**作业一分析报告**

1. 数据

[Wine Reviews](https://www.kaggle.com/zynicide/wine-reviews)：winemag-data\_first150k

[Trending YouTube Video Statistics](https://www.kaggle.com/datasnaek/youtube-new)：CAvideos （因为数值属性没有nan值，手动删除了部分数据属性的值）

语言：python

1. 数据一分析过程

2.1数据摘要

标称属性，给出每个可能取值的频数（以country为例）

US 62397

Italy 23478

France 21098

Spain 8268

Chile 5816

Argentina 5631

Portugal 5322

Australia 4957

New Zealand 3320

Austria 3057

Germany 2452

South Africa 2258

Greece 884

Israel 630

Hungary 231

Canada 196

Romania 139

Slovenia 94

Uruguay 92

Croatia 89

Bulgaria 77

Moldova 71

Mexico 63

Turkey 52

Georgia 43

Lebanon 37

Cyprus 31

Brazil 25

Macedonia 16

Serbia 14

Morocco 12

Luxembourg 9

England 9

India 8

Lithuania 8

Czech Republic 6

Ukraine 5

Switzerland 4

Bosnia and Herzegovina 4

South Korea 4

China 3

Egypt 3

Slovakia 3

Montenegro 2

Tunisia 2

Albania 2

Japan 2

US-France 1

Name: country, dtype: int64

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。（以price属性为例子）

count 137235.000000

mean 33.131482

std 36.322536

min 4.000000

25% 16.000000

50% 24.000000

75% 40.000000

max 2300.000000

Name: price, dtype: float64

数据含有缺省值情况：

country 5

description 0

designation 45735

points 0

price 13695

province 5

region\_1 25060

region\_2 89977

variety 0

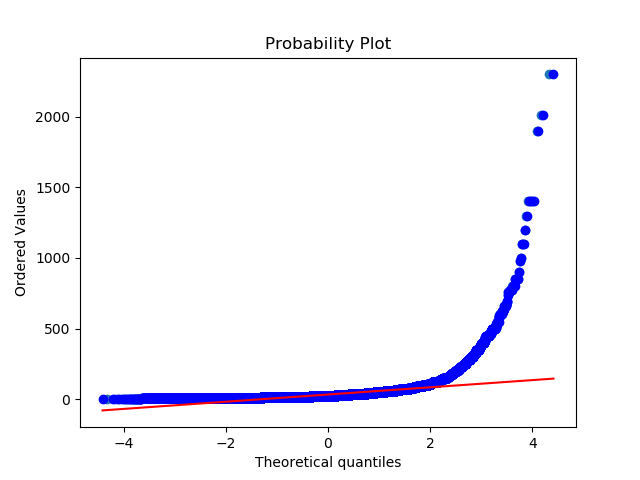
winery 0

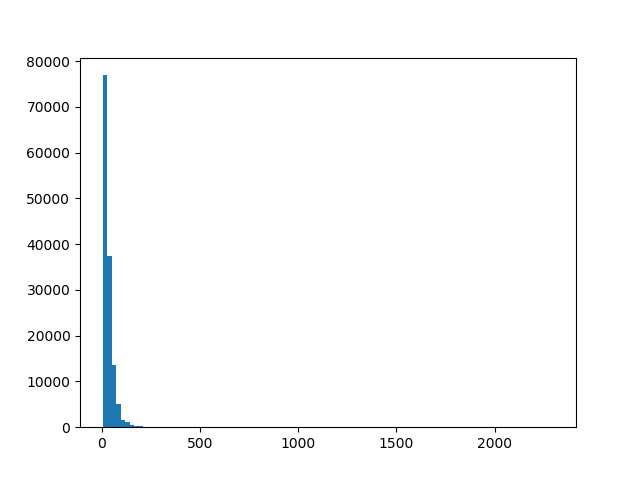
dtype: int64

2.2数据缺失处理

* 将缺失部分剔除

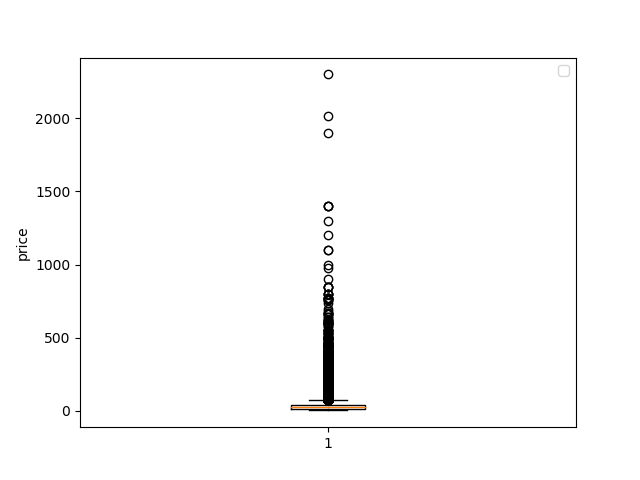
绘制直方图，并用qq图检测是否为正态分布



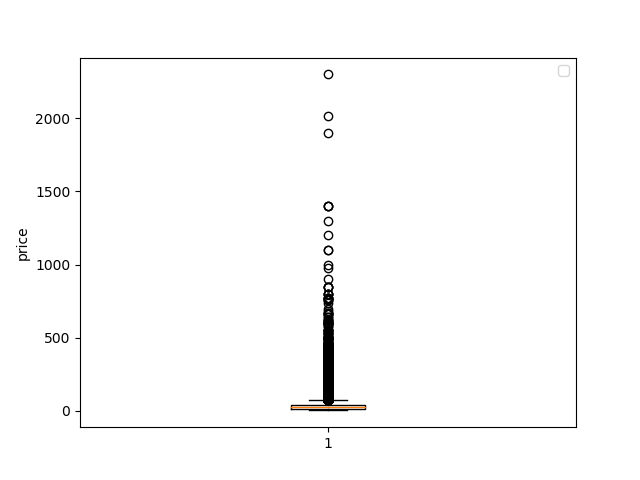
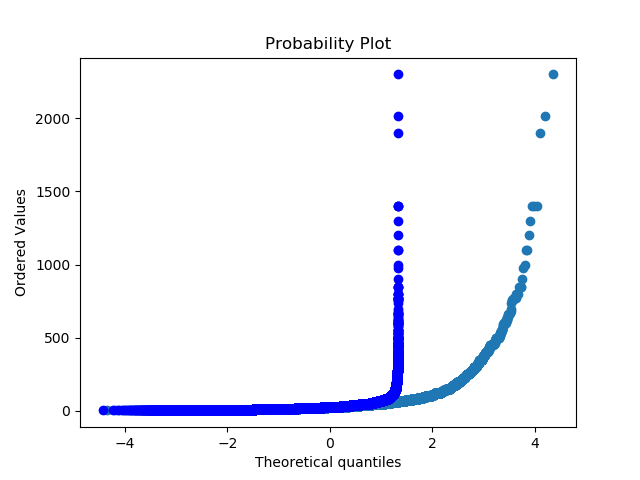


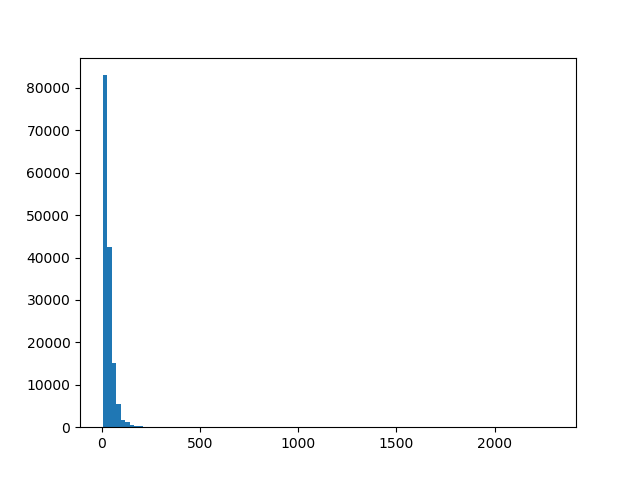
显然不符合正态分布

绘制盒图，对离群值进行识别

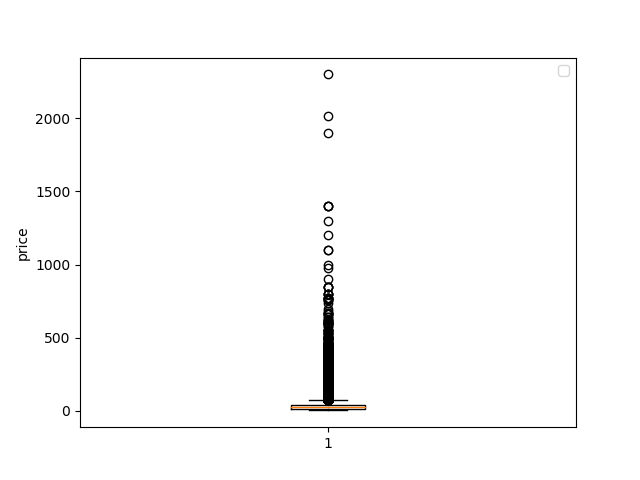
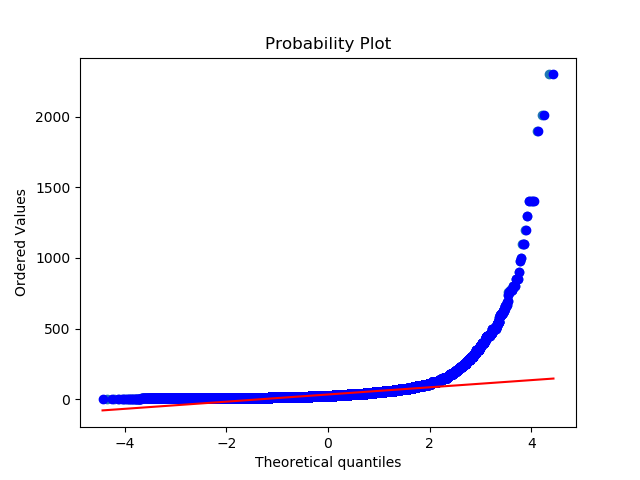
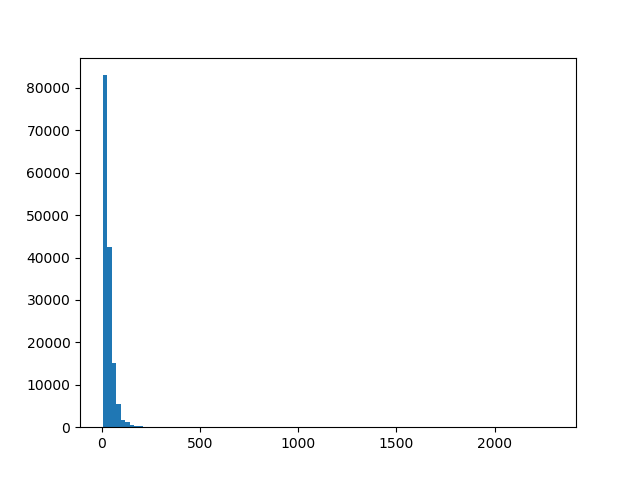


* 用最高频率值来填补缺失值

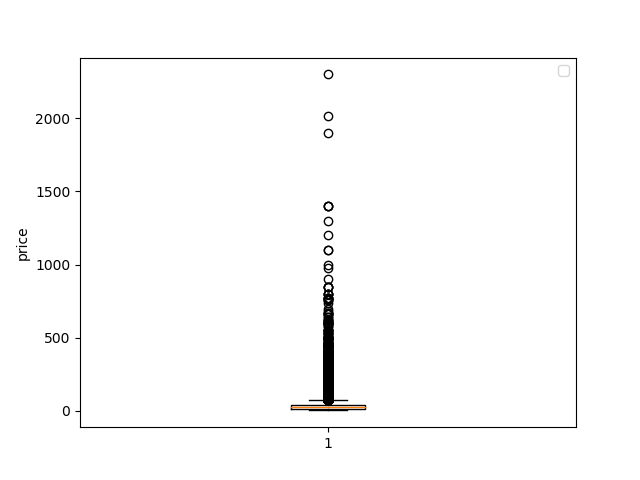
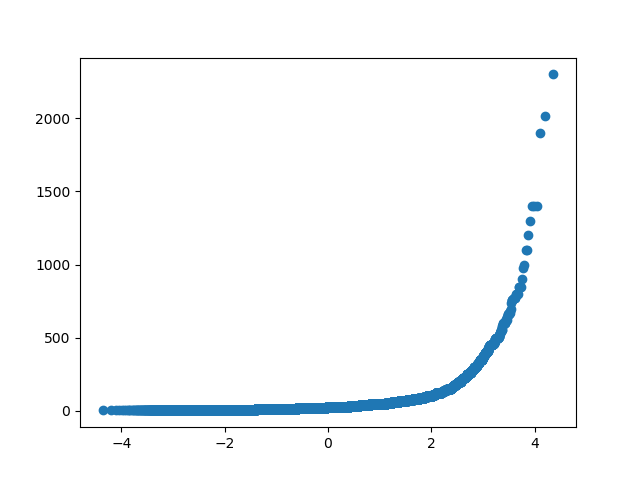
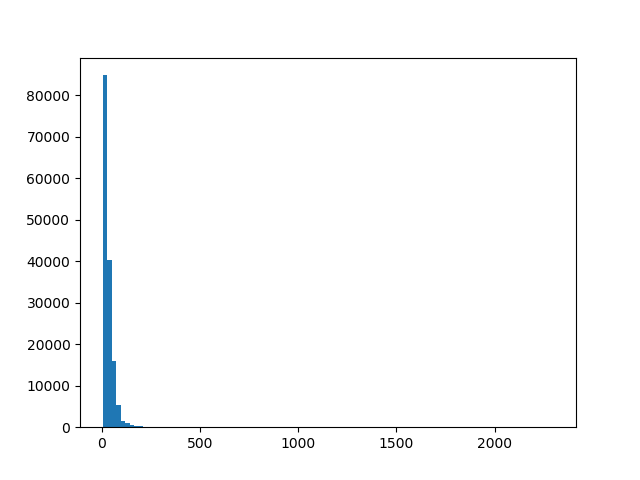




* 通过属性的相关关系来填补缺失值（均值）



* 通过数据对象之间的相似性来填补缺失值（随机森林预测可能的值填充缺值）



1. 数据二分析过程

3.1数据摘要

标称属性，给出每个可能取值的频数（以category\_id为例）

24 9840

10 6431

26 4067

23 3452

22 3164

25 2448

28 2377

1 2334

17 2164

27 1644

15 915

20 809

19 402

2 384

29 57

43 57

2018-05-04T01:47:42.000Z 21

2018-05-07T22:34:57.000Z 15

2018-04-19T23:20:40.000Z 13

2018-02-24T00:10:25.000Z 12

2018-04-16T19:32:09.000Z 12

2018-05-23T18:28:58.000Z 12

2018-04-25T23:11:00.000Z 10

2018-02-01T16:18:35.000Z 8

2018-02-13T19:59:13.000Z 8

2017-12-30T08:00:09.000Z 7

2017-12-01T09:06:02.000Z 6

2018-01-02T16:30:25.000Z 6

2018-01-09T12:52:52.000Z 5

2018-01-04T08:04:03.000Z 5

2017-11-15T15:17:28.000Z 5

2018-01-26T13:23:59.000Z 5

2017-12-06T22:31:32.000Z 5

2017-11-17T14:00:06.000Z 5

2018-01-24T03:25:21.000Z 5

2018-02-08T20:16:00.000Z 3

2018-02-09T23:07:28.000Z 3

94399 1

2017-12-19T02:06:23.000Z 1

2017-12-06T14:56:11.000Z 1

2018-02-25T15:00:00.000Z 1

The song is Carol Of The Bells 1

2018-01-06T16:08:49.000Z 1

2018-02-08T14:00:08.000Z 1

97535 1

Name: category\_id, dtype: int64

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。（以likes属性为例子）

count 4.065400e+04

mean 7.427396e+04

std 2.294329e+05

min 0.000000e+00

25% 5.293000e+03

50% 1.789950e+04

75% 5.523125e+04

max 5.613827e+06

Name: likes, dtype: float64

数据含有缺省值情况：

video\_id 0

trending\_date 0

title 0

channel\_title 0

category\_id 2

publish\_time 2

tags 3

views 3

likes 72

dislikes 3

comment\_count 3

thumbnail\_link 3

comments\_disabled 3

ratings\_disabled 5

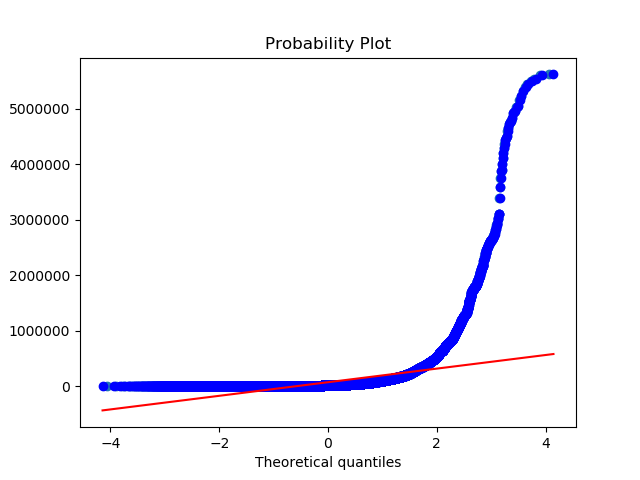
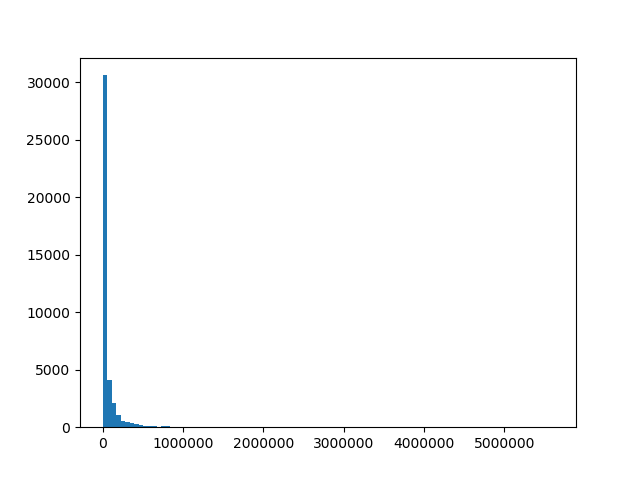
video\_error\_or\_removed 6

description 729

2.2数据缺失处理

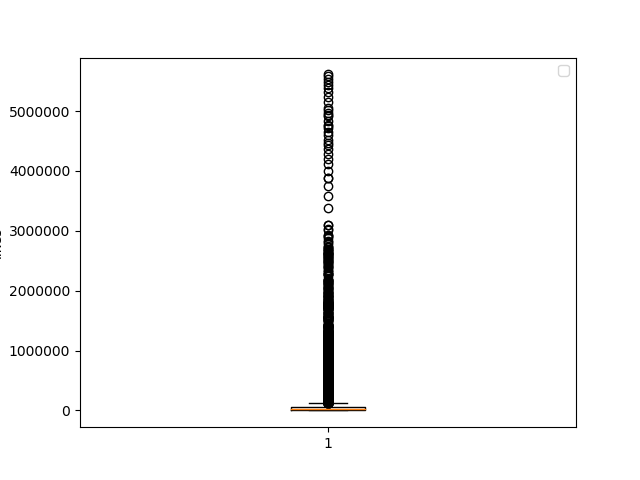
* 将缺失部分剔除

绘制直方图，并用qq图检测是否为正态分布

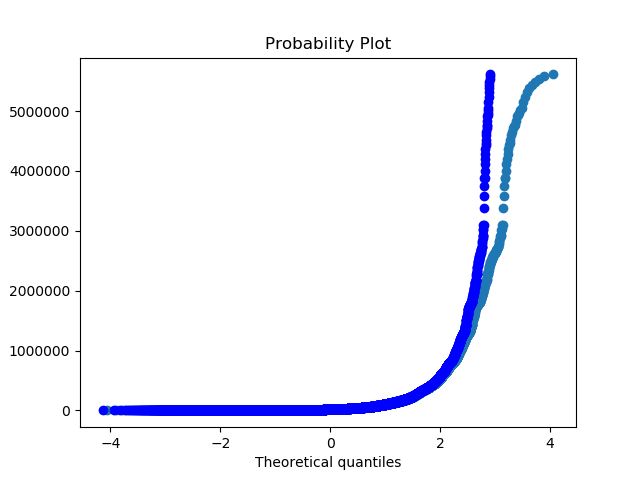
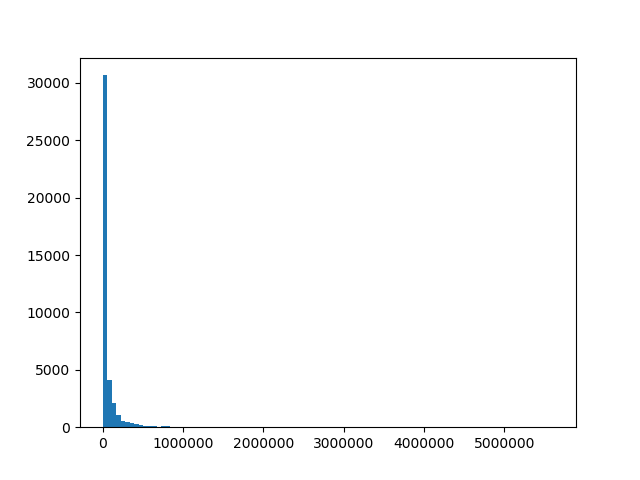


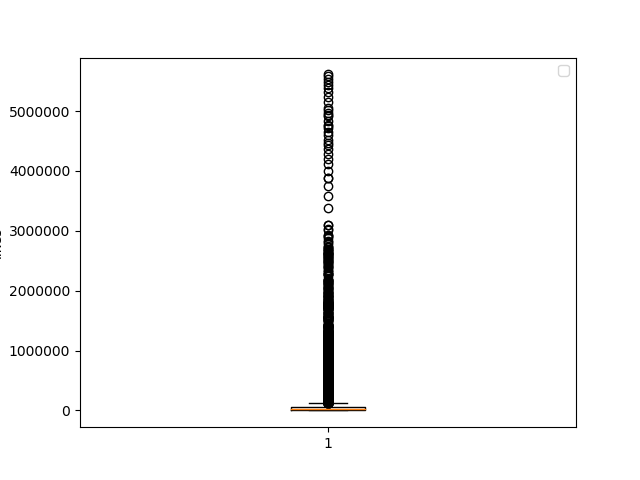
显然不符合正态分布

绘制盒图，对离群值进行识别

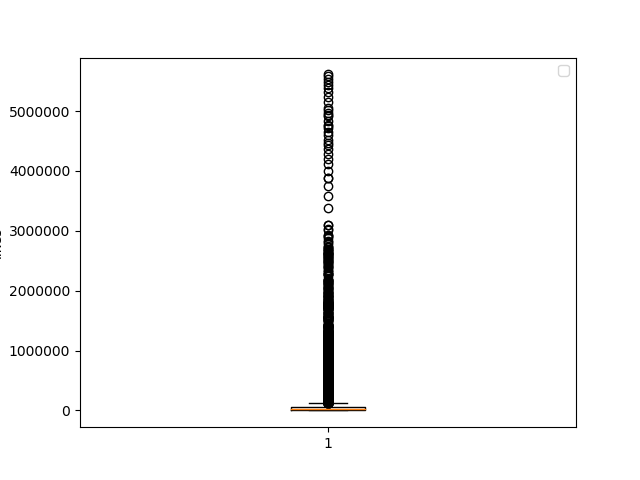
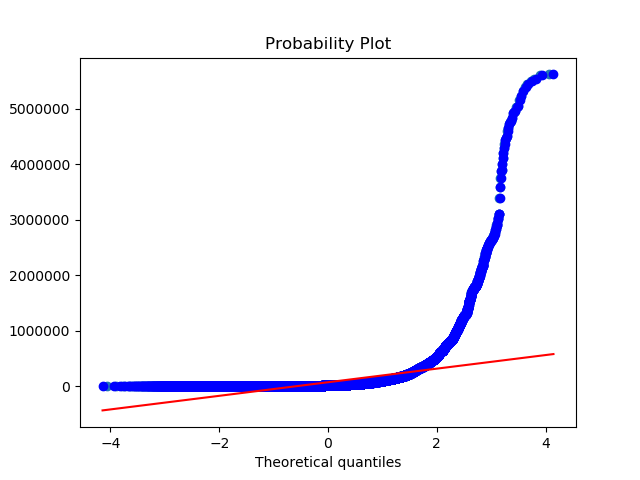
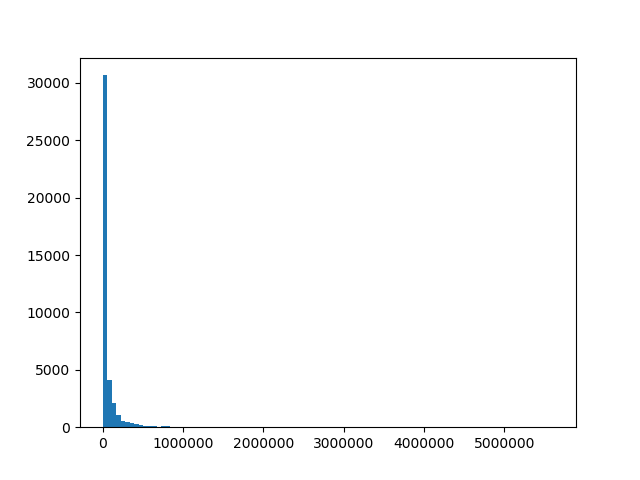


* 用最高频率值来填补缺失值





* 通过属性的相关关系来填补缺失值（均值）



* 通过数据对象之间的相似性来填补缺失值（随机森林预测可能的值填充缺值）

