

**数据挖掘**

**互评作业一**

**题　　目：数据探索性分析与数据预处理**

**学 院： 计算机学院**

**专业名称： 计算机科学与技术**

**学 号： 3120191033**

**姓 名： 彭成**

**任课教师： 汤世平老师**

2020年5月2日星期六

# 对于wine-reviews的数据分析与预处理

## 数据集说明

该数据集包含两个数据子集，分别是2017年6月15日爬取的150,000条数据，和2017年11月22日爬取的拥有更丰富信息的130,000条数据。我们的数据分析与预处理将会在这两个数据子集上进行。

数据集路径：

./data/wine-reviews/winemag-data\_first150k.csv

./data/wine-reviews/winemag-data-130k-v2.csv

## winemag-data\_first150k数据子集

我们可以从数据表中得知，每条数据包含{ 'country', 'description', 'designation', 'points', 'price', 'province', 'region\_1', 'region\_2', 'variety', 'winery'}共十个属性。

属性{'description'}是每条数据中具体的评论，不具备重复的性质，我们在这一阶段不进行分析和处理。

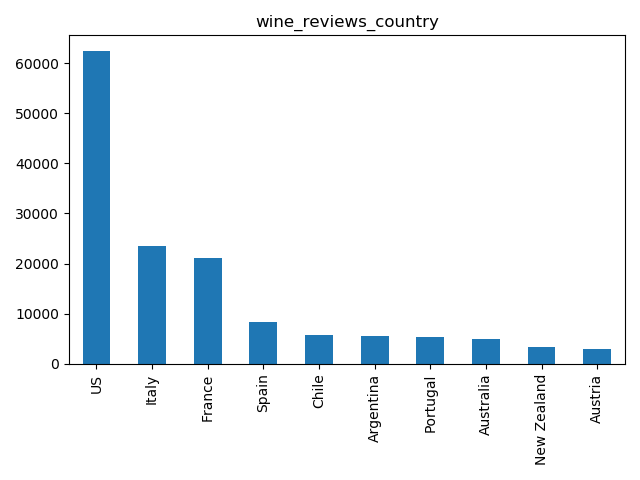
属性{'country', 'designation', 'province', 'region\_1', 'region\_2', 'variety', 'winery'}，分别表示国家名、葡萄园名、省名、产区1名、产区2名、葡萄类型、酿酒厂名。这些都是标称属性，我将对这些属性给出每个可能聚会的频数，用直方图表示数据的分布，每个属性显示出现最多的前十个。

属性{ 'points', 'price'}，分别表示每条数据对葡萄酒的评分以及这瓶葡萄酒的成本。这些是数值属性，我将对这些属性给出5数概括及缺失值的个数，用直方图表示数据的分布，用盒图检查离群点。

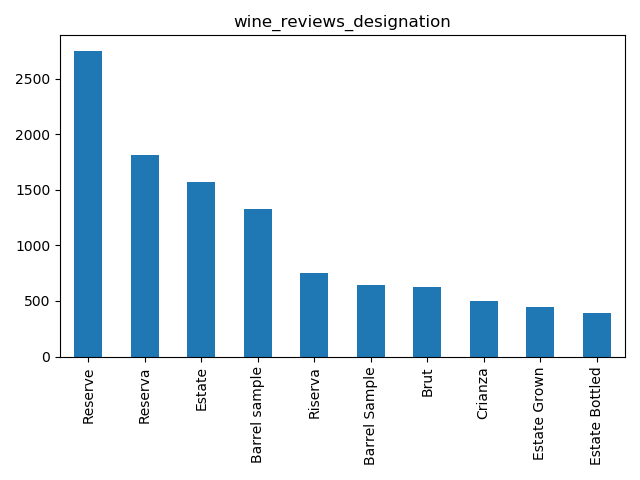
然后，对数据集中缺失数据，分析缺失的原因，使用四种方法对缺失值进行处理后，使用直方图对比新旧数据集。

### 数据可视化和摘要

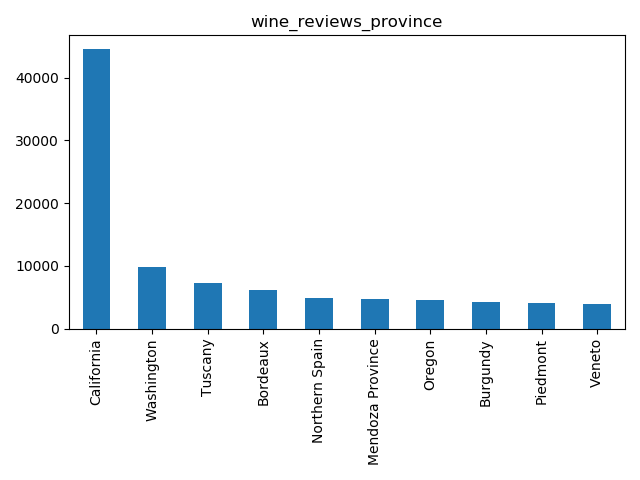
对标称属性'country'有：



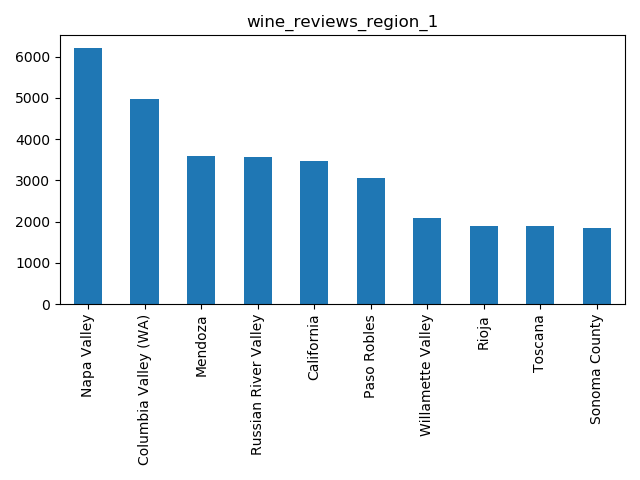
对标称属性'designation'有：



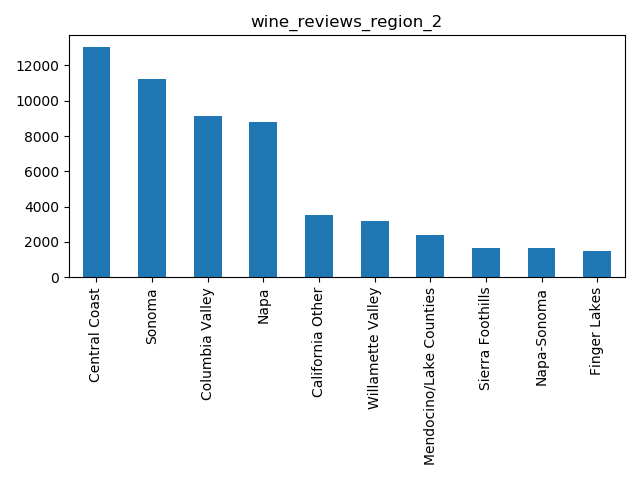
对标称属性'province'有：



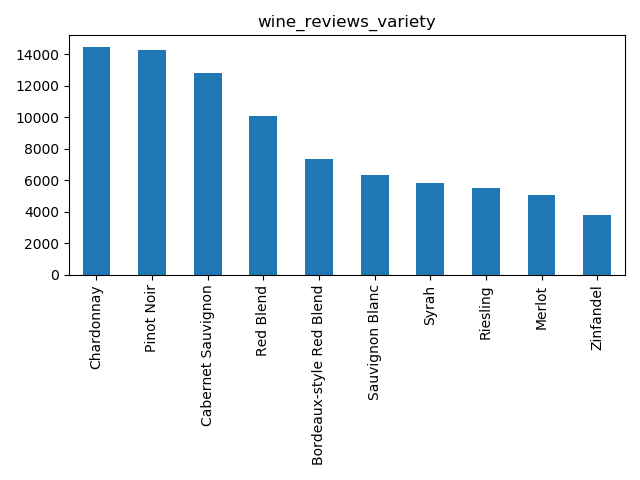
对标称属性'region\_1'有：



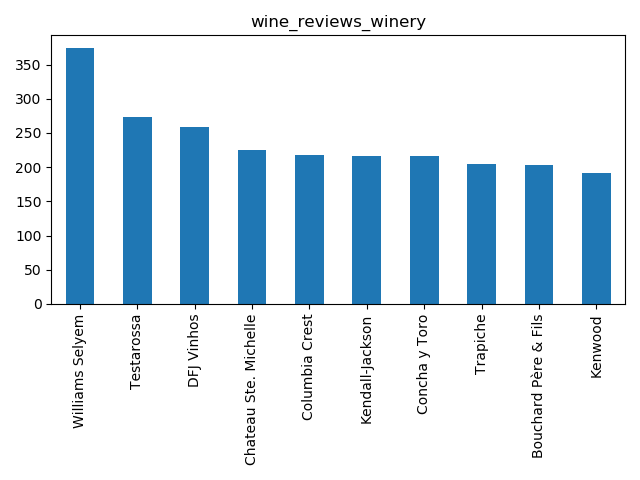
对标称属性'region\_2'有：



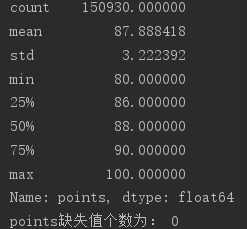
对标称属性'variety'有：

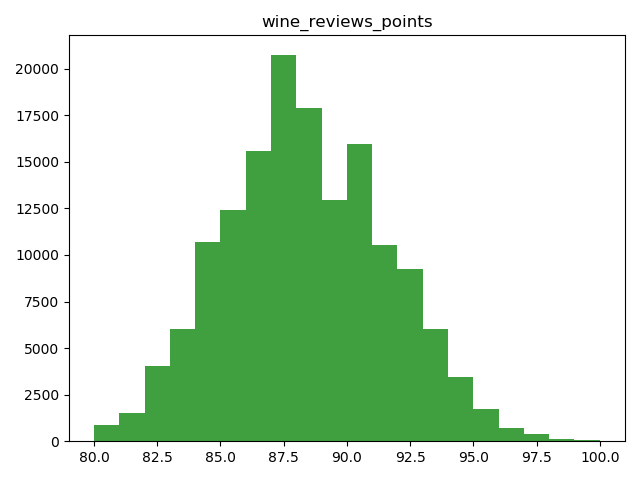


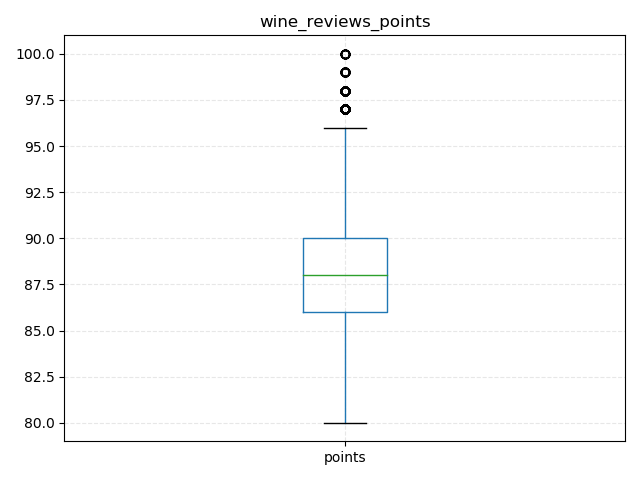
对标称属性'winery'有：



对数值属性'points'有：

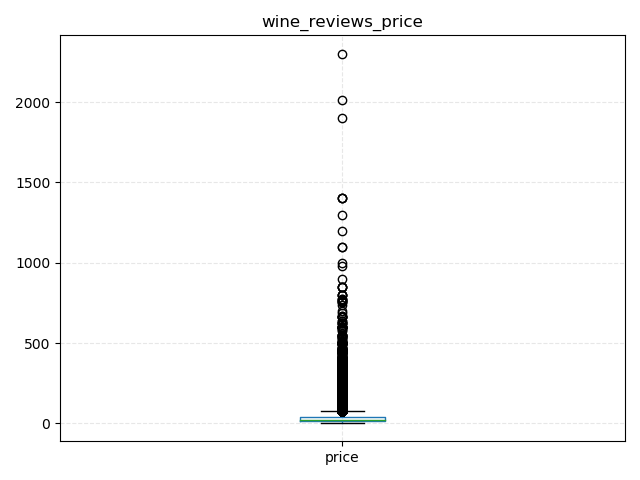
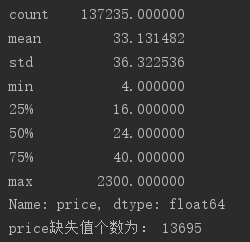


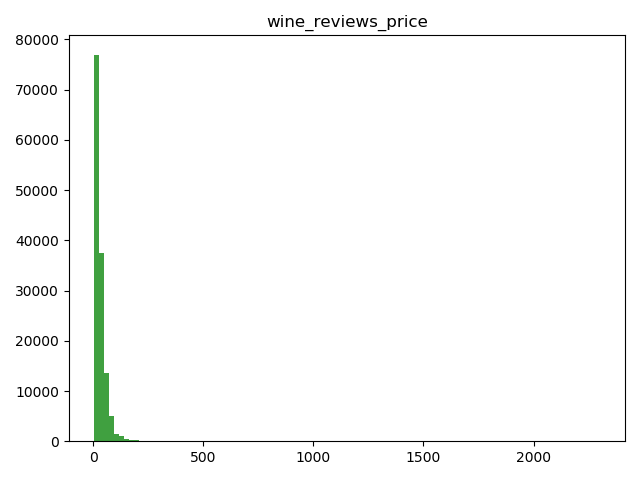




可以从第一个图看到五数概括分别为：80,86,88,90,100。缺失值的个数为0个。在盒图中，可以看到离群点的存在。

对数值属性'price'有：

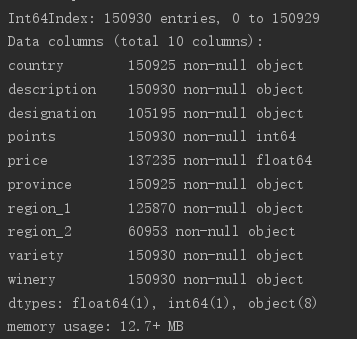




可以从第一个图看到五数概括分别为：4,16,24,40,2300。缺失值的个数为13695个。在盒图中，可以看到离群点的存在。

### 数据缺失的处理

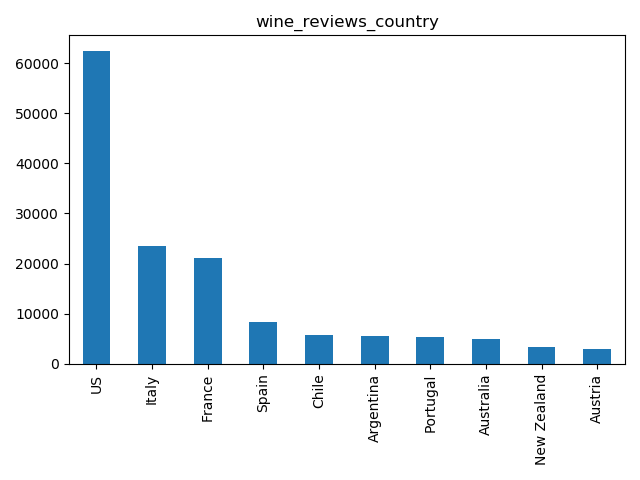
#### 分析缺失的原因



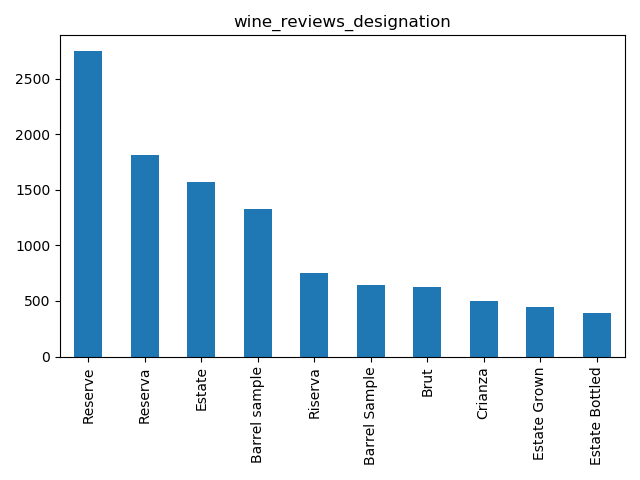
可以用.info()的方法得知每种属性的具体情况，我们可以看到，在我们要分析的属性中，points、variety、winery都是没有缺失值的，我们要进行填补的是country、designation、price、province、region\_1、region\_2这几个属性。缺失的原因很大一部分是可以不填写或者忘了填写了。

#### 将缺失部分剔除

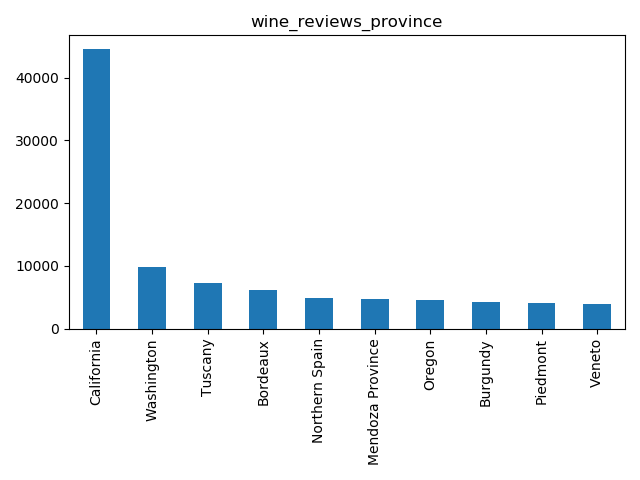
对标称属性'country'有：



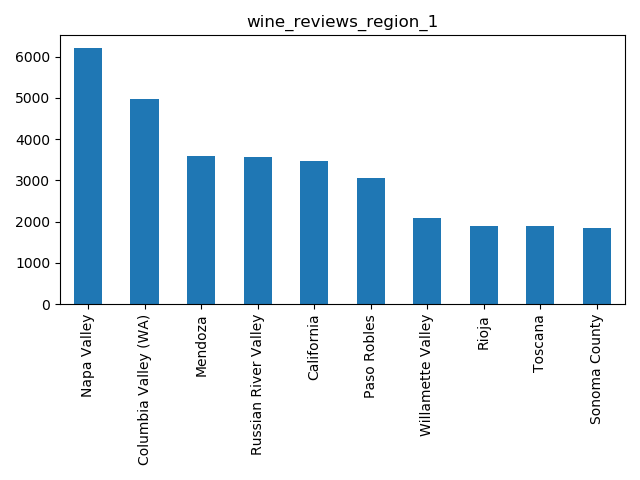
对标称属性'designation'有：



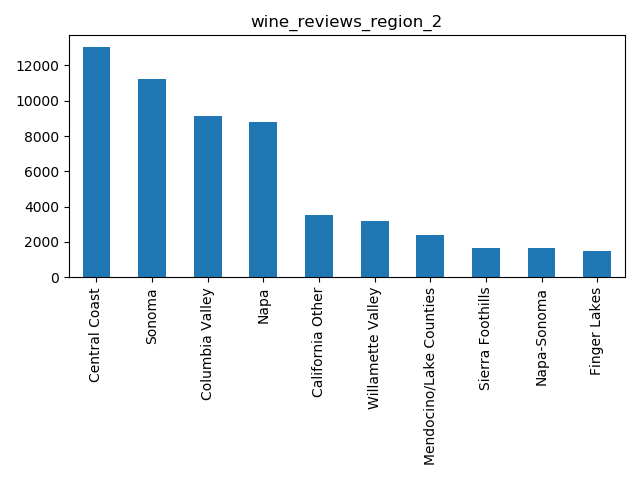
对标称属性'province'有：



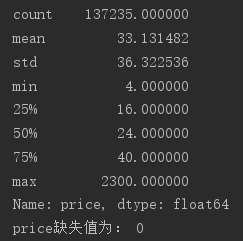
对标称属性'region\_1'有：

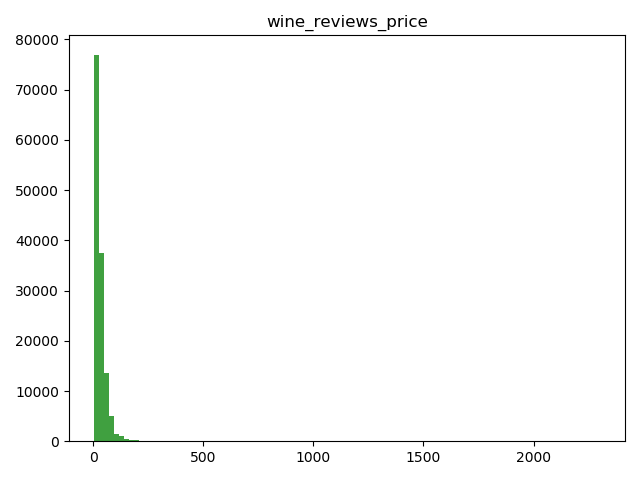


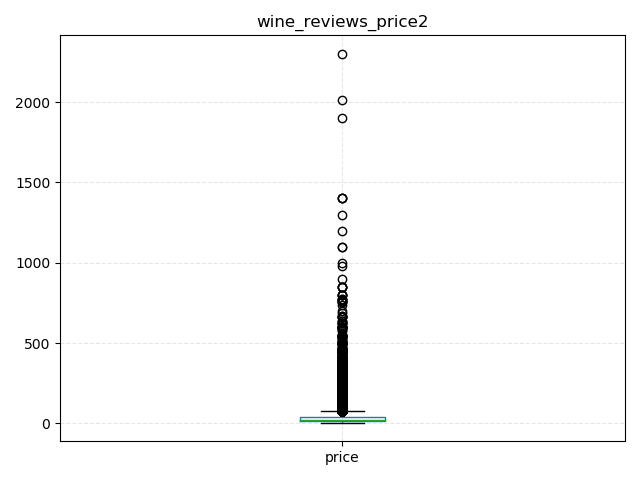
对标称属性'region\_2'有：



对数值属性'price'有：



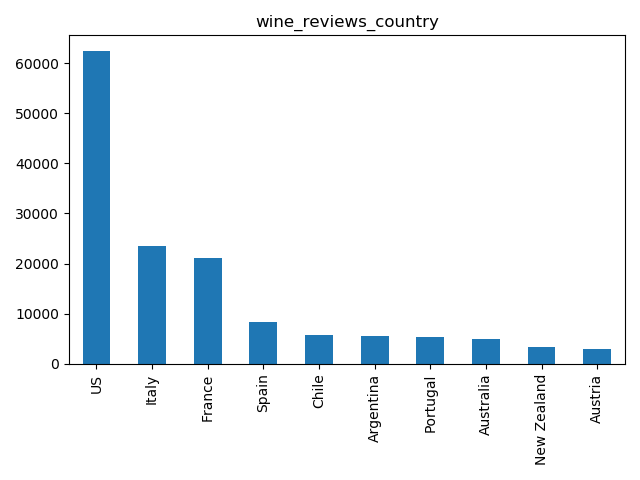




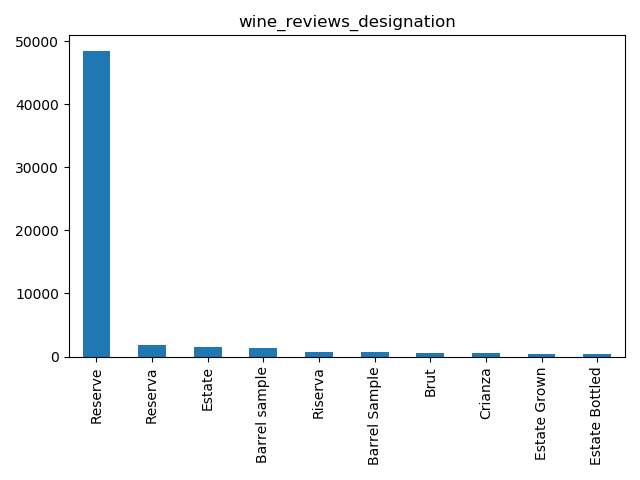
可以从第一个图看到五数概括分别为：80,86,88,90,100。缺失值的个数为0个。在盒图中，可以看到离群点的存在。

#### 用最高频率值来填补缺失值

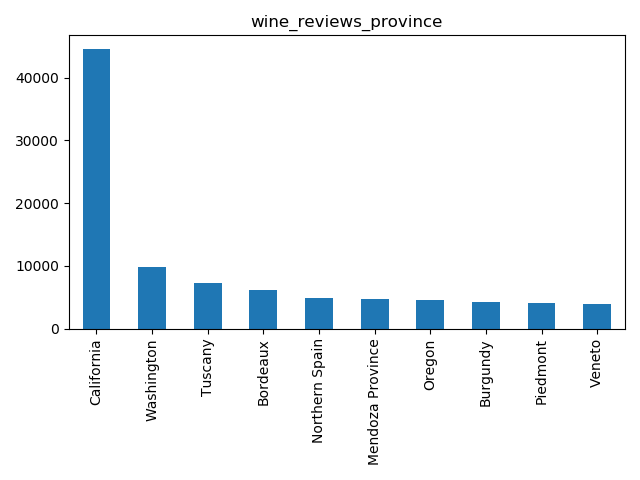
对标称属性'country'有：



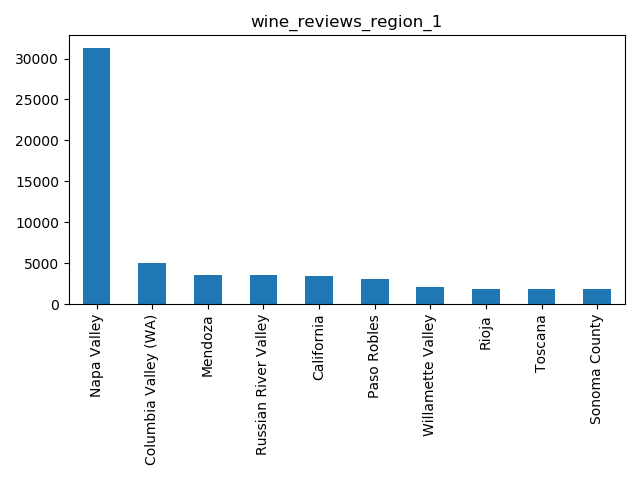
对标称属性'designation'有：



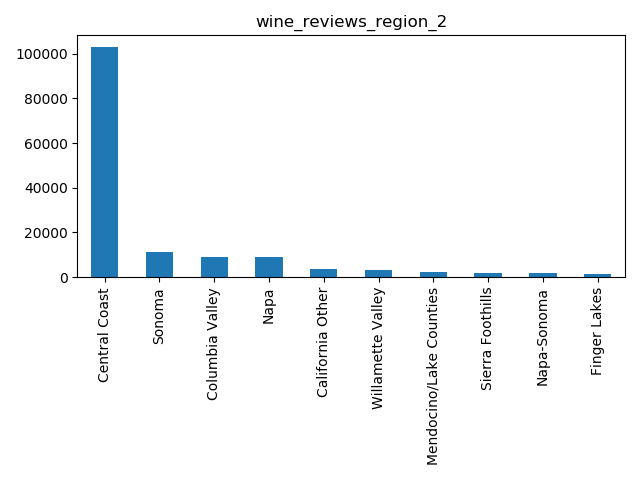
对标称属性'province'有：



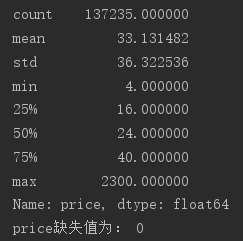
对标称属性'region\_1'有：



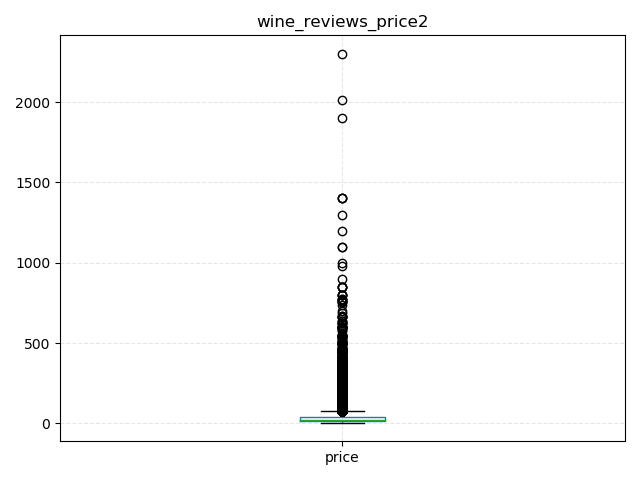
对标称属性'region\_2'有：



对数值属性'price'有：





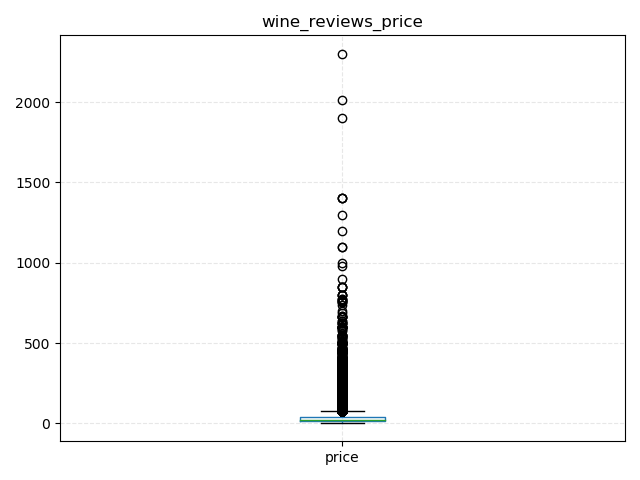
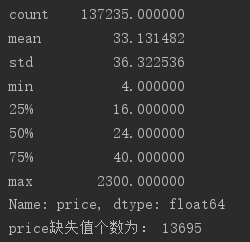


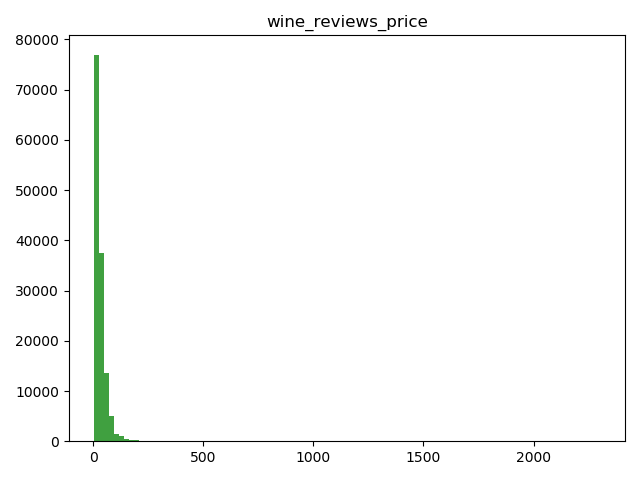
可以从第一个图看到五数概括分别为：80,86,88,90,100。缺失值的个数为0个。在盒图中，可以看到离群点的存在。

#### 通过属性的相关关系来填补缺失值

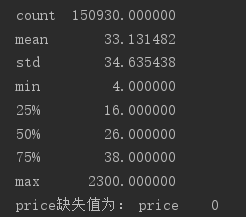
使用SimpleImputer，用来填充数据里面的缺失值。可以从盒图与直方图看到前后的对比。

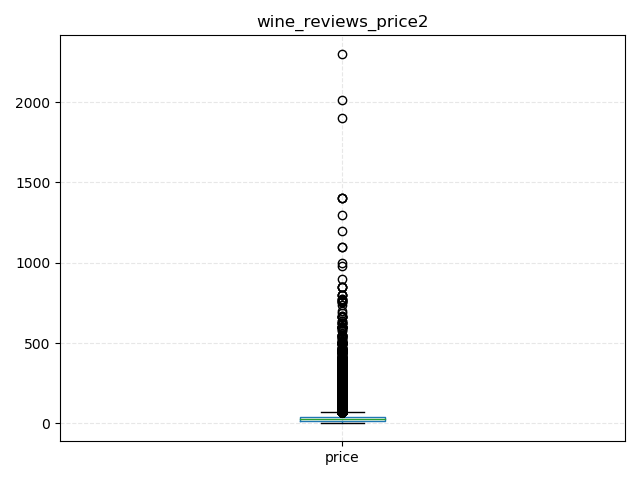
填补缺失值之前：

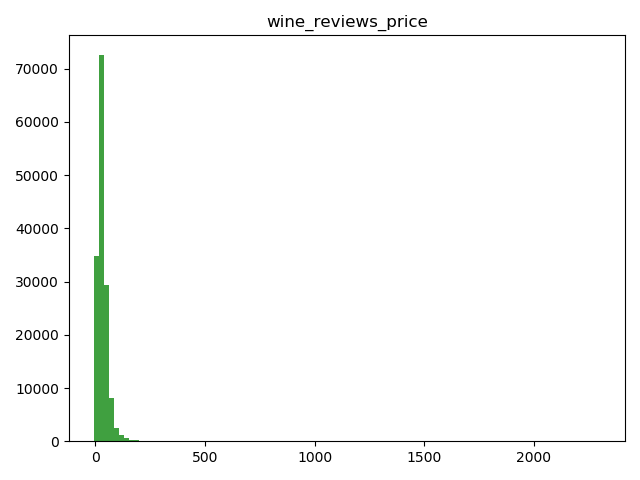




填补缺失值之后：





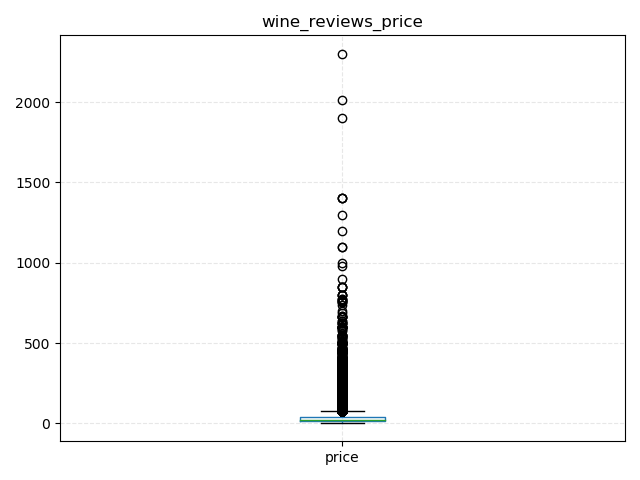
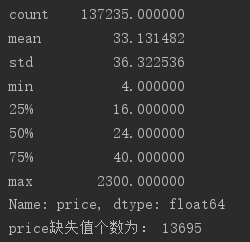


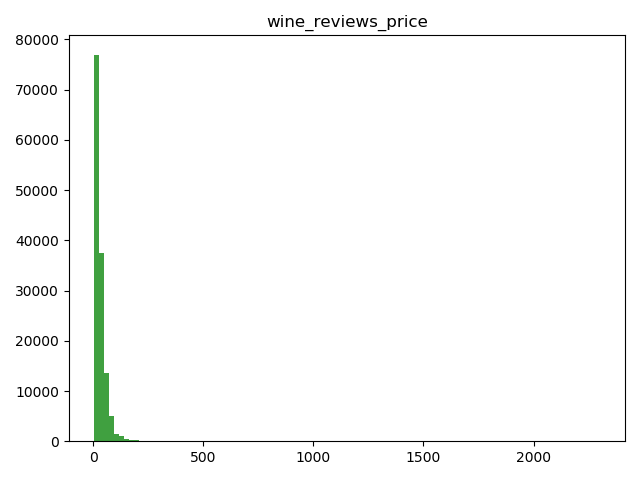
#### 通过数据对象之间的相似性来填补缺失值

使用线性回归分类器LinearRegression，拟合points属性和price属性之间的相关关系，用训练好的分类器，基于已知的points值预测price的值。

可以从盒图与直方图中看到前后的对比。

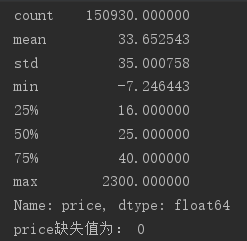
填补缺失值之前：

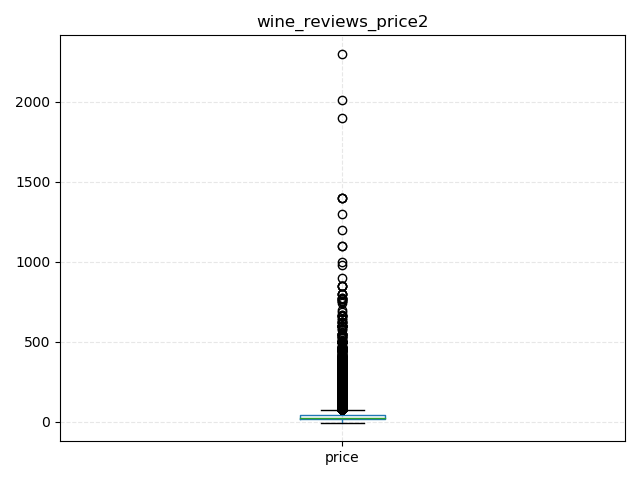


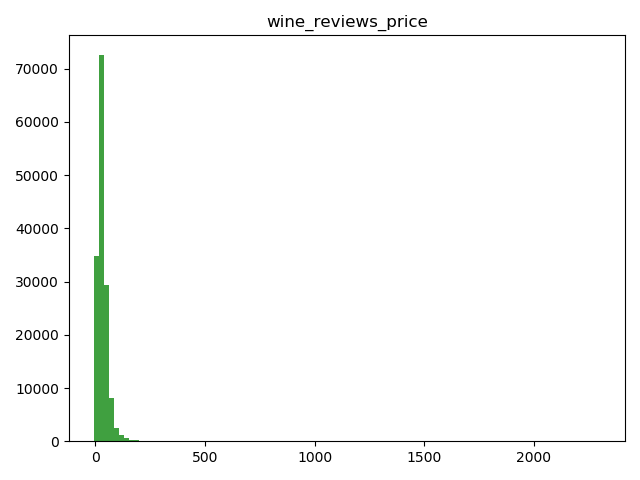


可以从第一个图看到五数概括分别为：4,16,24,40,2300。缺失值的个数为13695个。在盒图中，可以看到离群点的存在。

填补缺失值之后：







## winemag-data-130k-v2数据子集

我们可以从数据表中得知，每条数据包含{ 'country', 'description', 'designation', 'points', 'price', 'province', 'region\_1', 'region\_2', 'taster\_name', 'taster\_twitter\_handle', 'title', 'variety', 'winery'}共13个属性。

属性{'description', 'title'}是每条数据中具体的评论和标题，不具备重复的性质，我们在这一阶段不进行分析和处理。

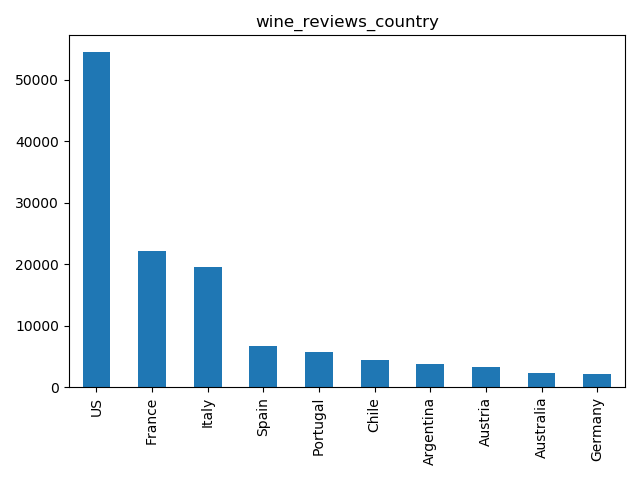
属性{'country', 'designation', 'province', 'region\_1', 'region\_2', 'taster\_name', 'taster\_twitter\_handle', 'variety', 'winery'}，分别表示国家名、葡萄园名、省名、产区1名、产区2名、品尝师名、品尝师Twitter名、葡萄类型、酿酒厂名。这些都是标称属性，我将对这些属性给出每个可能聚会的频数，用直方图表示数据的分布，每个属性显示出现最多的前十个。其中品尝师名和品尝师Twitter名一一对应，故我们只对品尝师名进行分析。

属性{ 'points', 'price'}，分别表示每条数据对葡萄酒的评分以及这瓶葡萄酒的成本。这些是数值属性，我将对这些属性给出5数概括及缺失值的个数，用直方图表示数据的分布，用盒图检查离群点。

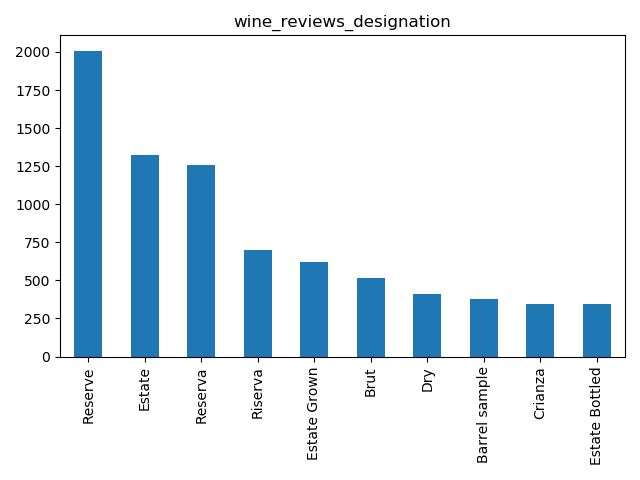
然后，对数据集中缺失数据，分析缺失的原因，使用四种方法对缺失值进行处理后，使用直方图对比新旧数据集。

### 数据可视化和摘要

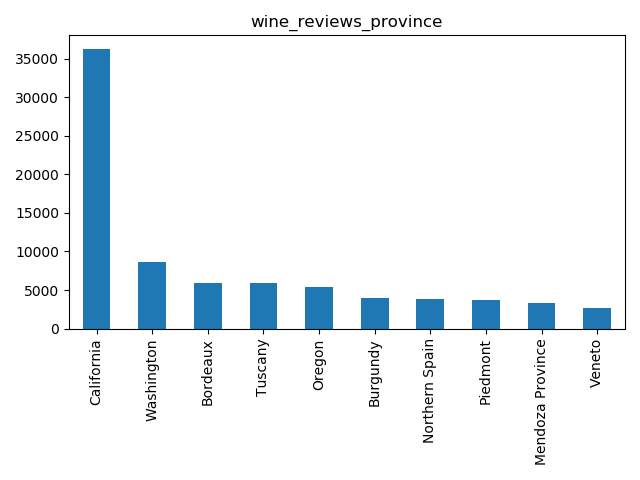
对标称属性'country'有：



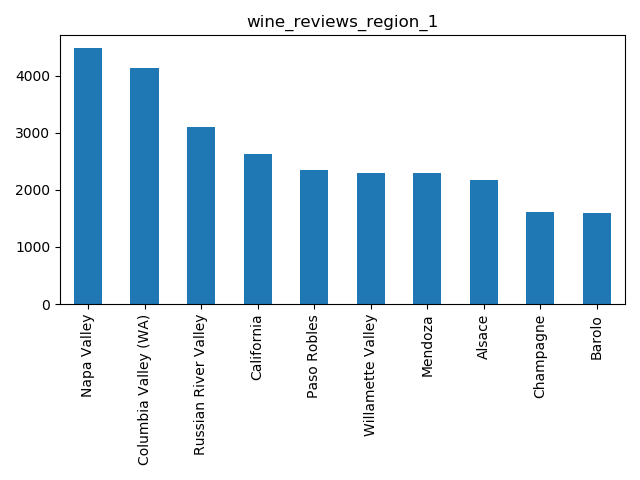
对标称属性'designation'有：



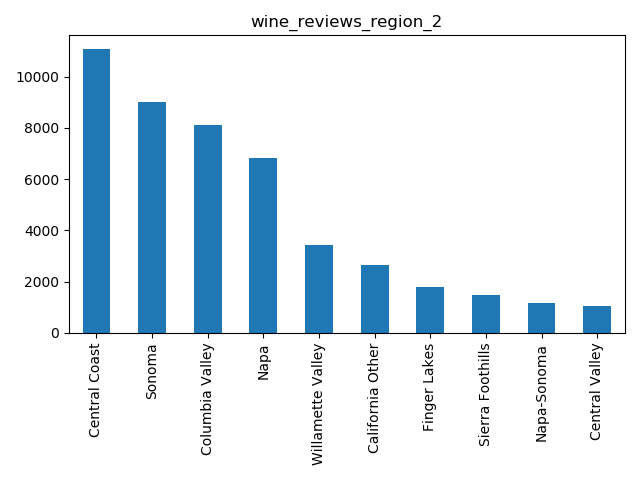
对标称属性'province'有：



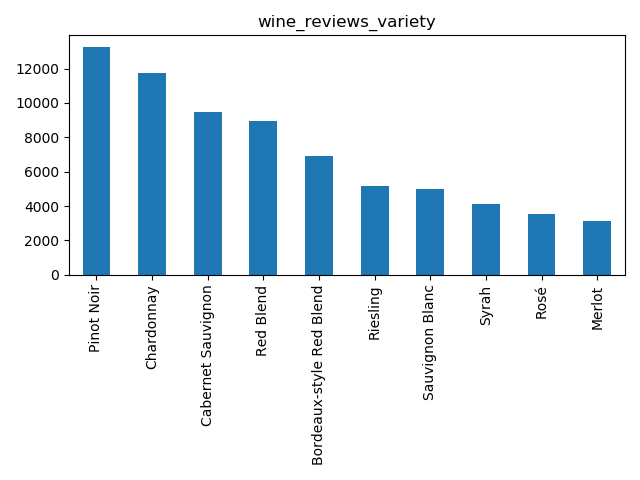
对标称属性'region\_1'有：



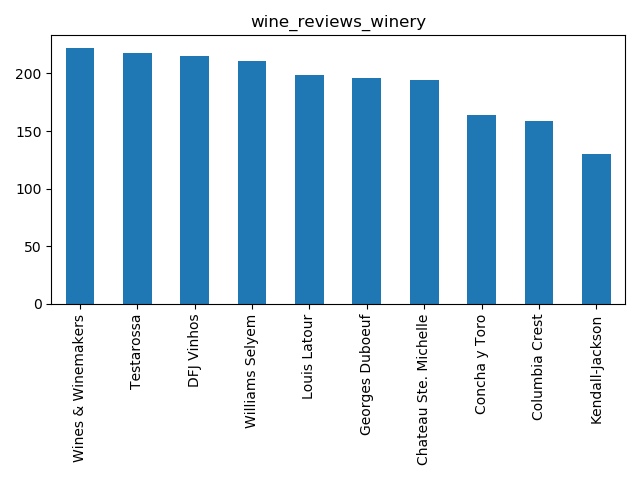
对标称属性'region\_2'有：



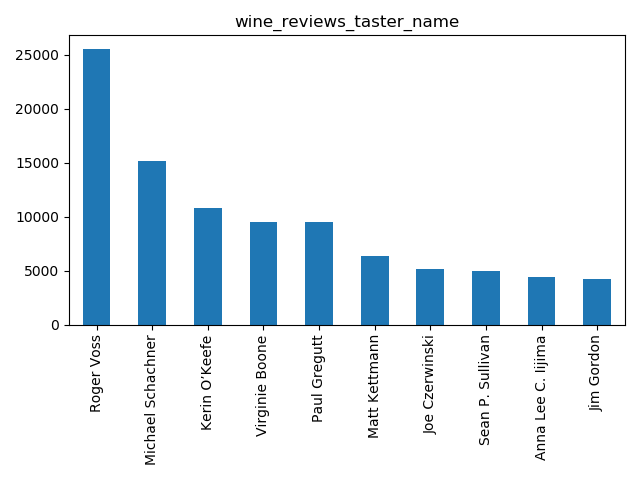
对标称属性'variety'有：



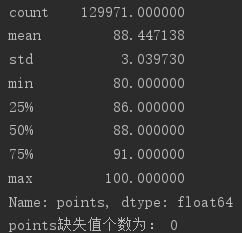
对标称属性'winery'有：

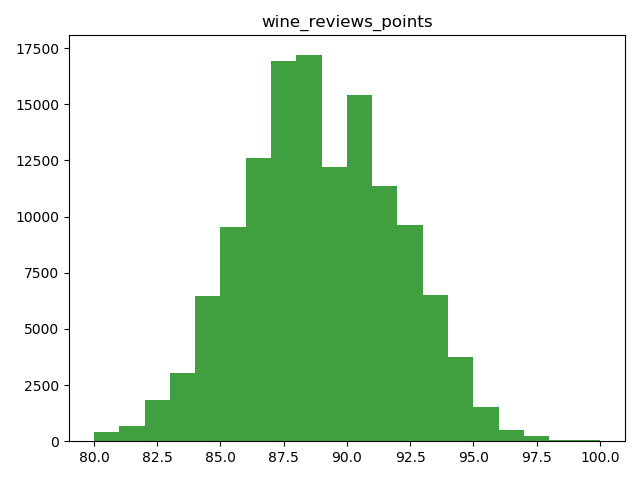


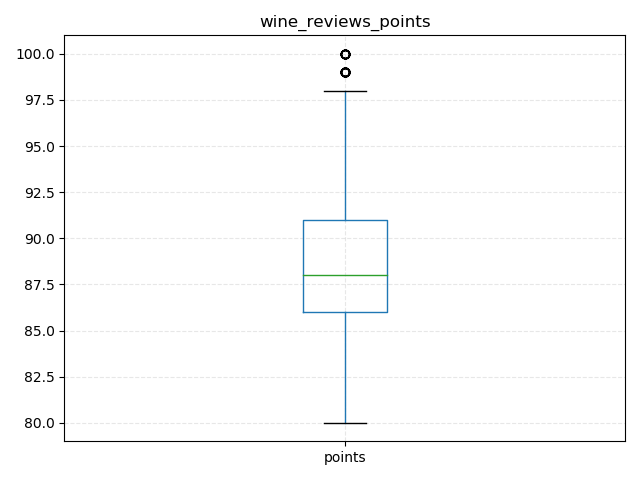
对标称属性'taster\_name'有：



对数值属性'points'有：

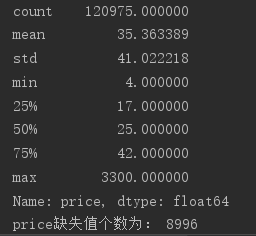
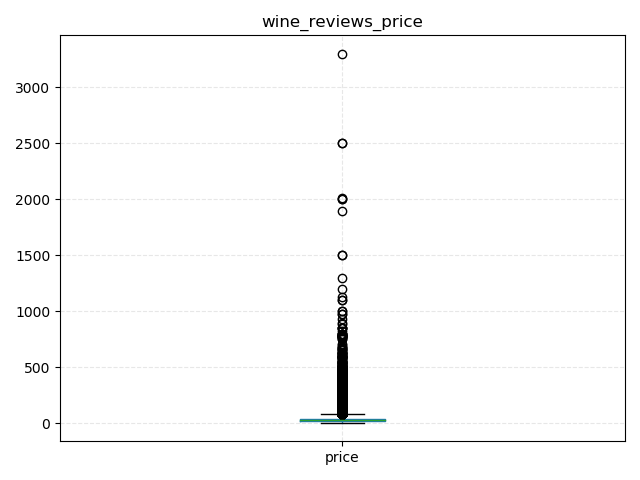






可以从第一个图看到五数概括分别为：80,86,88,91,100。缺失值的个数为0个。在盒图中，可以看到离群点的存在。

对数值属性'price'有：

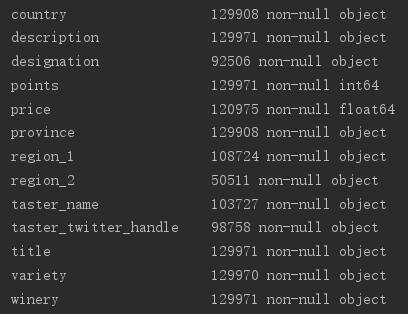
 



可以从第一个图看到五数概括分别为：4,17,25,42,3300。缺失值的个数为8996个。在盒图中，可以看到离群点的存在。

### 数据缺失的处理

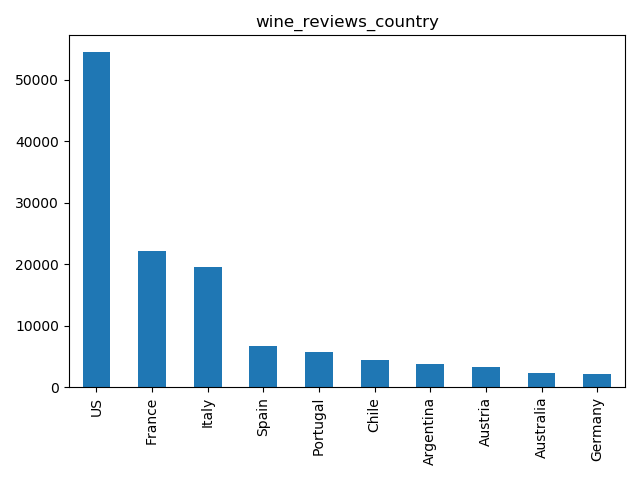
#### 分析缺失的原因



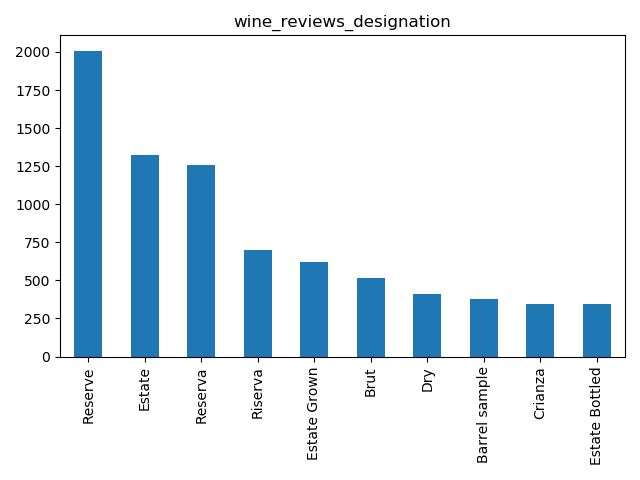
可以用.info()的方法得知每种属性的具体情况，我们可以看到，在我们要分析的属性中，points、winery都是没有缺失值的，我们要进行填补的是country、designation、price、province、taster\_name、region\_1、region\_2、variety这几个属性。缺失的原因很大一部分是可以不填写或者忘了填写了。

#### 将缺失部分剔除

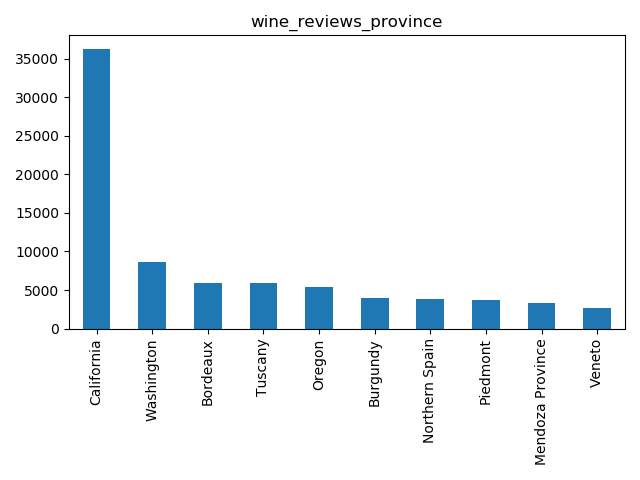
对标称属性'country'有：



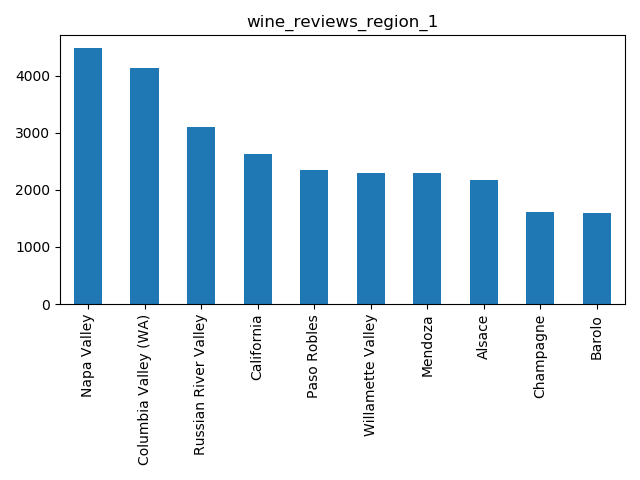
对标称属性'designation'有：



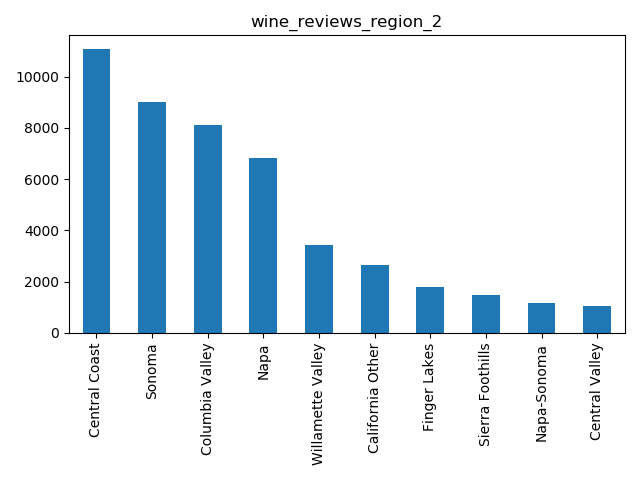
对标称属性'province'有：



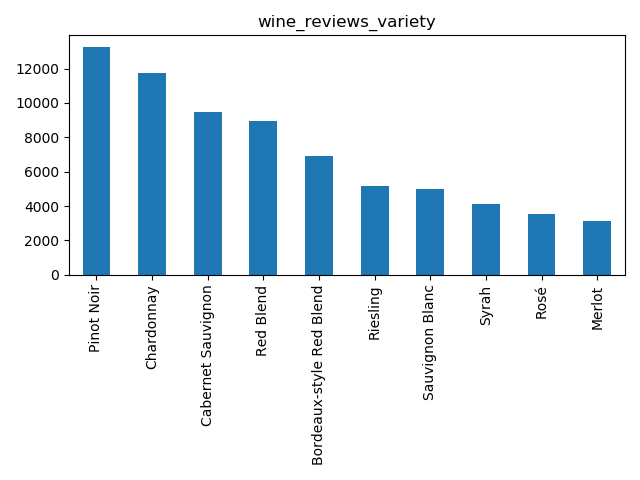
对标称属性'region\_1'有：



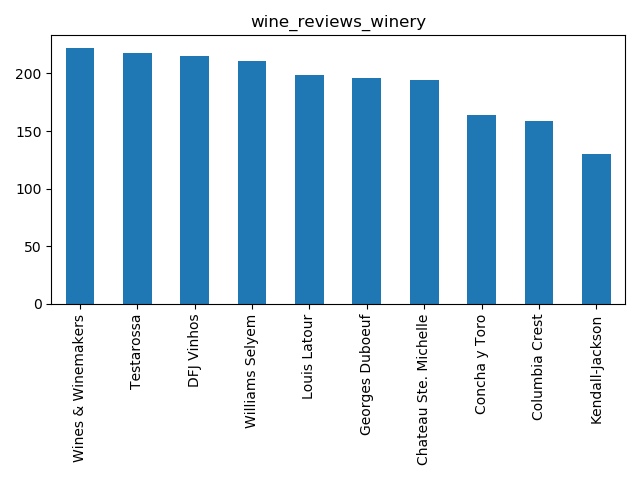
对标称属性'region\_2'有：



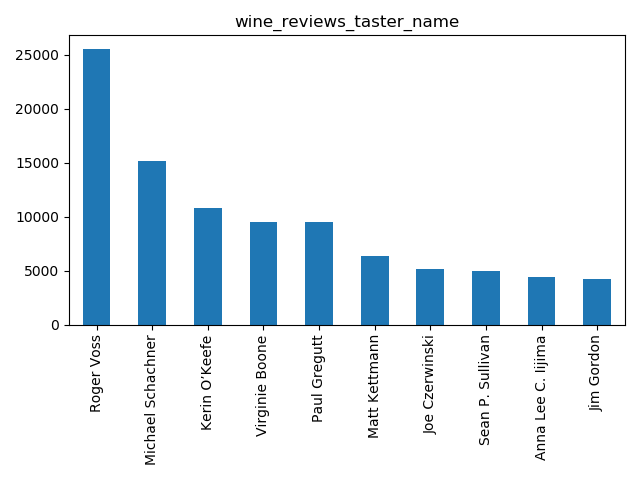
对标称属性'variety'有：



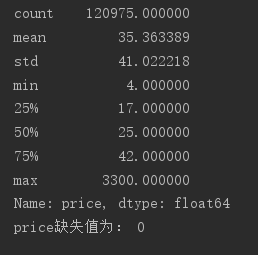
对标称属性'winery'有：

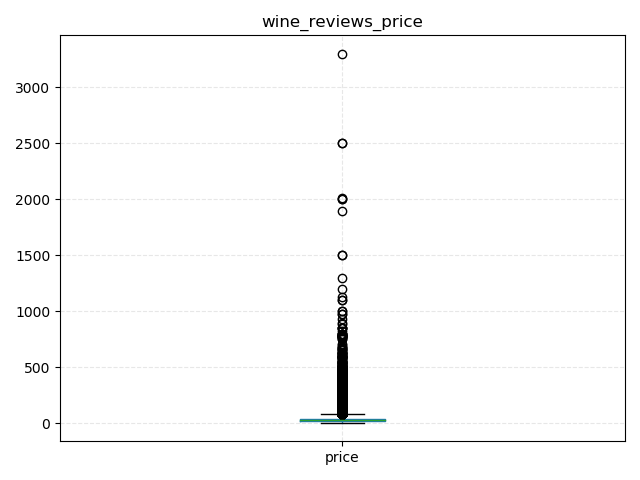


对标称属性'taster\_name'有：



对数值属性'price'有：



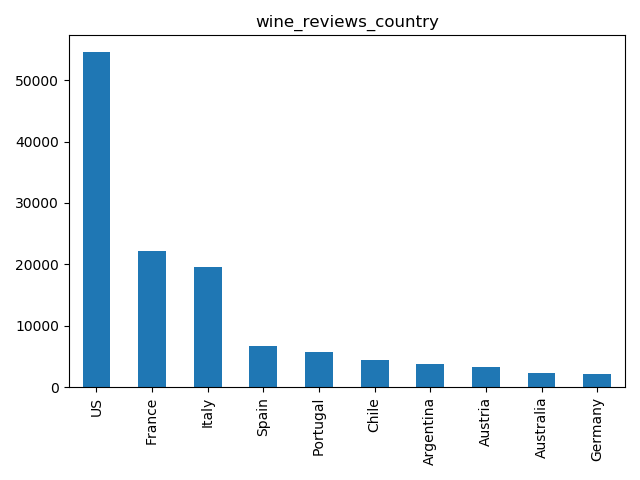




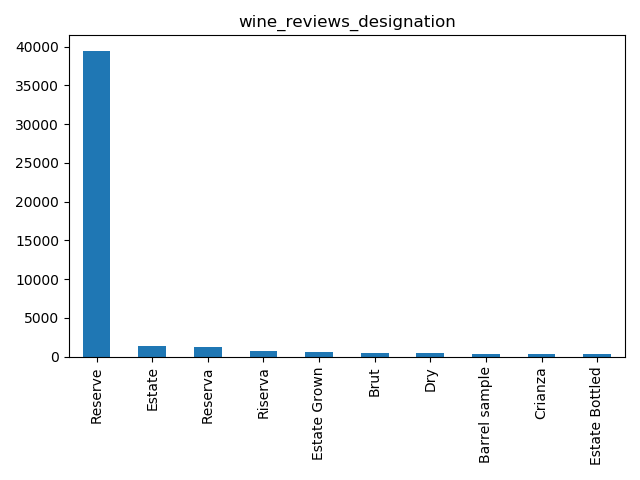
可以从第一个图看到五数概括分别为：4,17,25,42,3300。缺失值的个数为0个。在盒图中，可以看到离群点的存在。

#### 用最高频率值来填补缺失值

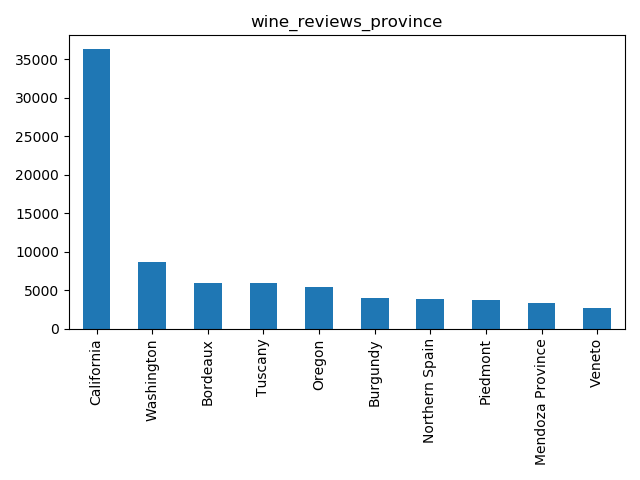
对标称属性'country'有：



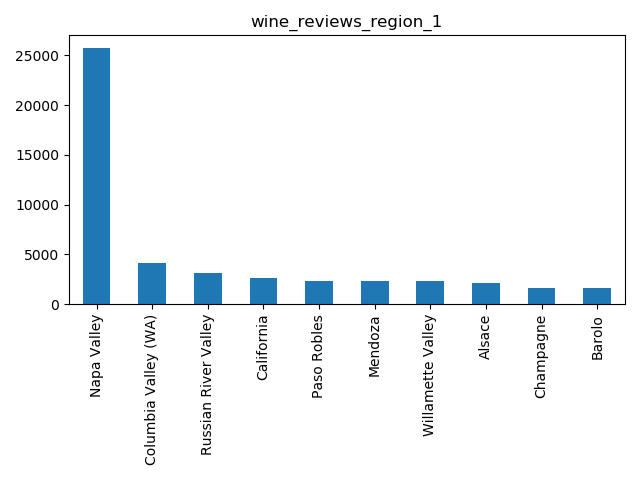
对标称属性'designation'有：



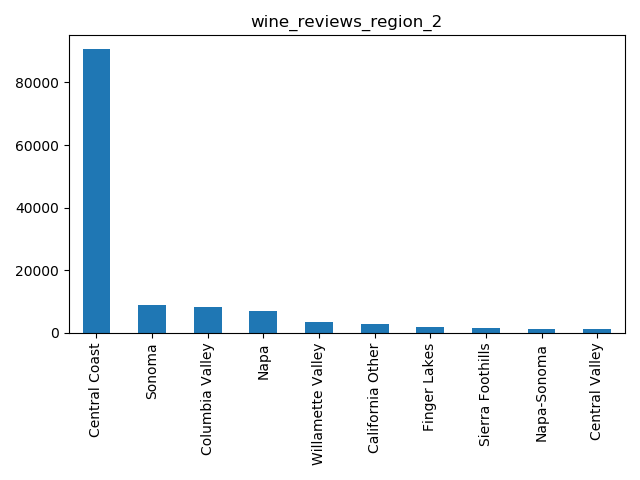
对标称属性'province'有：



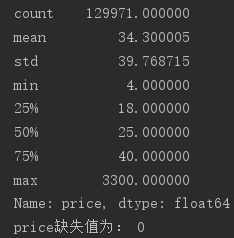
对标称属性'region\_1'有：

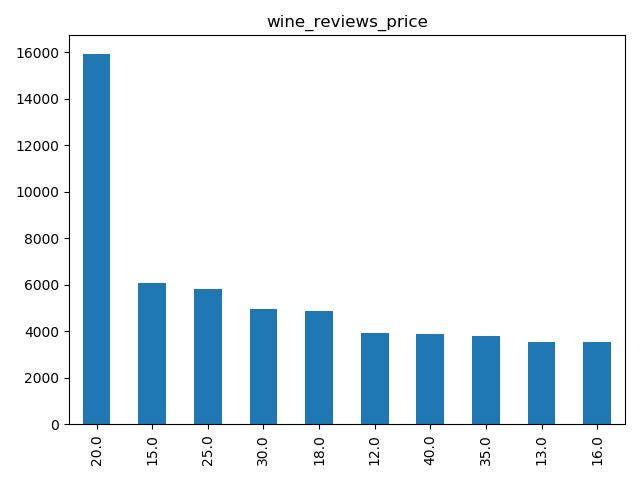


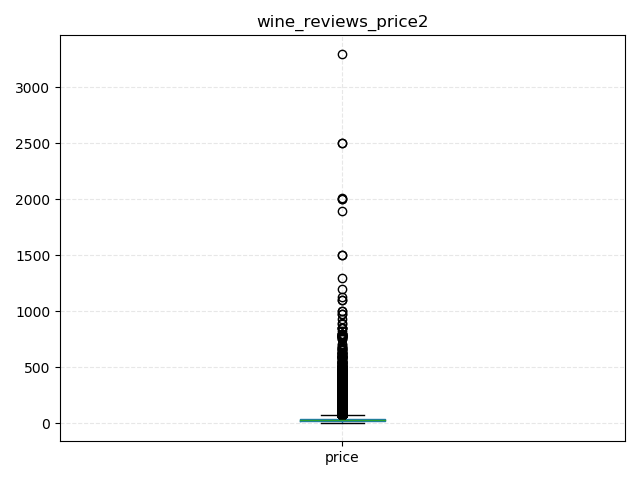
对标称属性'region\_2'有：



对数值属性'price'有：





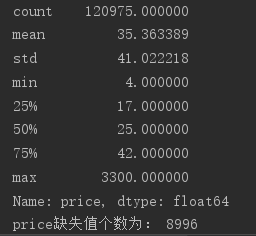
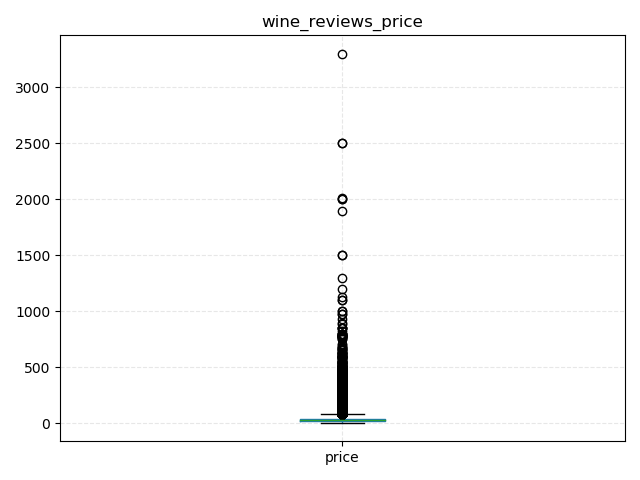


可以从第一个图看到五数概括分别为：4,18,25,40,3300。缺失值的个数为0个。在盒图中，可以看到离群点的存在。

#### 通过属性的相关关系来填补缺失值

使用SimpleImputer，用来填充数据里面的缺失值。可以从盒图与直方图看到前后的对比。

填补缺失值之前：

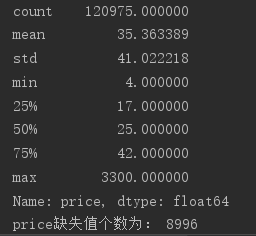
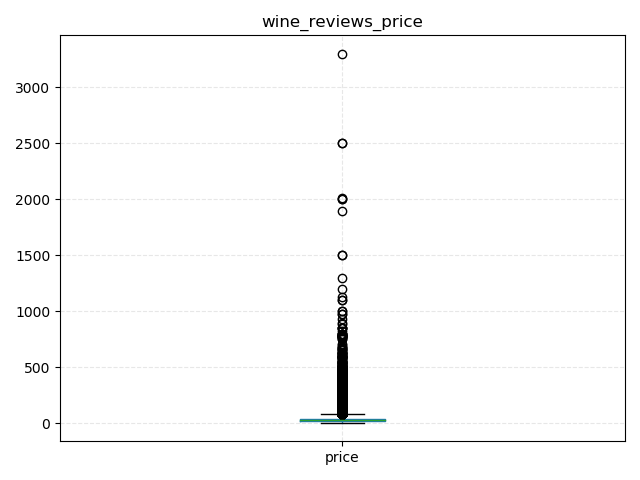


#### 通过数据对象之间的相似性来填补缺失值

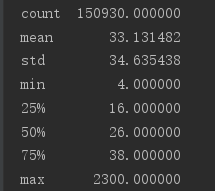
使用线性回归分类器LinearRegression，拟合points属性和price属性之间的相关关系，用训练好的分类器，基于已知的points值预测price的值。

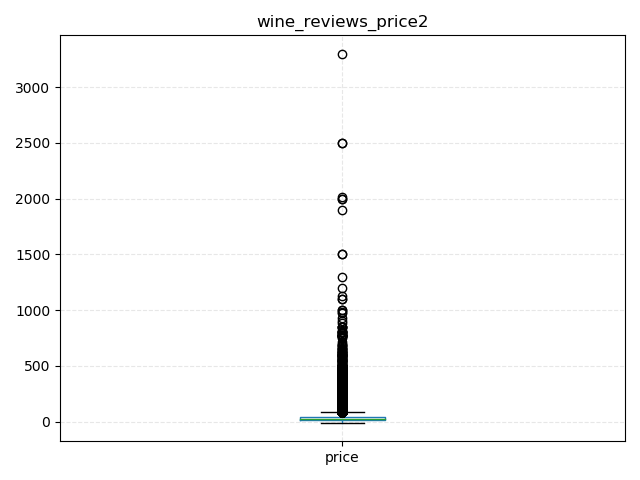
可以从盒图与直方图中看到前后的对比。

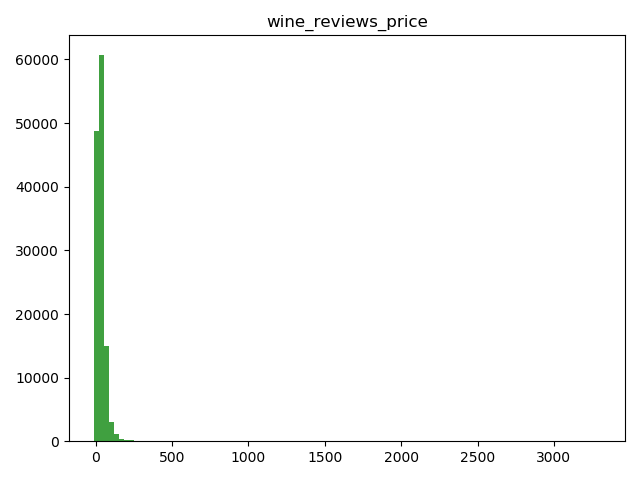
填补缺失值之前：

填补缺失值之后：

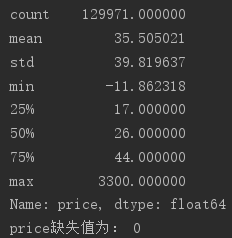


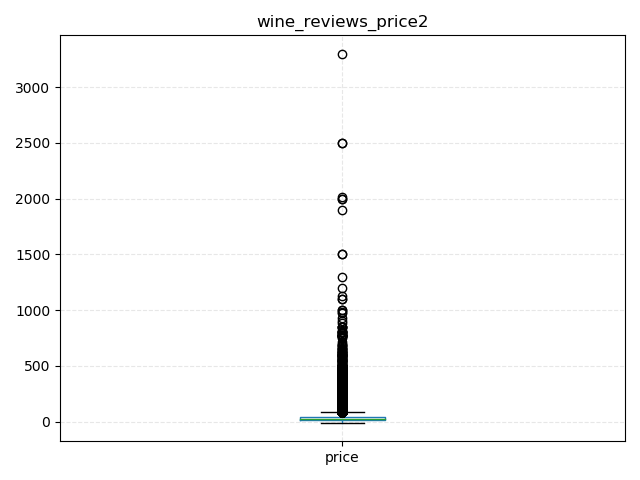


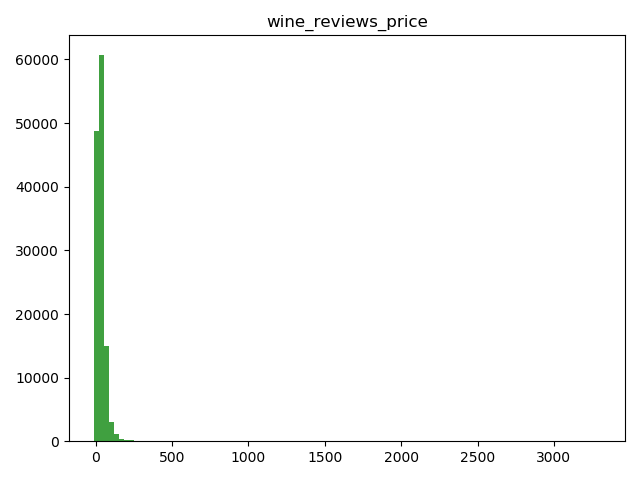




填补缺失值之后：





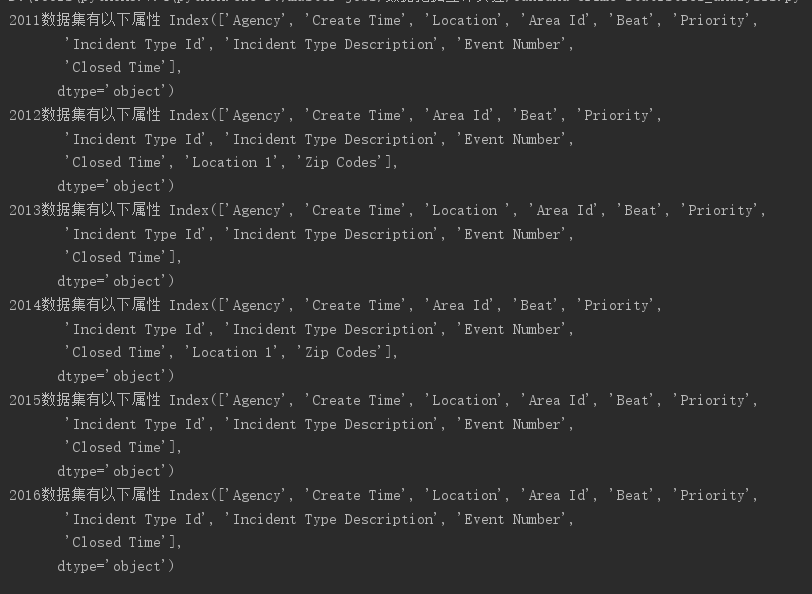


# 对于oakland-crime-statistics-2011-to-2016的数据分析与预处理

## 数据集说明

在这个数据集中，一共包含6个数据子集，分别为2011-2016年度的奥克兰犯罪情况。通过观察数据集可得知，这六年的数据属性基本一样，值得进行分析与预处理的有如下几个属性：Location, Area id, Incident Type id，Incident Type Descripe, EventNumber，其中2012年和2014年的属性为Location 1，单独拿出来进行讨论。由于Incident Type id与Incident Type Descripe一一对应，我们只对Incident Type id进行分析。EventNumber对应每一行数据，不具备重复性，不对齐进行分析。

在这个数据集上没有严格意义上的数值属性，只有标称属性，我们对标称属性进行分析与预处理，同时，由于没有数值属性，故用属性的相关关系和数据对象之间的相似性来填充缺失数据不可行，我们只使用剔除和最高频率两种方法进行填充。



数据集路径：

./data/Oakland-Crime-Statistics-2011-to-2016/records-for-2011.csv

./data/Oakland-Crime-Statistics-2011-to-2016/records-for-2012.csv

./data/Oakland-Crime-Statistics-2011-to-2016/records-for-2013.csv

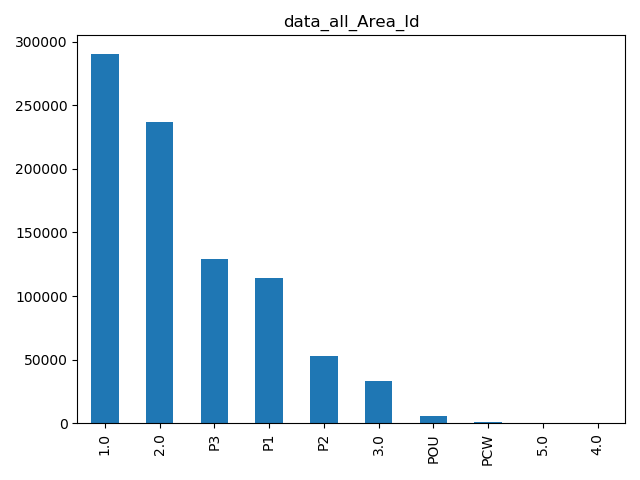
./data/Oakland-Crime-Statistics-2011-to-2016/records-for-2014.csv

./data/Oakland-Crime-Statistics-2011-to-2016/records-for-2015.csv

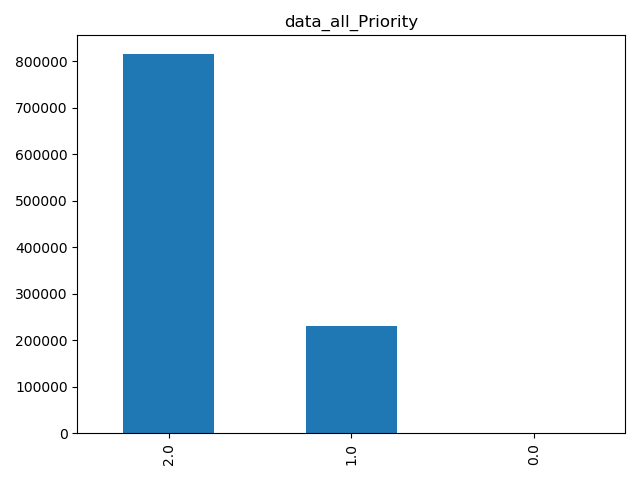
./data/Oakland-Crime-Statistics-2011-to-2016/records-for-2016.csv

## 数据可视化和摘要

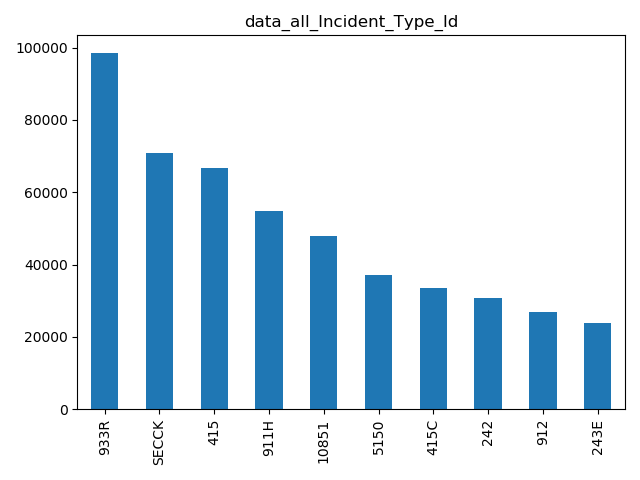
对标称数据Area Id有：



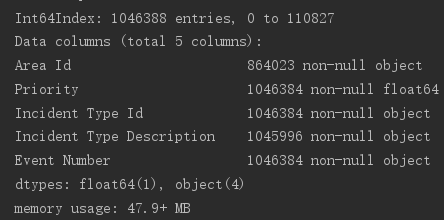
对标称数据Priority有：



对标称数据Incident Type Id有：



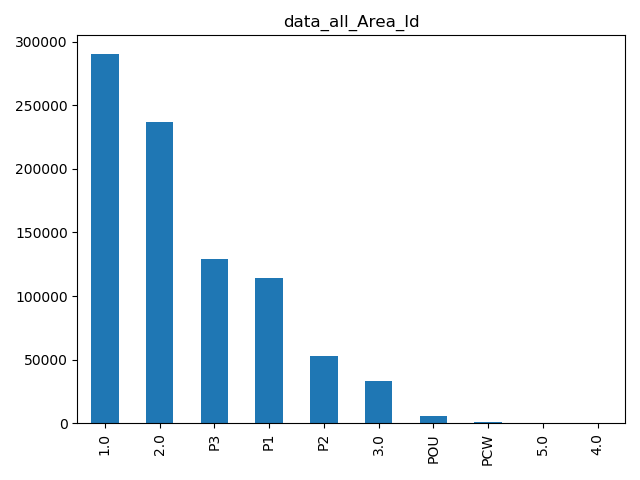
## 数据缺失的处理



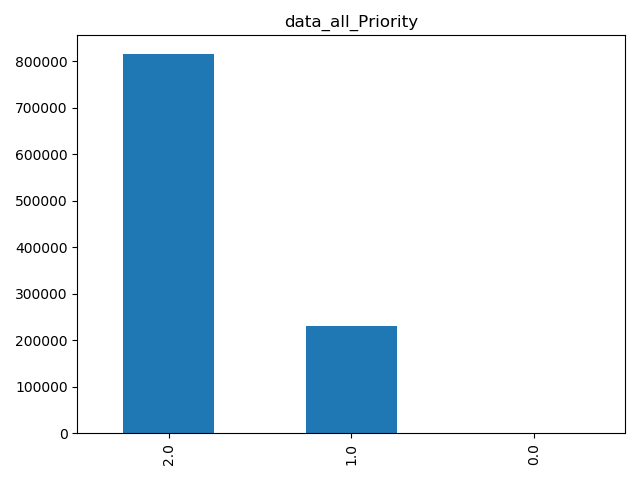
我们可以从.info()函数得知，在我们要分析的三个属性都有所缺失，我们都要对其进行填补。缺失的原因很大一部分是可以不填写或者忘了填写了。

### 将缺失部分剔除

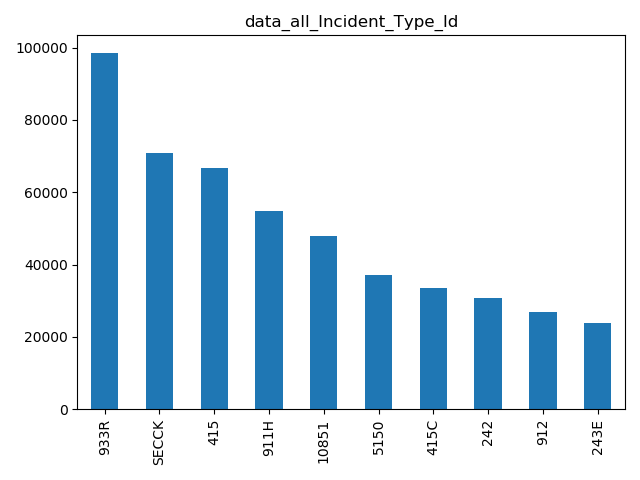
对标称数据Area Id有：



对标称数据Priority有：

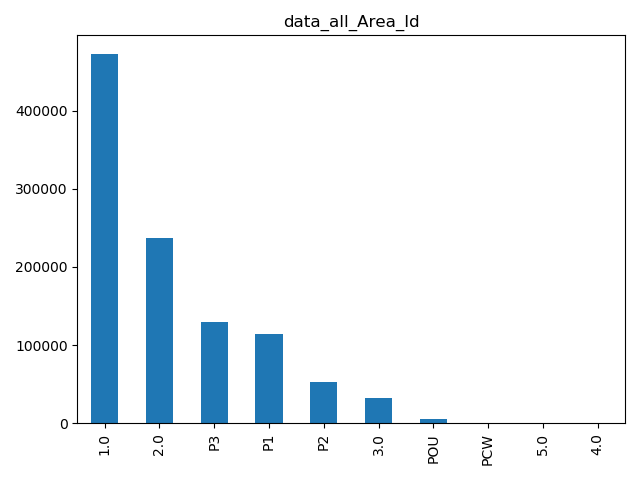


对标称数据Incident Type Id有：

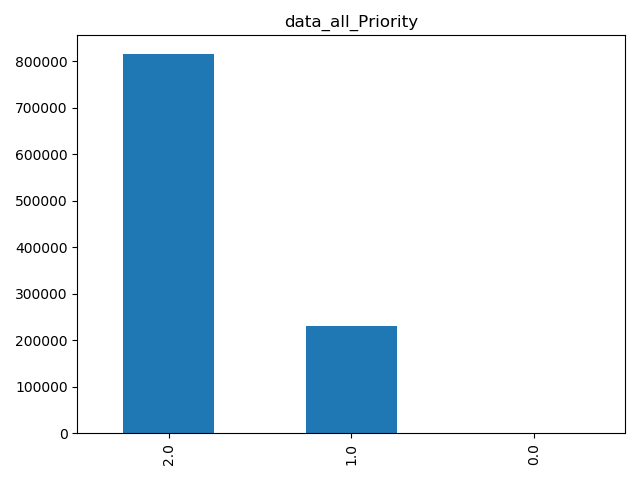


### 用最高频率值来填补缺失值

对标称数据Area Id有：



对标称数据Priority有：



对标称数据Incident Type Id有：

