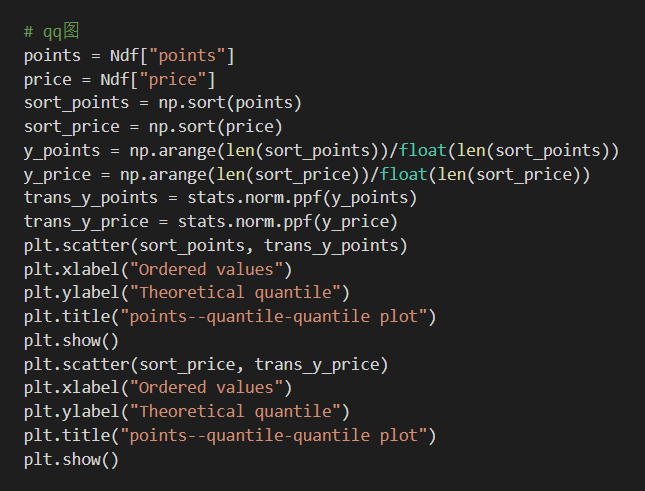
1. 数据可视化（**以数据集 Wine Reviews为例**）

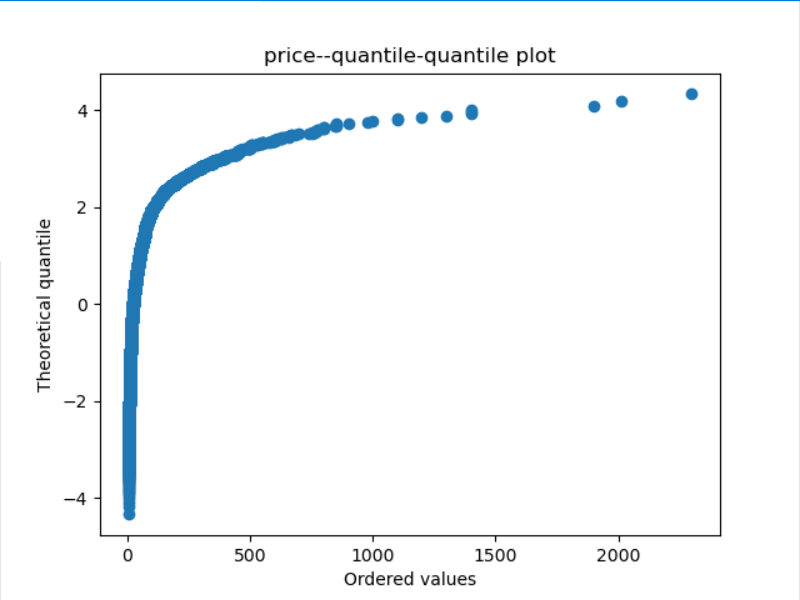
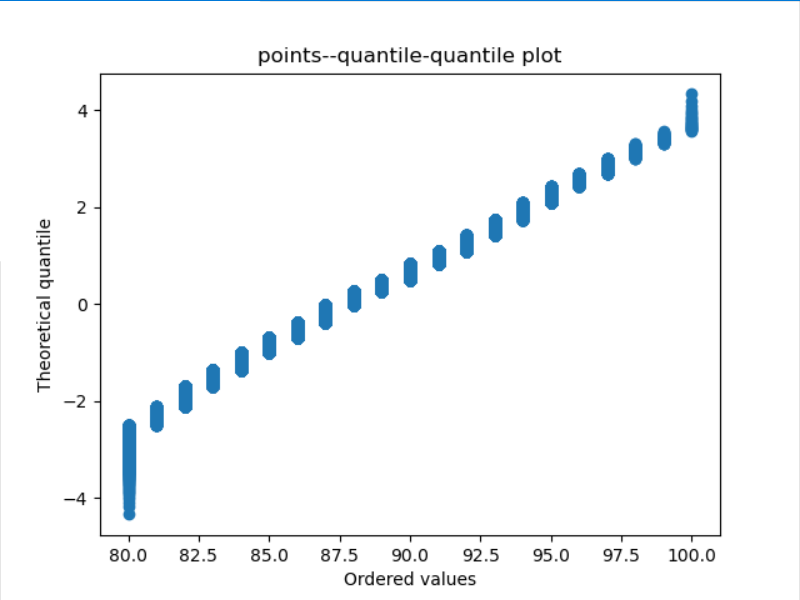
绘制直方图：



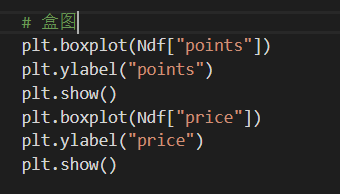


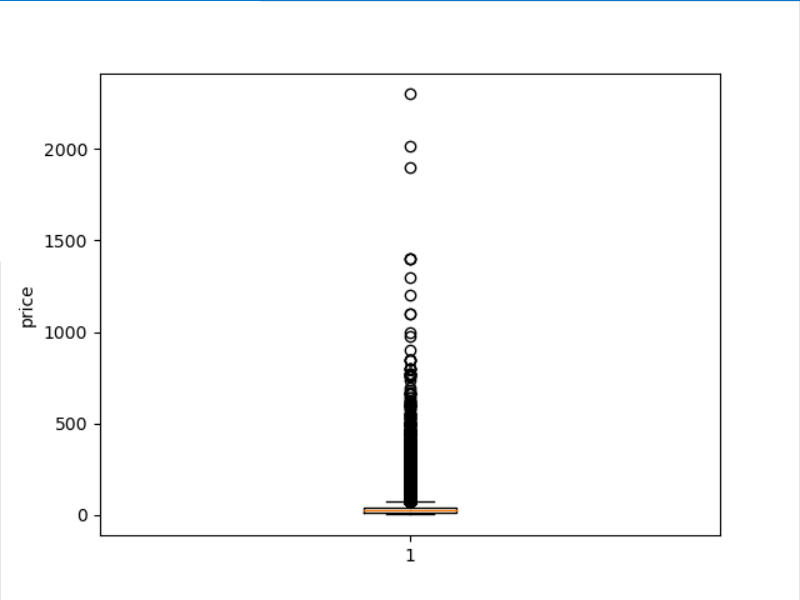
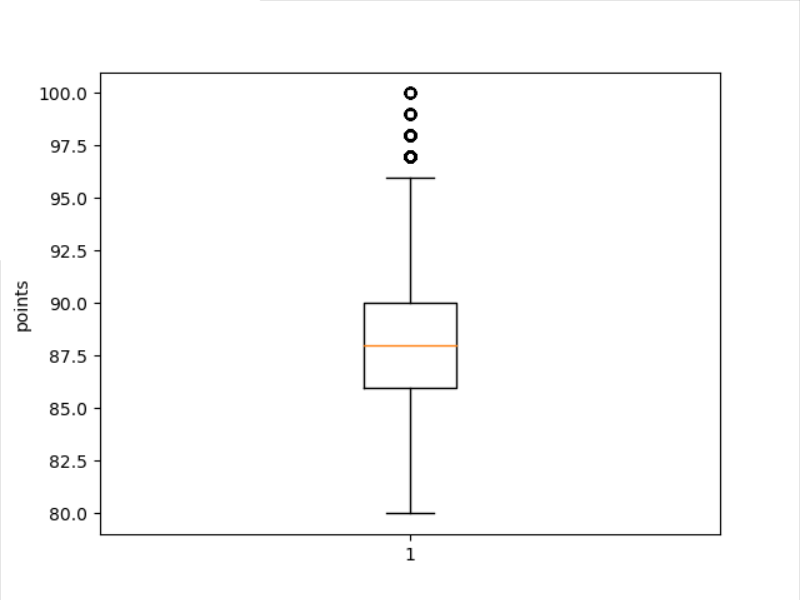
绘制QQ图检验数据分布是否为正态分布：





绘制盒图，对离群值进行识别：





1. 数据缺失的处理（**以数据集 Wine Reviews为例**）

**文件1：winemag-data\_first150k.csv**

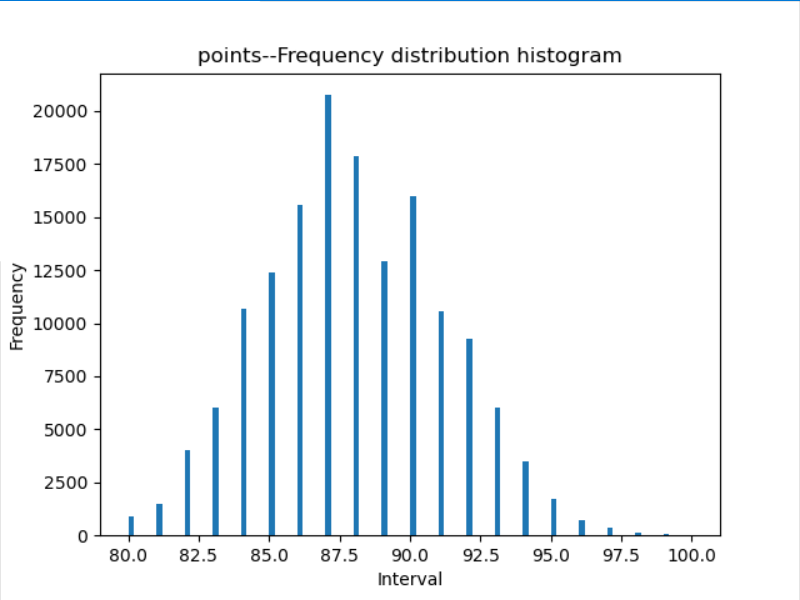
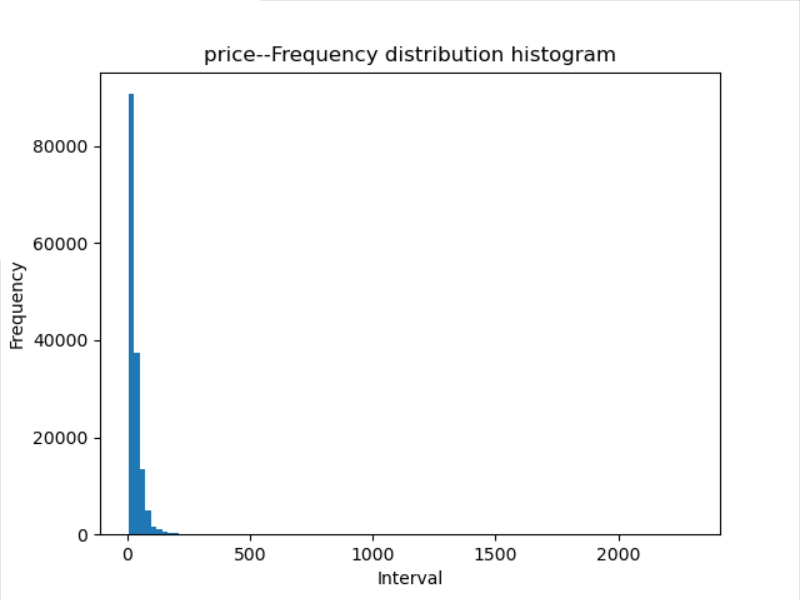
1. 将缺失部分剔除

上文中使用的过程就是剔除缺失值的方法，将空值皆去除掉。

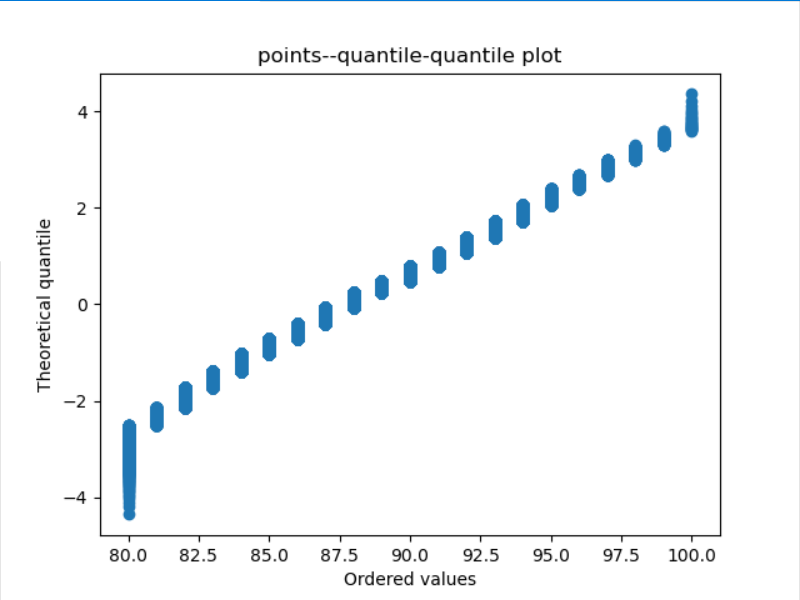
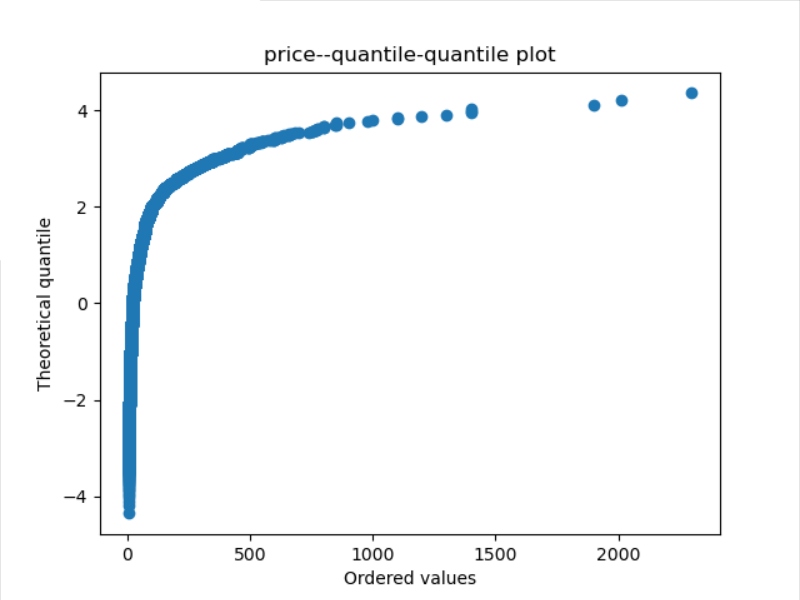
1. 用最高频率值来填补缺失值



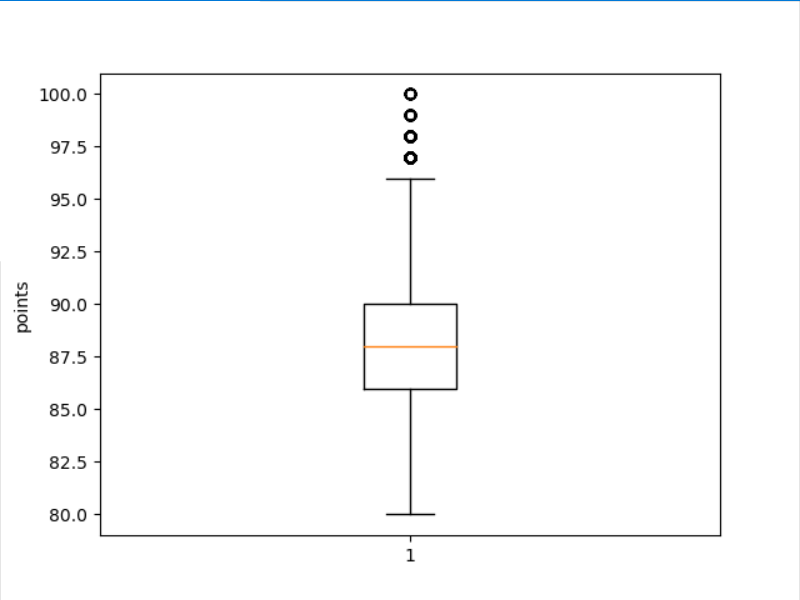
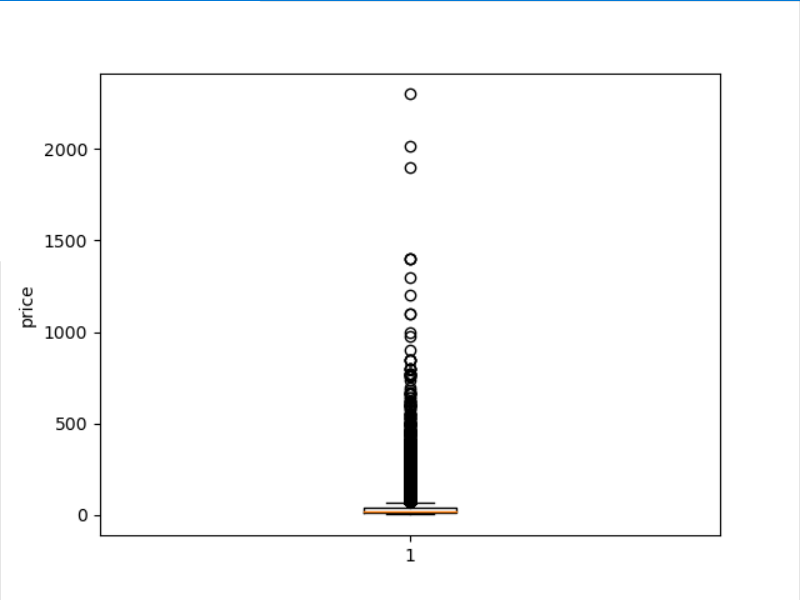
直方图：

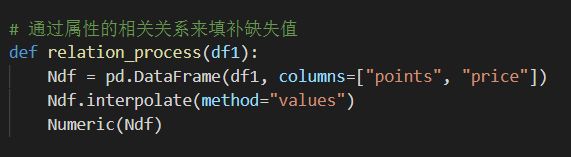
QQ图

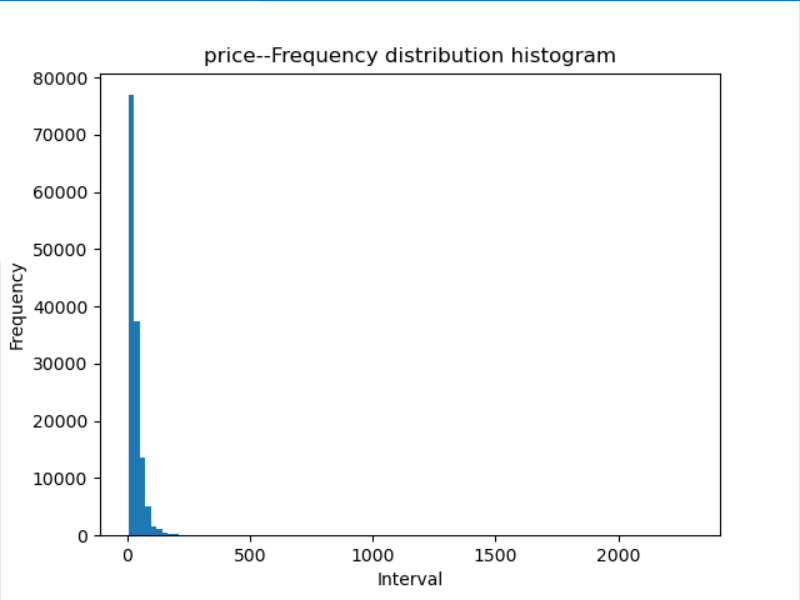
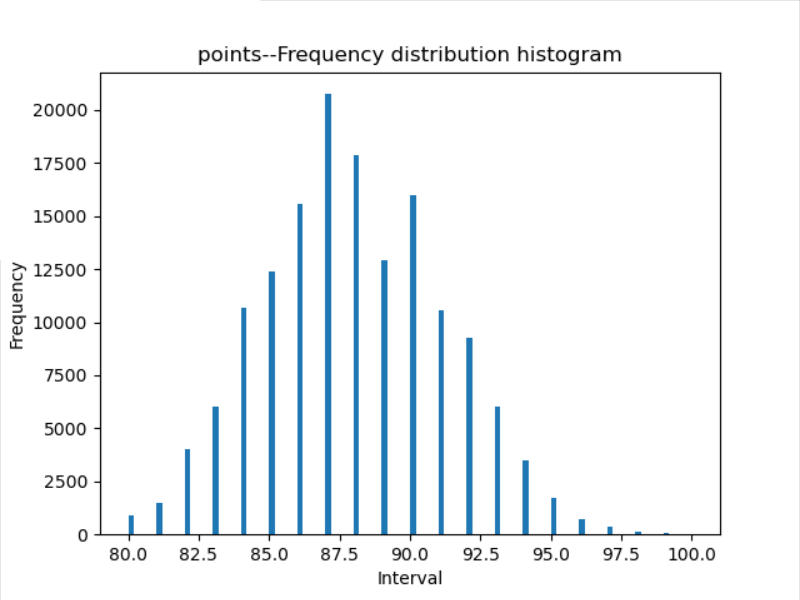
盒图：

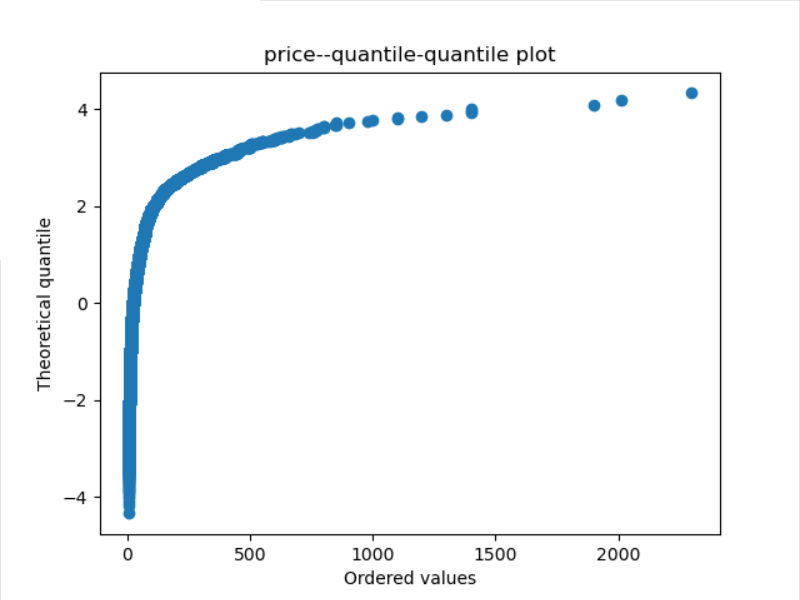
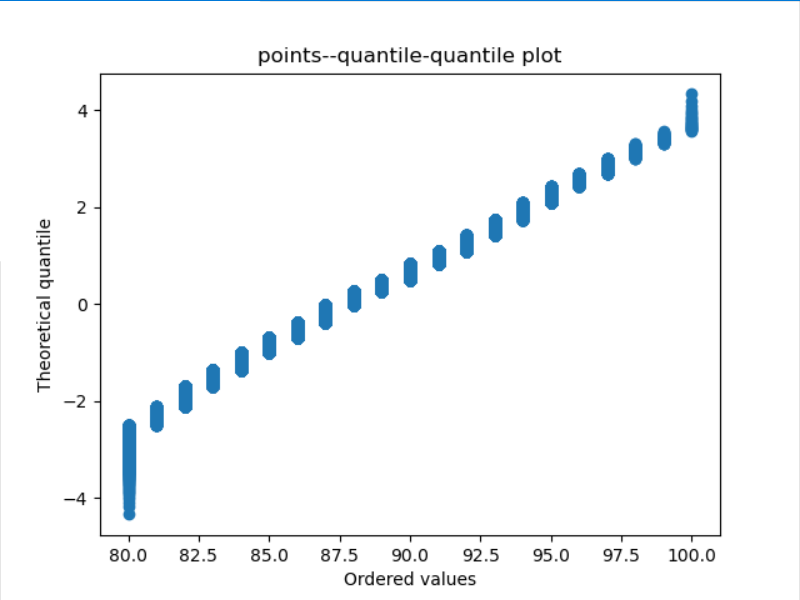
1. 通过属性的相关关系来填补缺失值



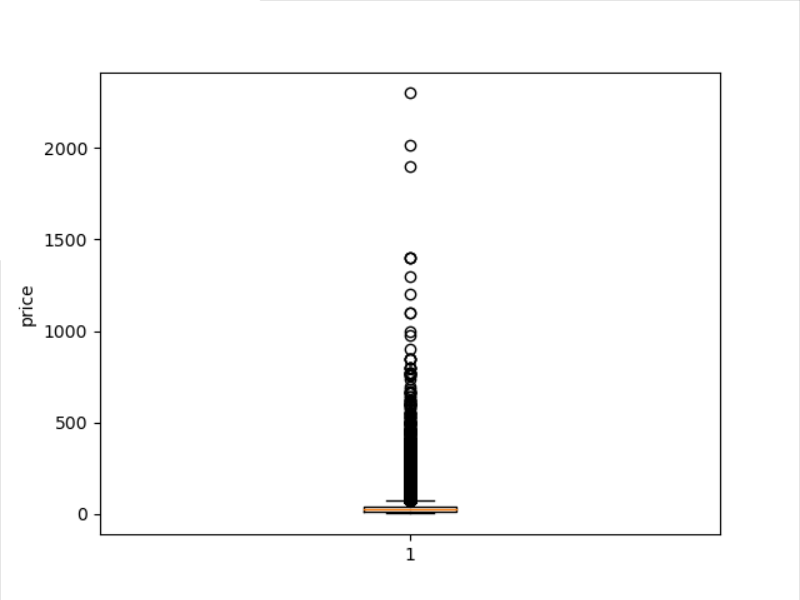
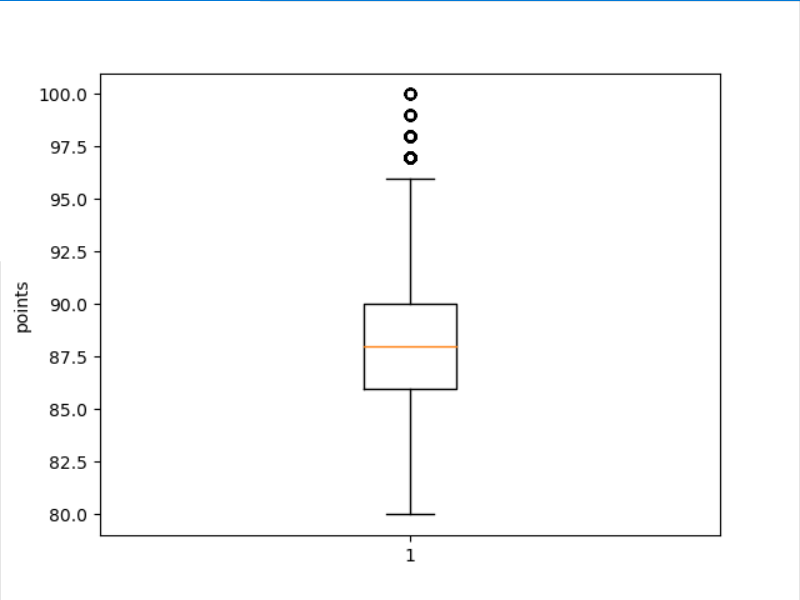
直方图：



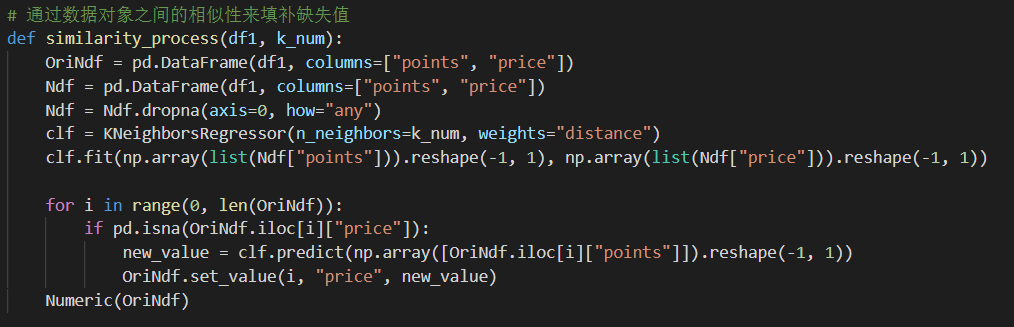
QQ图：



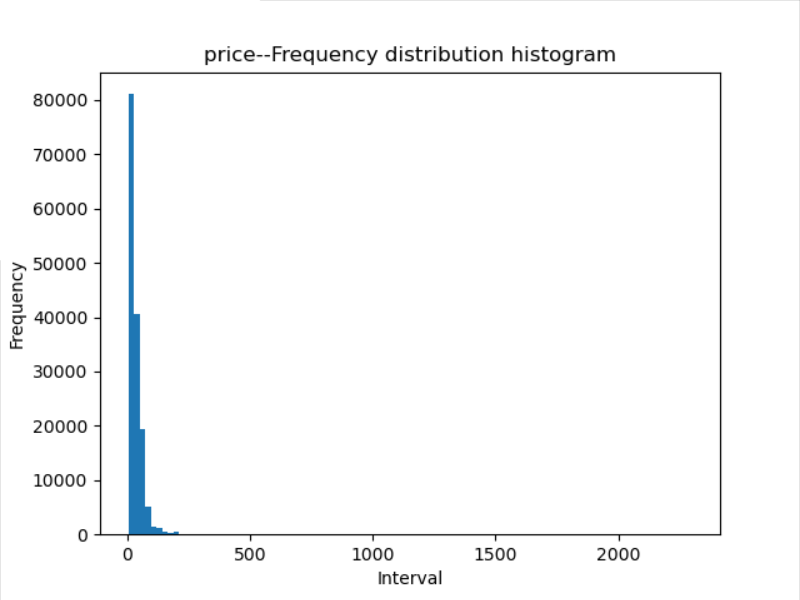
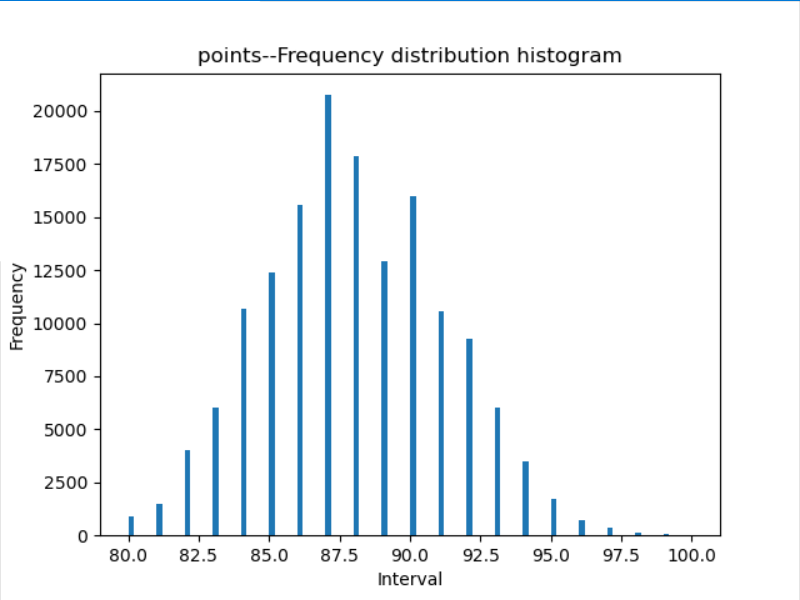
盒图：



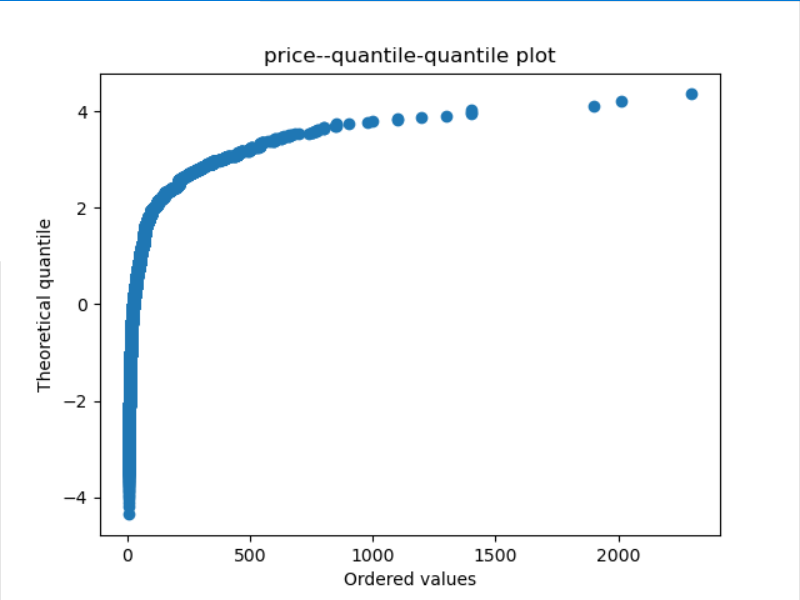
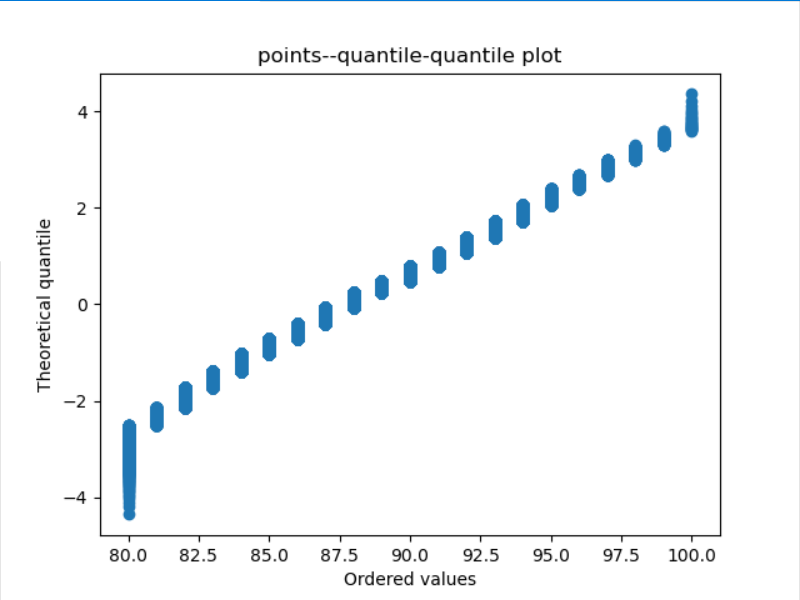
1. 通过数据对象之间的相似性来填补缺失值



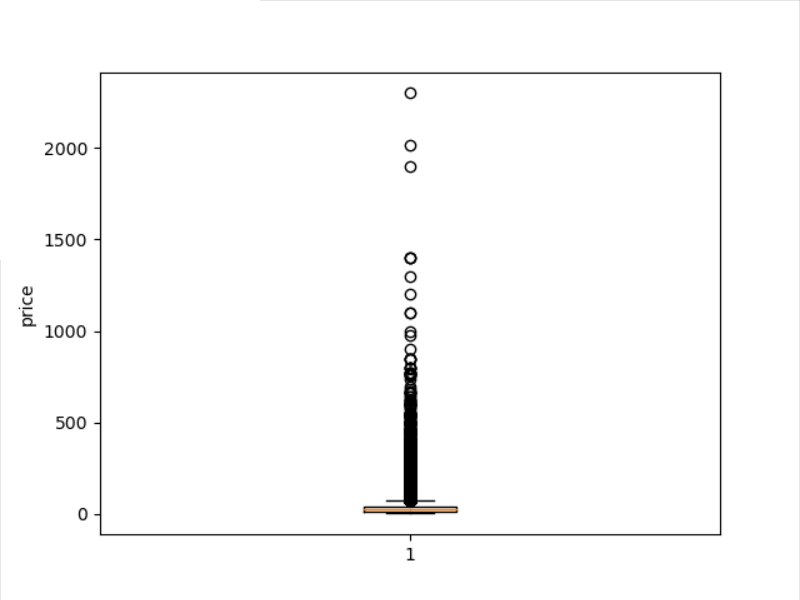
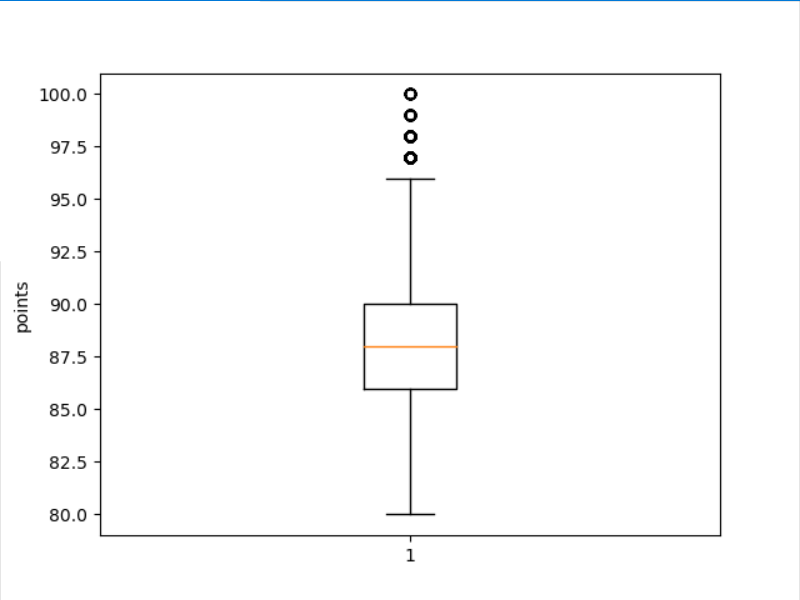
直方图：



QQ图：

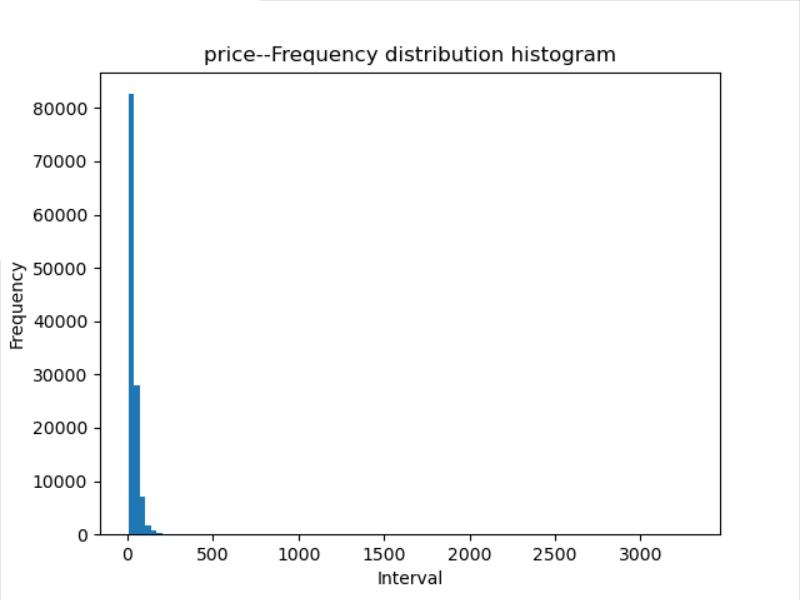
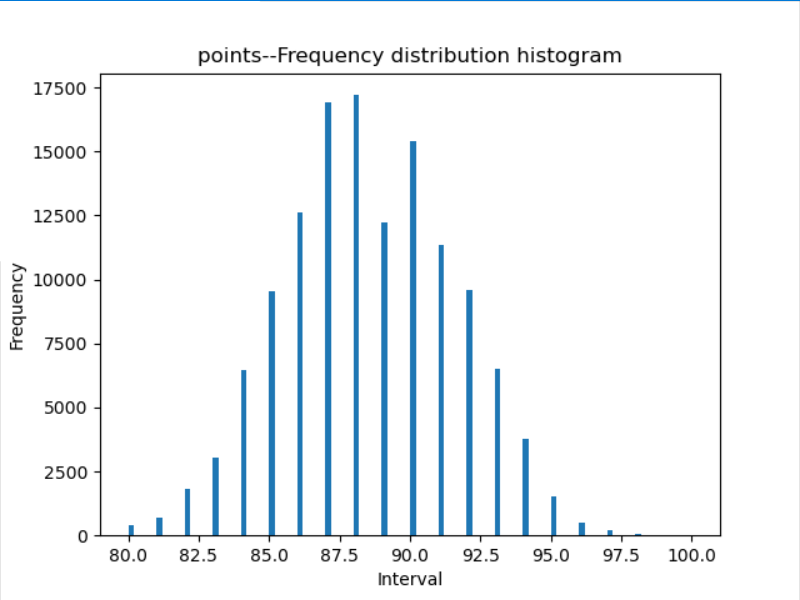


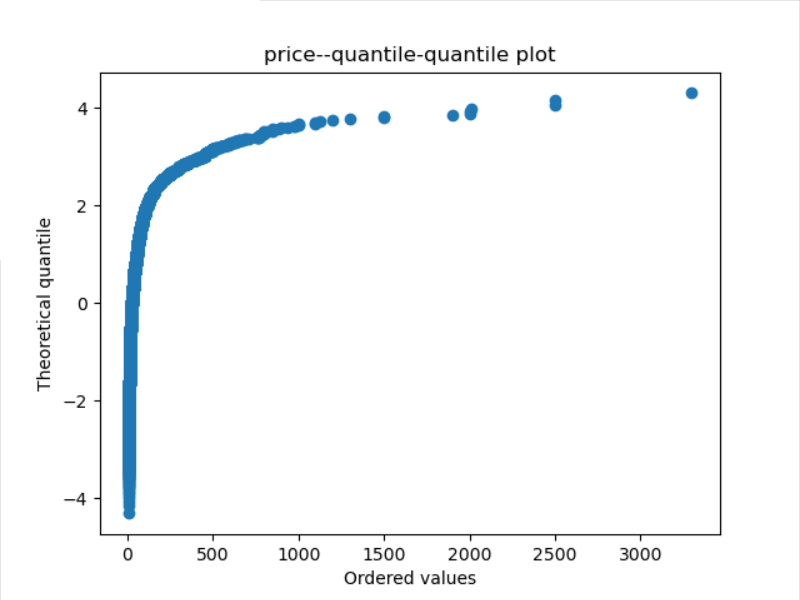
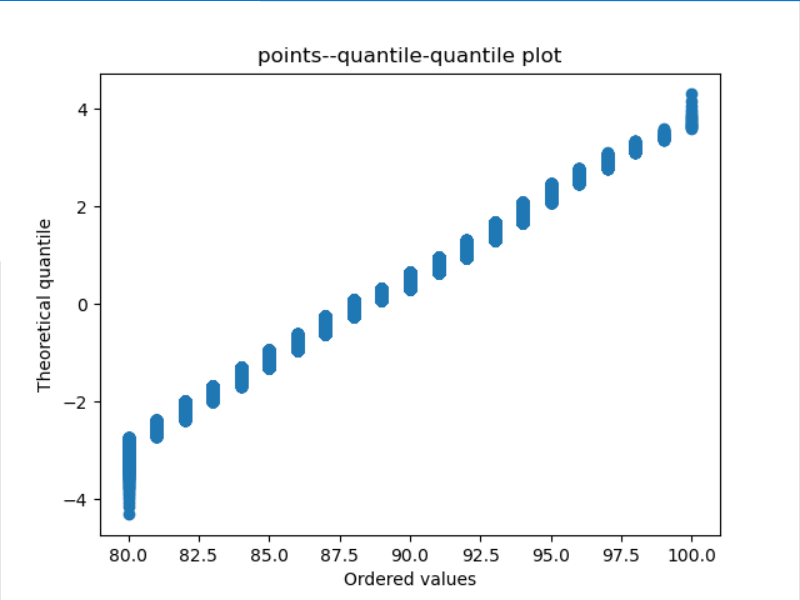
盒图：

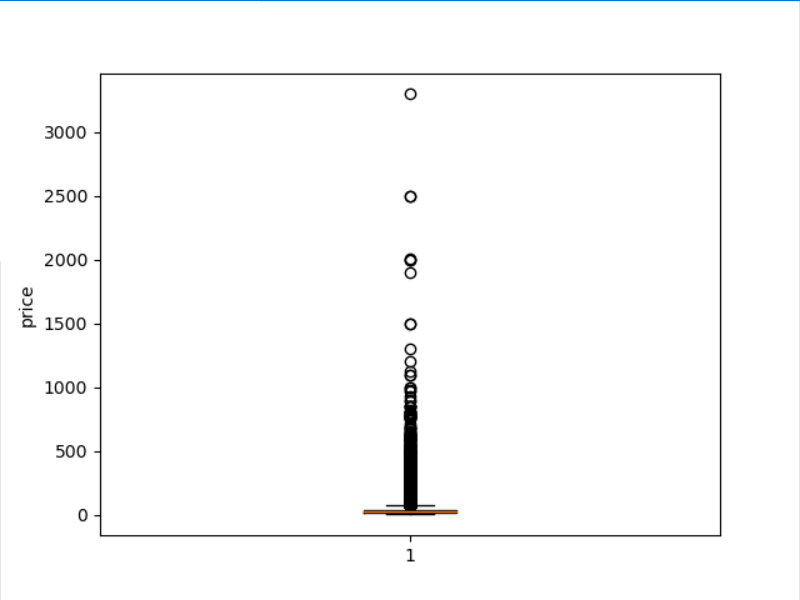
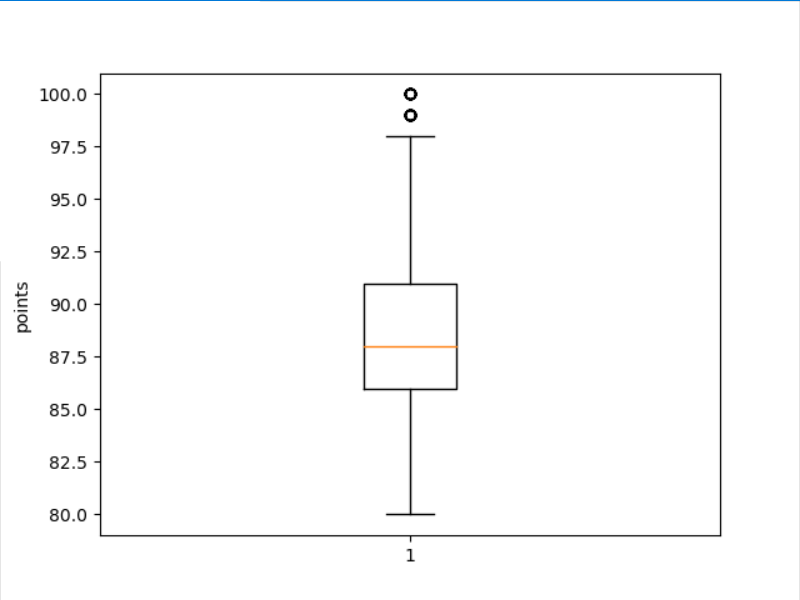


**文件2：winemag-data-130k-v2.csv**

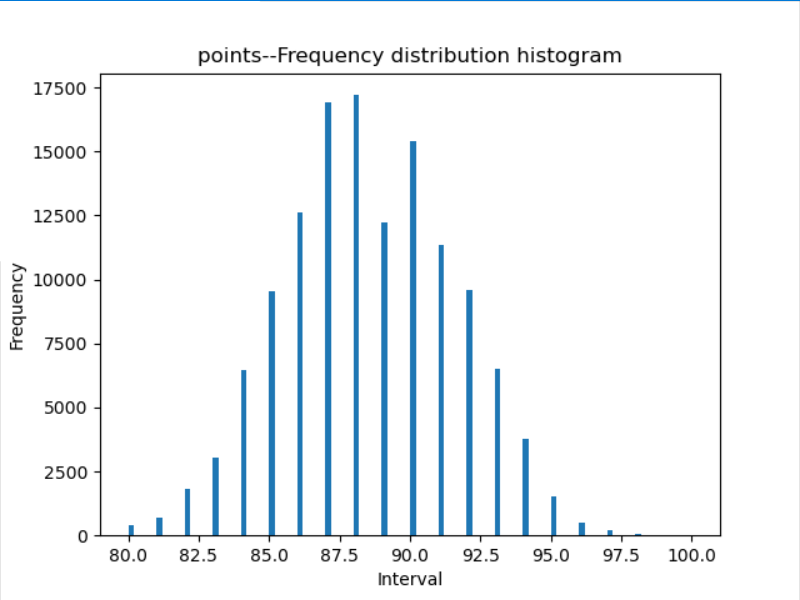
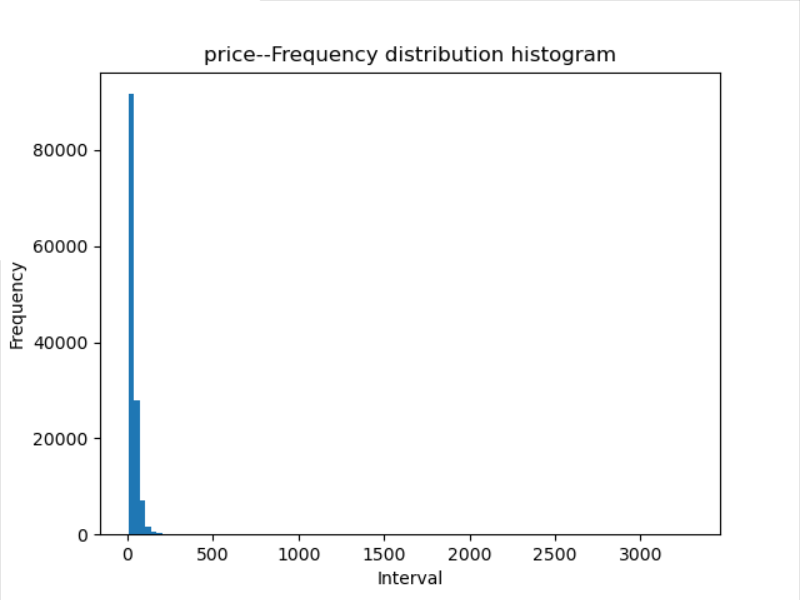
1. 原始数据（将缺失部分剔除）：

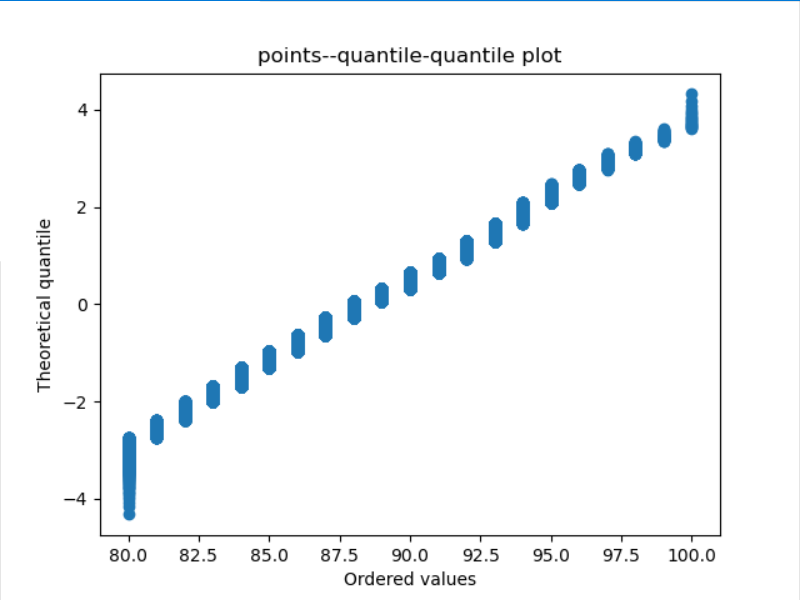
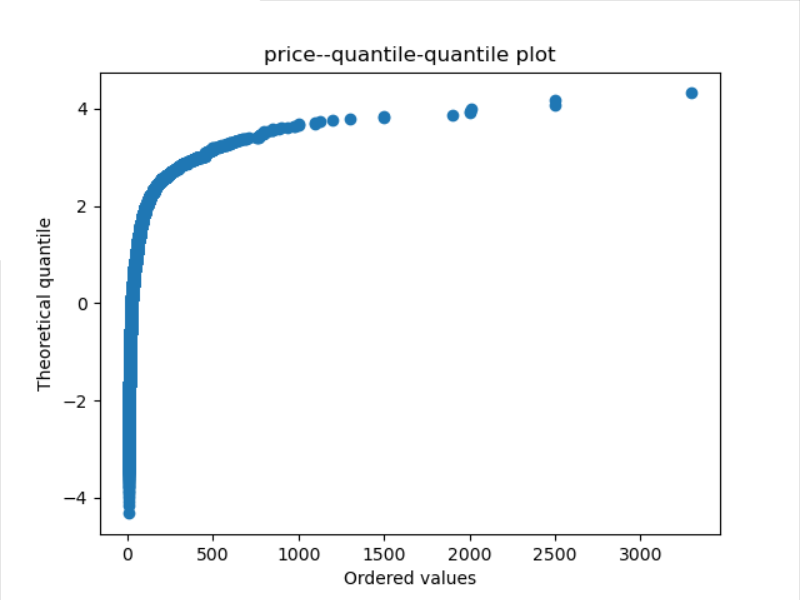


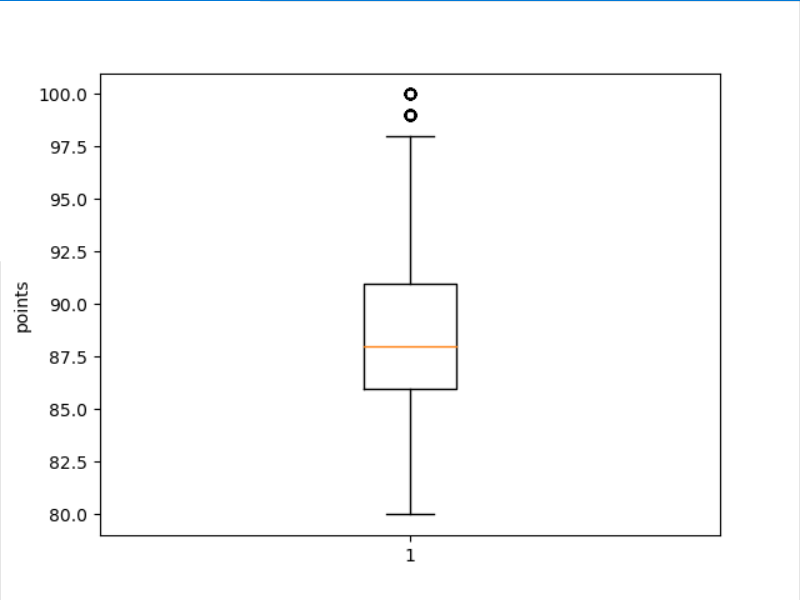
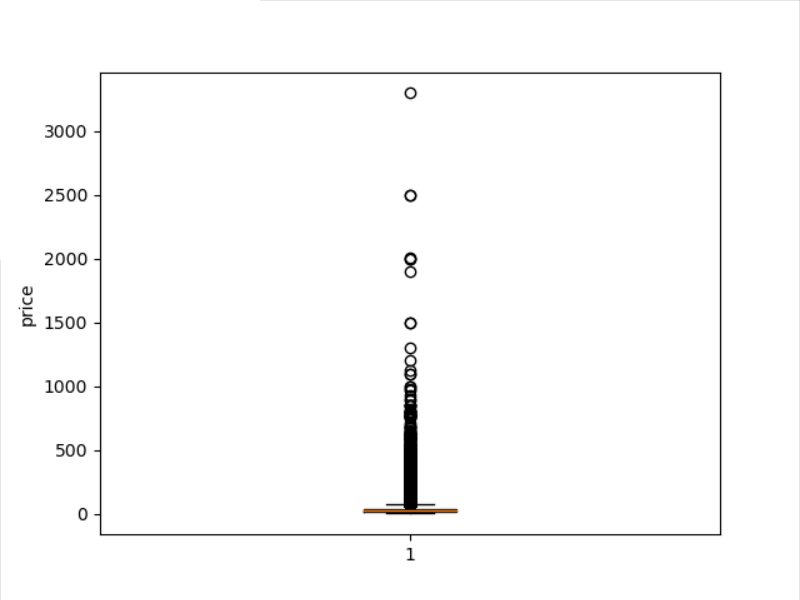




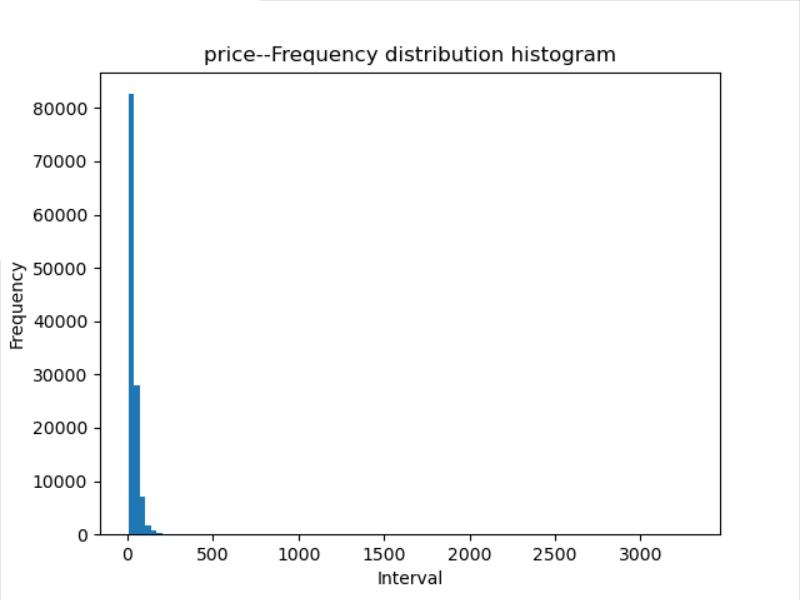
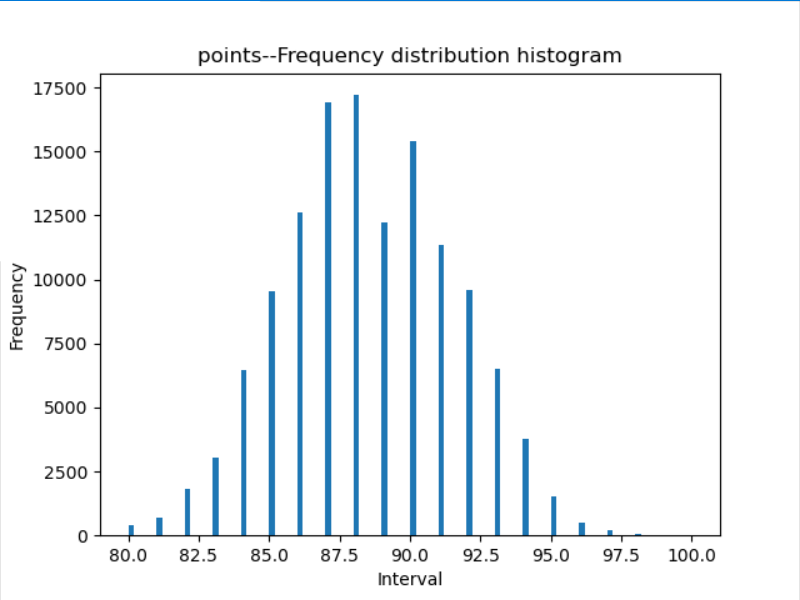
1. 用最高频率值来填补缺失值

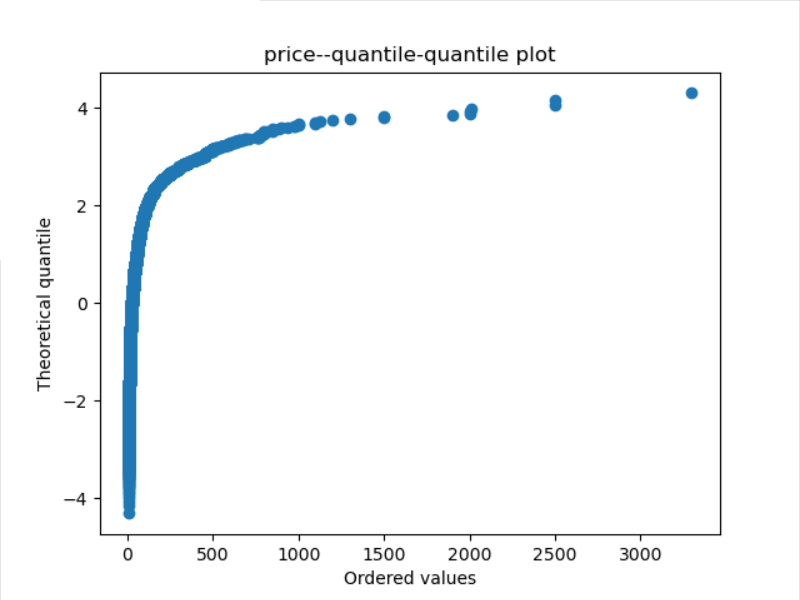
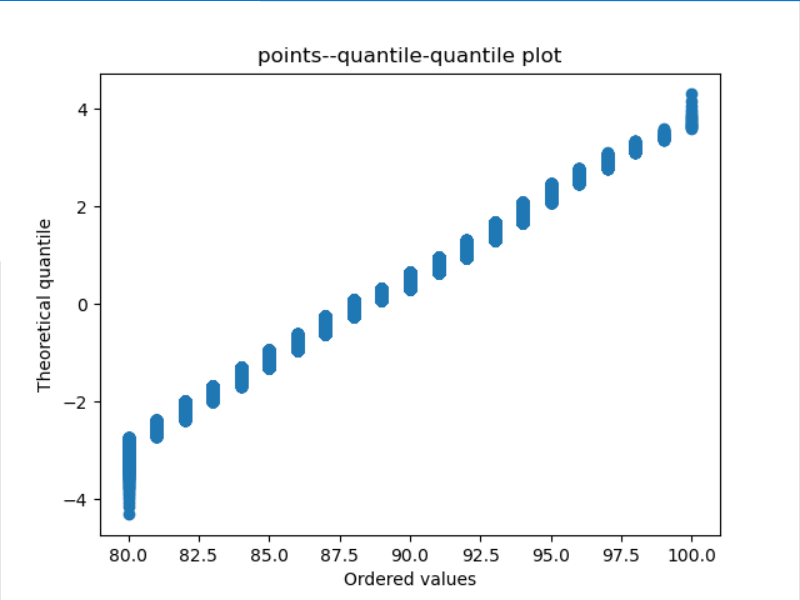
 

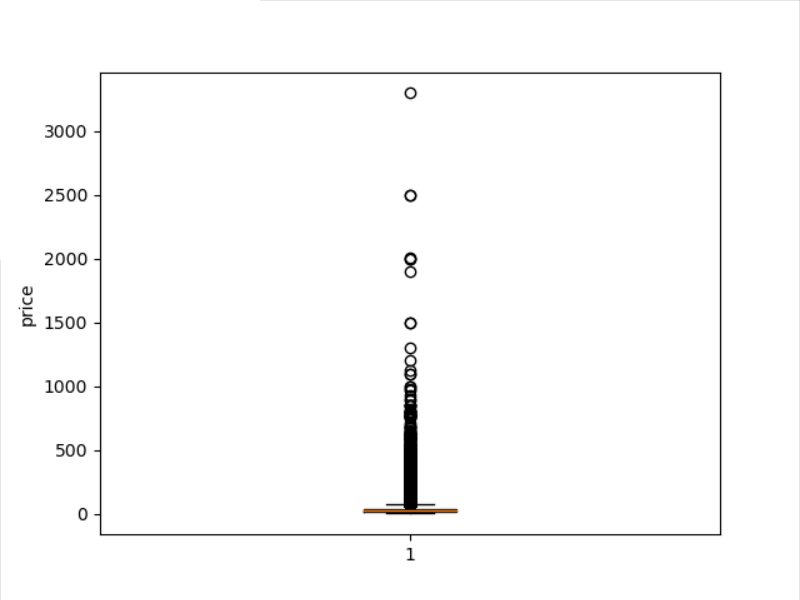
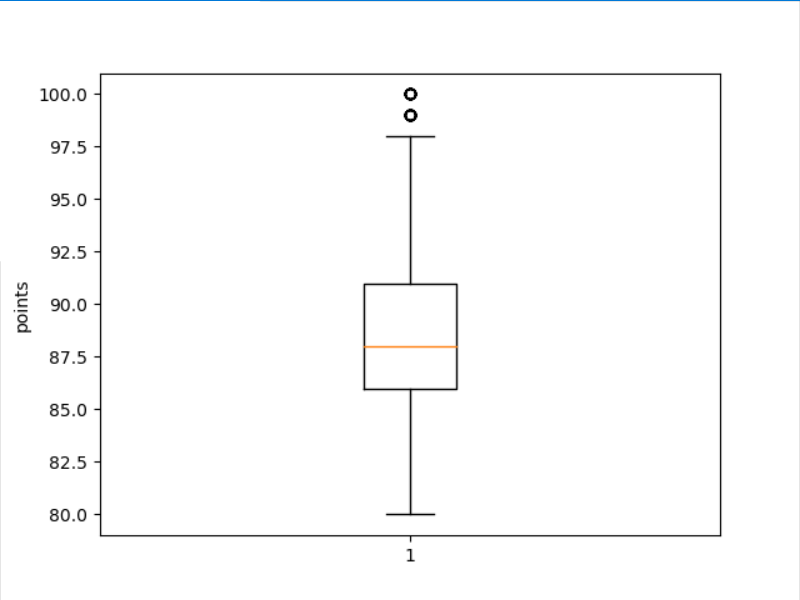
 

1. 通过属性的相关关系来填补缺失值：







1. 通过数据对象之间的相似性来填补缺失值：

