作业1 —**winemag-data\_first150k**

通过观察数据集，发现该数据集标称属性包括：

'country',' description', 'designation',

'province', 'region\_1', 'region\_2', 'variety' , 'winery'

数值属性,包括：

'points' ,'price'

1 以‘province’为例，计算标称属性的频数如下（只展示部分结果）：

California 44508

Washington 9750

Tuscany 7281

Bordeaux 6111

Northern Spain 4892

Mendoza Province 4742

Oregon 4589

Catalonia 1352

Rhône Valley 1318

Other 889

….

Krania Olympus 1

Sterea Ellada 1

Dolenjska 1

Maipo Valley-Colchagua Valley 1

Langenlois 1

Pannon 1

Nevada 1

Terasele Dunarii 1

Zitsa 1

Vino da Tavola della Svizzera Italiana 1

Name: province, dtype: int64

2 计算全部数值属性的最大最小值等，如：

最大值：

points 100.0

price 2300.0

dtype: float64

最小值：

points 80.0

price 4.0

dtype: float64

均值：

points 87.888418

price 33.131482

dtype: float64

中位数：

points 88.0

price 24.0

dtype: float64

四分位数：

points 86.0

price 16.0

Name: 0.25, dtype: float64

缺失值的个数：

points 0

price 13695

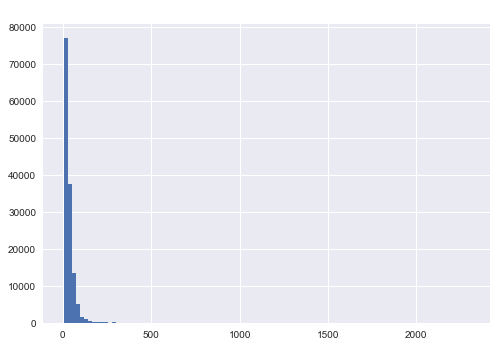
dtype: int64

3 直方图+qq图+盒图

以‘price’为例

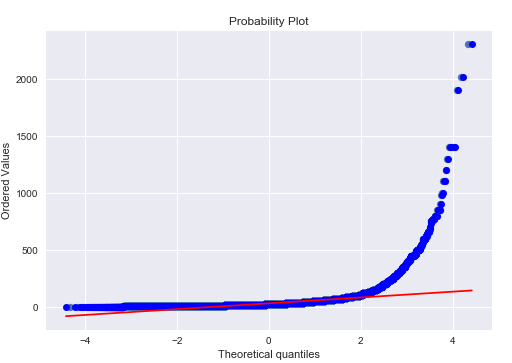
3.1 剔除缺失数据

**直方图（以price\_hist\_1.png存储）：**

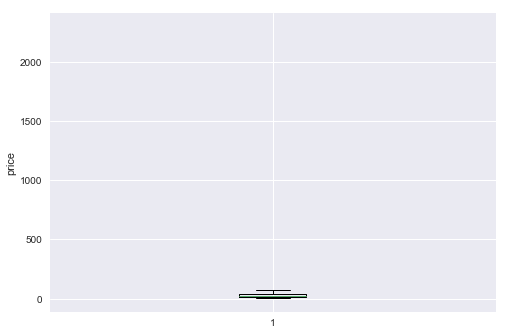


**QQ图（'price\_QQ\_1.png'存储）：**

并不是所有点基本上在一条直线上，所以这两个分布不是同一分布，也即不服从正态分布。

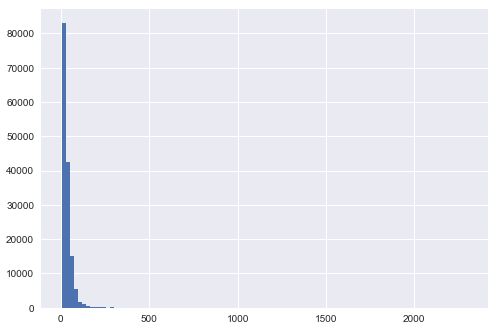


**盒图（'price\_box\_1.png'存储）：**

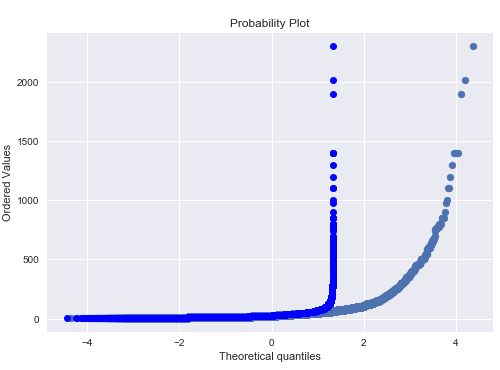


3.2 以最高频率值来填补缺失值

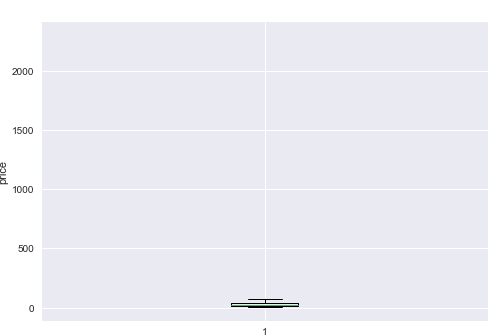
直方图：



QQ图：

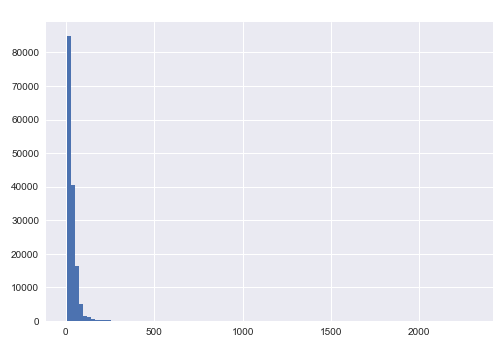


盒图：

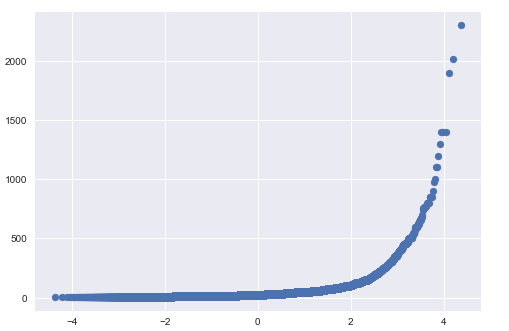


3.3 以属性间的相关关系来填补缺失值

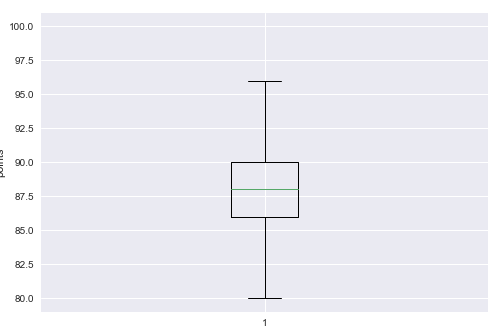
直方图：



QQ图：



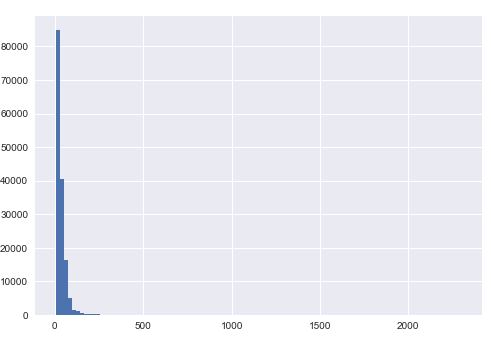
盒图：



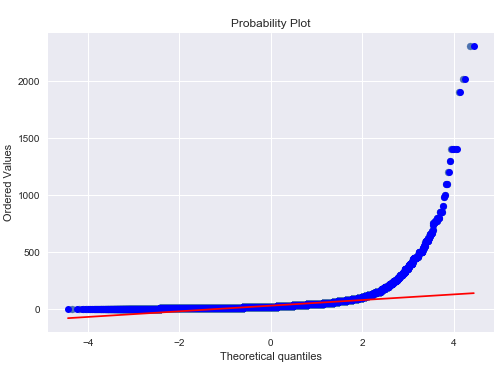
3.4 通过数据对象之间的相似性来填补缺失值

同3.2，但使用mean值

直方图：



QQ图：



盒图：

