互评作业1: 数据探索性分析与数据预处理-1

3120215523 金玉卿

1. 问题描述

自行选择2个数据集进行探索性分析与预处理

所选数据集： Wine Reviews

2. 数据集

wine reviews

包含2个csv文件

- winemag-data_first150k.csv

  包含10列和15万条葡萄酒评论

- winemag-data-130k-v2.csv

  包含10列和13万行葡萄酒评论

导入库

In [5]:

```
import matplotlib
import numpy as np
import pandas as pd
%matplotlib inline
```

载入数据

In [6]:

```
path_15k = "data/wine reviews/winemag-data_first150k.csv"
path_13k = "data/wine reviews/winemag-data-130k-v2.csv"
data_15k = pd.read_csv(path_15k)
```

数据属性

In [7]:

```
1 data_15k.head(5)
```

Out[7]:

	Unnamed: 0	country	description	designation	points	price	province	region_1	region
0	0	US	This tremendous 100% varietal wine hails from ...	Martha's Vineyard	96	235.0	California	Napa Valley	Na
1	1	Spain	Ripe aromas of fig, blackberry and cassis are ...	Carodorum Selección Especial Reserva	96	110.0	Northern Spain	Toro	Na
2	2	US	Mac Watson honors the memory of a wine once ma...	Special Selected Late Harvest	96	90.0	California	Knights Valley	Sonor
3	3	US	This spent 20 months in 30% new French oak, an...	Reserve	96	65.0	Oregon	Willamette Valley	Willame Vall
4	4	France	This is the top wine from La Bégude, named aft...	La Brûlade	95	66.0	Provence	Bandol	Na

In [8]:

```
1 data_15k.dtypes
```

Out[8]:

```
Unnamed: 0       int64
country         object
description     object
designation     object
points           int64
price          float64
province        object
region_1        object
region_2        object
variety         object
winery          object
dtype: object
```

- country 国家
- desprition 描述
- designation 葡萄酒庄

- points 得分
- price 价格
- province 省份
- region_1 区域1
- region_2 区域2
- variety 葡萄种类
- winery 酿酒厂

3. 数据分析

3.1 数据可视化和摘要

3.1.1 country属性

标称属性，给出每个可能取值的频数

In [9]:

```
attri = "country"
data_15k[attri].value_counts(dropna=False)
```

Out[9]:

```
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
NaN                             5
Ukraine                         5
Bosnia and Herzegovina          4
South Korea                     4
Switzerland                     4
Slovakia                        3
Egypt                           3
China                           3
Tunisia                         2
Montenegro                      2
Albania                         2
Japan                           2
US-France                       1
Name: country, dtype: int64
```

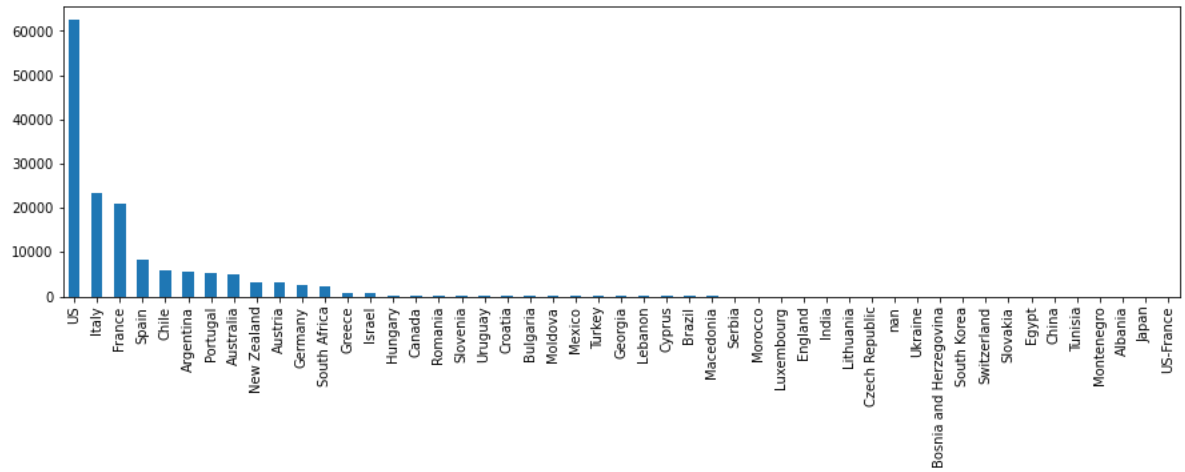
使用直方图可视化

In [10]:

```
1 data_15k[attri].value_counts(dropna = False).plot(kind="bar",figsize=(15,4))
```

Out[10]:

<AxesSubplot:>



3.1.2 designation属性

标称属性，给出每个可能取值的频数

In [11]:

```
1 attri = "designation"
2 data_15k[attri].value_counts(dropna = False)
```

Out[11]:

```
NaN                                   45735
Reserve                                2752
Reserva                                1810
Estate                                 1571
Barrel sample                          1326
                                      ...
Extra Dry Rosè                            1
Meredith Estate Méthode a l'Ancienne      1
Mythos                                    1
Taberner                                  1
Flying Rooster Vineyard                   1
Name: designation, Length: 30622, dtype: int64
```

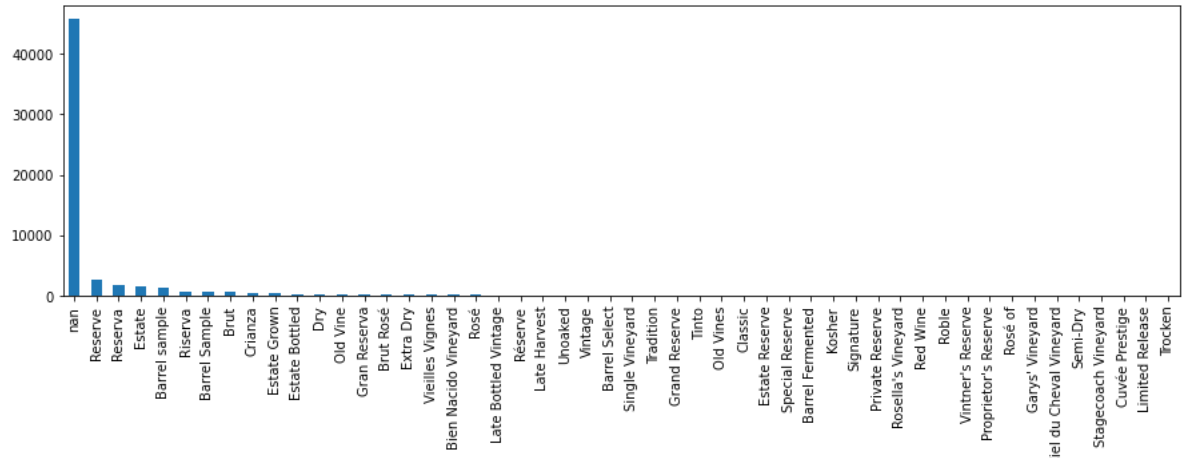
使用直方图可视化，只显示前50项

In [12]:

```
1 data_15k[attri].value_counts(dropna = False)[:50].plot(kind="bar",figsize=(15,4))
```

Out[12]:

<AxesSubplot:>



3.1.3 province属性

标称属性，给出每个可能取值的频数

In [13]:

```
1 attri = "province"
2 data_15k[attri].value_counts(dropna = False)
```

Out[13]:

```
California          44508
Washington           9750
Tuscany              7281
Bordeaux             6111
Northern Spain       4892
                    ...
Viile Timis             1
Pannon                  1
Colchagua Costa         1
Valais                  1
Neuchâtel               1
Name: province, Length: 456, dtype: int64
```

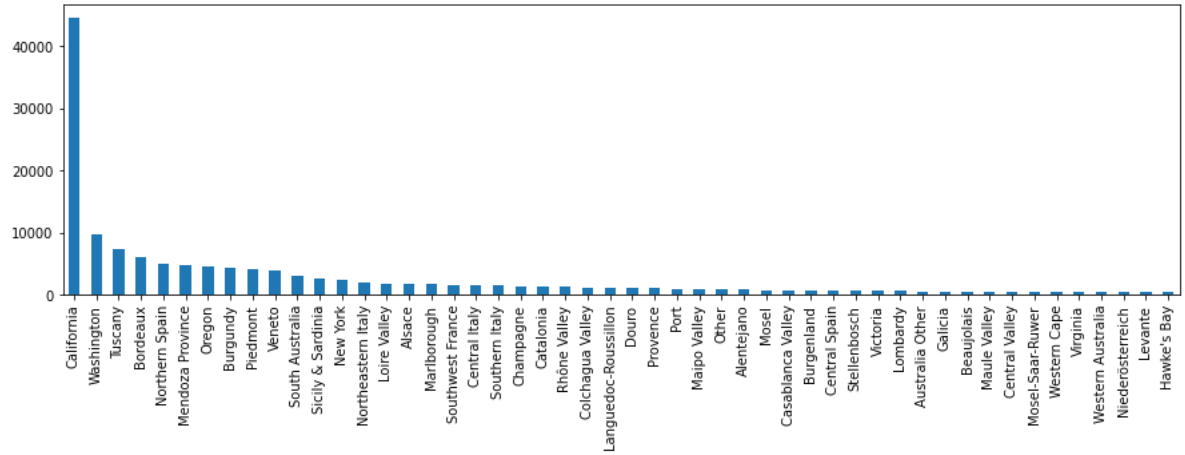
使用直方图可视化，只显示前50项

In [14]:

```
1 data_15k[attri].value_counts(dropna = False)[:50].plot(kind="bar",figsize=(15,4))
```

Out[14]:

<AxesSubplot:>



3.1.4 region_1属性

标称属性，给出每个可能取值的频数

In [15]:

```
1 attri = "region_1"
2 data_15k[attri].value_counts(dropna = False)
```

Out[15]:

```
NaN                                        25060
Napa Valley                                 6209
Columbia Valley (WA)                        4975
Mendoza                                     3586
Russian River Valley                        3571
                                           ...
Mitterberg                                     1
Mazoyeres-Chambertin                           1
Monterey County-Napa County-Sonoma County      1
Prince Edward County                           1
Mâcon-Mancey                                   1
Name: region_1, Length: 1237, dtype: int64
```

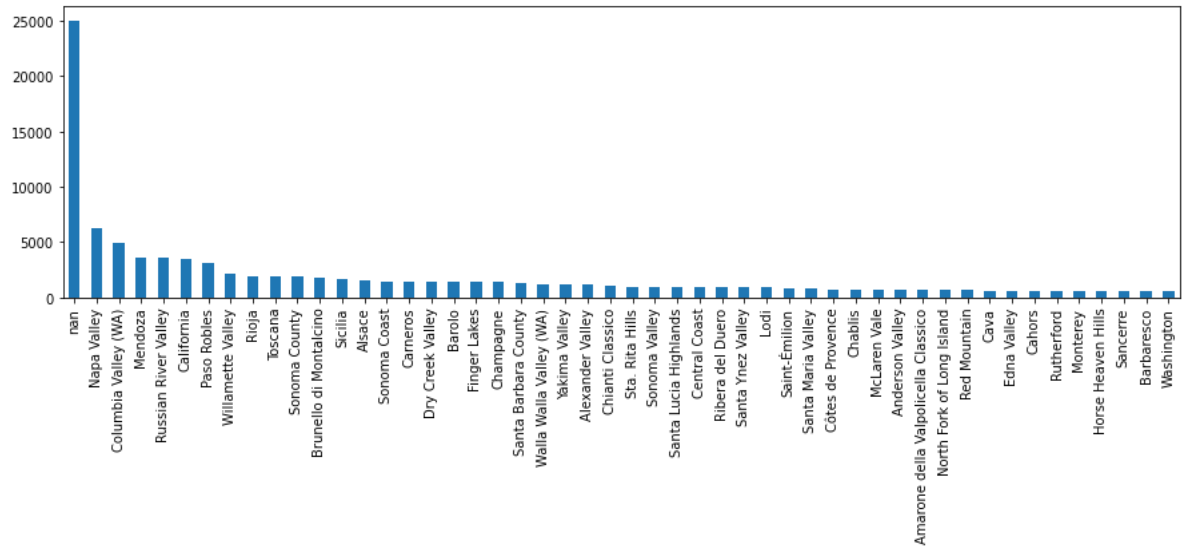
使用直方图可视化,只显示前50项

In [16]:

```
1 data_15k[attri].value_counts(dropna = False)[:50].plot(kind="bar",figsize=(15,4))
```

Out[16]:

<AxesSubplot:>



3.1.5 region_2属性

标称属性，给出每个可能取值的频数

In [17]:

```
attri = "region_2"
data_15k[attri].value_counts(dropna = False)
```

Out[17]:

```
NaN                      89977
Central Coast            13057
Sonoma                   11258
Columbia Valley           9157
Napa                      8801
California Other          3516
Willamette Valley         3181
Mendocino/Lake Counties   2389
Sierra Foothills          1660
Napa-Sonoma               1645
Finger Lakes              1510
Central Valley            1115
Long Island                771
Southern Oregon            662
Oregon Other               661
North Coast                632
Washington Other           593
South Coast                198
New York Other             147
Name: region_2, dtype: int64
```

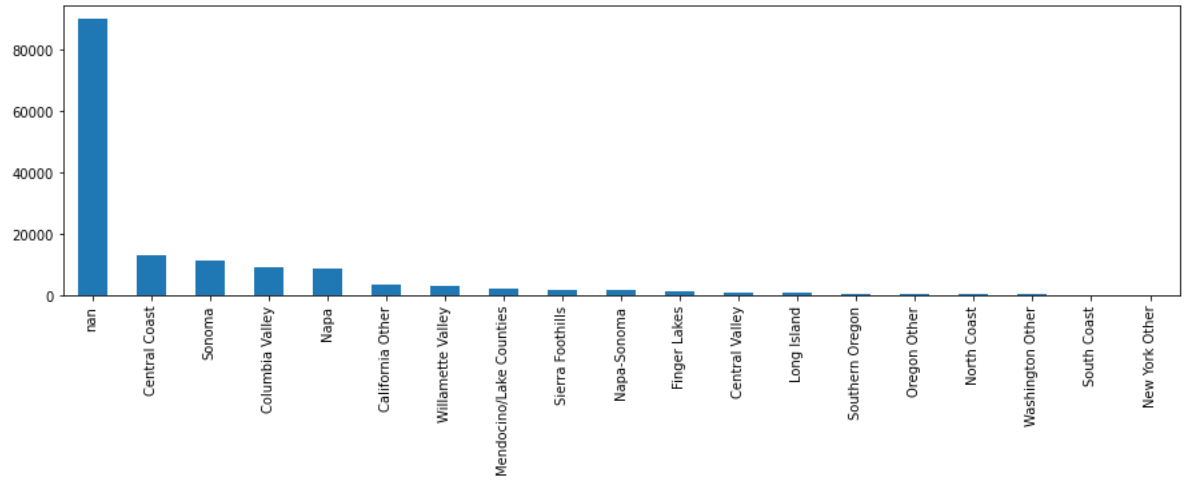
使用直方图可视化

In [18]:

```
data_15k[attri].value_counts(dropna = False).plot(kind="bar",figsize=(15,4))
```

Out[18]:

<AxesSubplot:>



3.1.6 variety属性

标称属性，给出每个可能取值的频数

In [19]:

```
attri = "variety"
data_15k[attri].value_counts(dropna = False)
```

Out[19]:

```
Chardonnay                  14482
Pinot Noir                  14291
Cabernet Sauvignon          12800
Red Blend                   10062
Bordeaux-style Red Blend     7347
                            ...
Jacquez                         1
Moscato di Noto                 1
Moscatel de Alejandría          1
Forcallà                        1
Syrah-Carignan                  1
Name: variety, Length: 632, dtype: int64
```

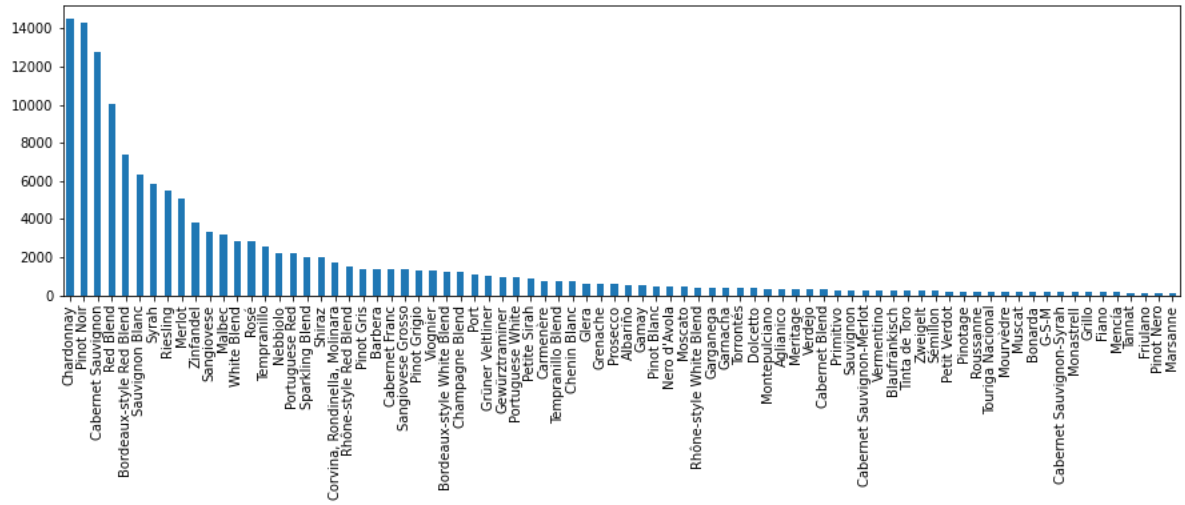
使用直方图可视化,只显示前80项

In [20]:

```
data_15k[attri].value_counts(dropna = False)[:80].plot(kind="bar",figsize=(15,4))
```

Out[20]:

<AxesSubplot:>



3.1.7 winery属性

标称属性，给出每个可能取值的频数

In [21]:

```
attri = "winery"
data_15k[attri].value_counts(dropna = False)
```

Out[21]:

```
Williams Selyem          374
Testarossa               274
DFJ Vinhos               258
Chateau Ste. Michelle    225
Columbia Crest           217
                        ...
Château Capbern            1
Dovev                      1
Poggioventoso              1
Château d'Estoublon        1
Costa del Sol              1
Name: winery, Length: 14810, dtype: int64
```

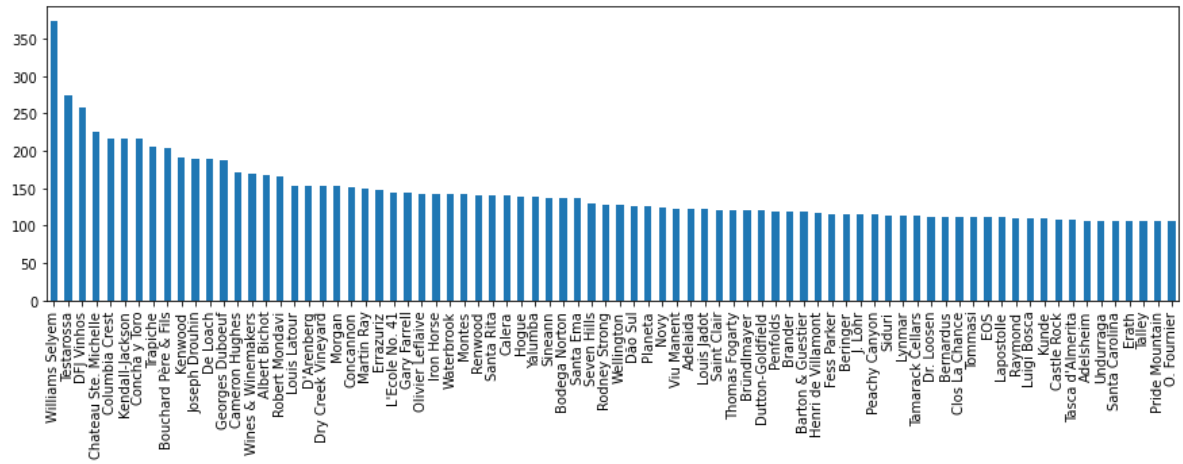
使用直方图可视化,只显示前80项

In [22]:

```
data_15k[attri].value_counts(dropna = False)[:80].plot(kind="bar",figsize=(15,4))
```

Out[22]:

<AxesSubplot:>



3.1.8 points属性

数值属性，给出五数概括

In [23]:

```
attri = "points"
for i in range(0,5):
    q = i * 0.25
    print("Q%d:"%(i),end=" ")
    print(data_15k[attri].quantile(q))
```

```
Q0: 80.0
Q1: 86.0
Q2: 88.0
Q3: 90.0
Q4: 100.0
```

给出缺失值个数

In [24]:

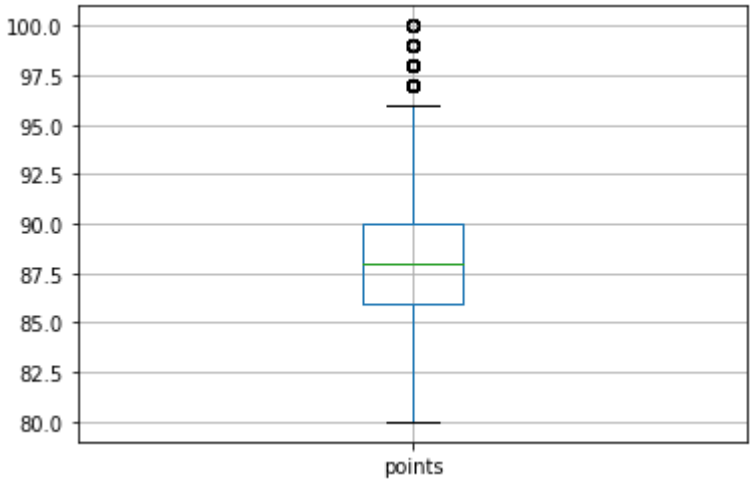
```
print(data_15k[attri].isnull().sum(axis=0))
```

```
0
```

绘制盒图

In [25]:

```
p = data_15k.boxplot([attri],return_type='dict')
```



检查离群点

In [26]:

```
print(p['fliers'][0].get_ydata())
print("MIN: ",end="")
print(min(p['fliers'][0].get_ydata()))
```

```
[100  99  98  98  98  97  97  97  97  97  97  97  97  97  97  97  97  97
  97  97  97  97  97  98  97  98  97  97  97  97  97  97  98  97  97  97
  97  97  97  98  97  97  97  97  97  97  97  97 100 100  99  99  98  98
  98  98  98  98  97  97  97  97  97  97  98  98  98  98  97  97  97  97
  97  97  97  98  97  97  97  97 100  99  99  98  98  98  98  98  98  97
  97  97  97 100  99  98  97  97  97  97  97  97  97  97  97  99  97  98
  97  97  97  97  97  97 100  98  98  97  97  97  97  97  97  97  97  97
  97  97  97  97  97  97  97  97  97  99  97  97  99  99  99  98  98  98
  98  98  97  97  97  97  97  97  99  97  97  97  97  97  97  97  97  97
 100  99  99  98  98  98  98  98  98  97  97  97  97  97  99  99  98  97
  97  97  97  97  98  98  97  97  97  97  97 100  99  98  97  97  97  97
  97  97  98  97  97  97  97  97  97  97  99  99  99  98  98  98  98  97
  97  97  97  97  97  97  97  97  97  97  97  97  98  97  97  97  98  98
  97  98  99  98  98  97  97  97  97  97  99  98  97  97  97  97  98  97
  97  97  97  97  97  97  97  97  97  97 100  98  98  97  97  97  97  97
  97  97  97  97  97  97  97  99  99  99  98  98  97  97  97  97  97  97
 100  99  98  97  97  97  97  97  97 100 100  99  99  98  98  98  98  98
  98  97  97  97  97  97  98  98  97  97  97  97  97 100  98  97  97  97
  97  99  99  98  97  97  97  98  97 100  98  97  97  97  97  97  97  97
  97  97  97  98  97  97  97  97  98  97  97  97 100  98  98  97  97  97
  97  97  98  97  97  97  99  98  97  97  98  98  98  98  97  97  98  98
  97  97  97  97  98  97  97  97  97  99  99  99  98  98  98  98  97  97
  97  97  97  97  97  97  98  97  97  97  97  97  97  98  97  97  97  97
  97  97  98 100  97  97  97  99  98  97  97 100  99  98  98  97  97  97
  97  97  97  97  99  98  98  97  97  97  97  97 100 100  99  99  98  98
  98  98  98  98  97  97  97  97  97  97 100  99  99  98  98  98  98  98
  98  97  97  97  97  98  97  97  97  99  98  98  97  97  97  97  97 100
  98  98  97  97  97  97  98  97  99  97  97  97  97  97  97  97  97  97
  99  98  97  97  98  97  97  97  97  99  99  99  98  98  98  98  97  97
  97  97  97  97  97  97  97  97  97  97  97  99  98  97  97  97 100  98
  97  97  97  97 100  97  98  98  97  97  97  97  99  98  98  98  97 100
  99  98  98  97  97  97  97  97  97  97  97  97]
MIN: 97
```

因此，points中大于等于97的项被识别为离群点。

3.1.9 price属性

数值属性，给出五数概括

In [27]:

```
attri = "price"
for i in range(0,5):
    q = i * 0.25
    print("Q%d:"%(i),end=" ")
    print(data_15k[attri].quantile(q))
```

```
Q0: 4.0
Q1: 16.0
Q2: 24.0
Q3: 40.0
Q4: 2300.0
```

给出缺失值个数

In [28]:

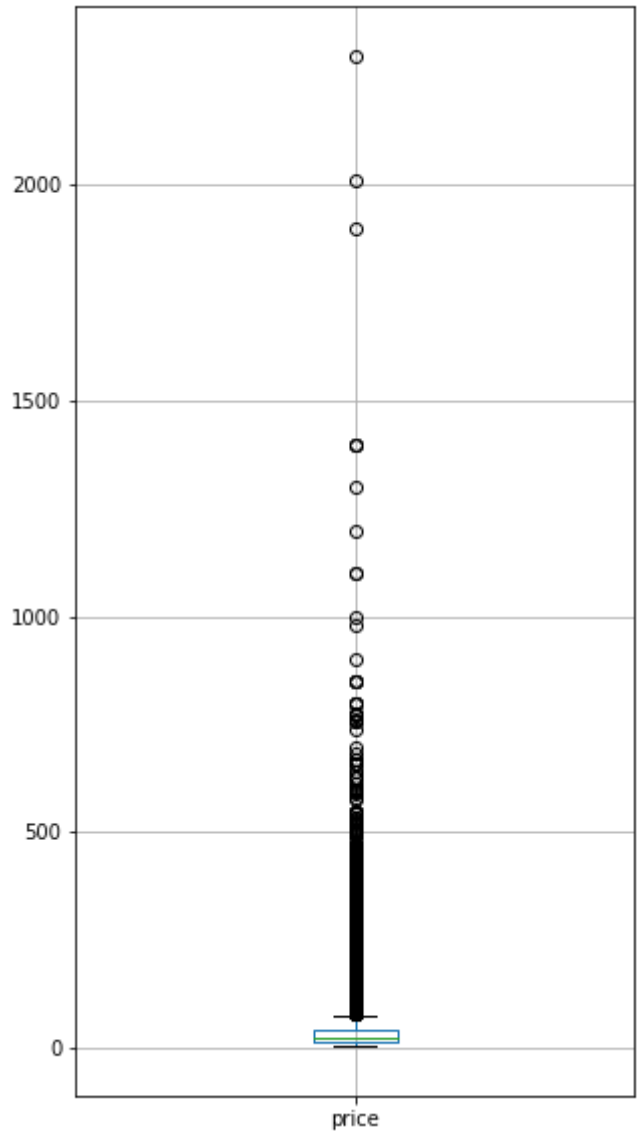
```
print(data_15k[attri].isnull().sum(axis=0))
```

```
13695
```

绘制盒图

In [29]:

```
p = data_15k.boxplot([attri],return_type='dict',figsize=(5,10))
```



检查离群点

In [30]:

```
print(p['fliers'][0].get_ydata())
print("MIN: ",end="")
print(min(p['fliers'][0].get_ydata()))
```

```
[235. 110.  90. ...  83. 100.  87.]
MIN: 77.0
```

因此，price中大于等于77的项被识别为离群点。

3.2 处理数据缺失

首先统计所有属性的缺失值

In [31]:

```
print(data_15k.isnull().sum(axis=0))
```

```
Unnamed: 0         0
country            5
description        0
designation    45735
points             0
price          13695
province           5
region_1       25060
region_2       89977
variety            0
winery             0
dtype: int64
```

3.2.1 处理country属性缺失

可能原因是人为失误，通过属性的相关关系来填补缺失值，使用designation的属性来判断所属国家

根据空值的分布，定义一个从designation到country的转换字典

In [32]:

```
attri = "country"
designation2country = {
    "Askitikos":"Greece",
    "Shah":"US",
    "Piedra Feliz":"Chile",
}
```

处理缺失

In [33]:

```
data_15k_new = data_15k.iloc[:,:]
for i in range(0,len(data_15k_new)):
    tmp = data_15k_new.iloc[i,1]
    if pd.isnull(tmp):
        designation = data_15k_new.iloc[i,3]
        data_15k_new.iloc[i,1] = designation2country[designation]
data_15k_new[attri].value_counts(dropna = False)
```

Out[33]:

```
US                        62398
Italy                     23478
France                    21098
Spain                      8268
Chile                      5819
Argentina                  5631
Portugal                   5322
Australia                  4957
New Zealand                3320
Austria                    3057
Germany                    2452
South Africa               2258
Greece                      885
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
Lithuania                     8
India                         8
Czech Republic                6
Ukraine                       5
South Korea                   4
Bosnia and Herzegovina        4
Switzerland                   4
Slovakia                      3
China                         3
Egypt                         3
Tunisia                       2
Montenegro                    2
Albania                       2
Japan                         2
US-France                     1
Name: country, dtype: int64
```

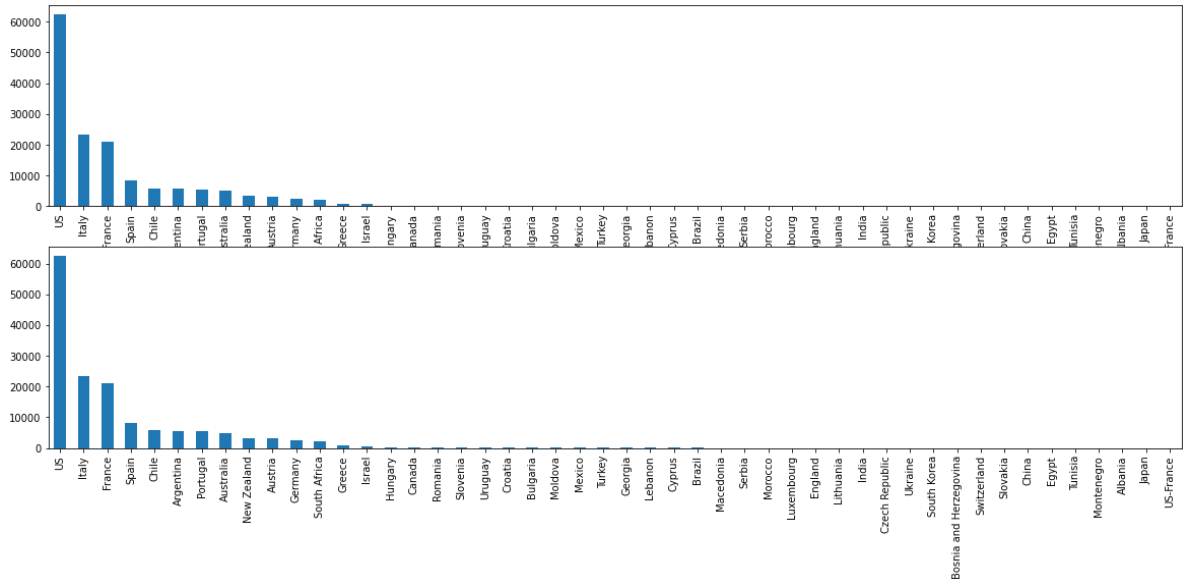
可视化对比

In [*]:

```
attri = "country"
matplotlib.pyplot.subplot(2,1,1)
data_15k[attri].value_counts(dropna = False).plot(kind='bar',figsize=(20,8))
matplotlib.pyplot.subplot(2,1,2)
data_15k_new[attri].value_counts(dropna = False).plot(kind='bar',figsize=(20,8))
```

Out[34]:

<AxesSubplot:>



3.2.2处理designation属性缺失

将缺失部分剔除

In [29]:

```
attri = "designation"
d = data_15k.dropna(subset=[attri])
d
```

Out[29]:

	Unnamed: 0	country	description	designation	points	price	province	region_1
0	0	US	This tremendous 100% varietal wine hails from ...	Martha's Vineyard	96	235.0	California	Napa Valley
1	1	Spain	Ripe aromas of fig, blackberry and cassis are ...	Carodorum Selección Especial Reserva	96	110.0	Northern Spain	Toro
2	2	US	Mac Watson honors the memory of a wine once ma...	Special Selected Late Harvest	96	90.0	California	Knights Valley
3	3	US	This spent 20 months in 30% new French oak, an...	Reserve	96	65.0	Oregon	Willamette Valley
4	4	France	This is the top wine from La Bégude, named aft...	La Brûlade	95	66.0	Provence	Bandol
...	...	...	...	...	...	...	...	...
150923	150923	France	Rich and toasty, with tiny bubbles. The bouque...	Demi-Sec	91	30.0	Champagne	Champagne
150924	150924	France	Really fine for a low-acid vintage, there's an...	Diamant Bleu	91	70.0	Champagne	Champagne
150926	150926	France	Offers an intriguing nose with ginger, lime an...	Cuvée Prestige	91	27.0	Champagne	Champagne
150927	150927	Italy	This classic example comes from a cru vineyard...	Terre di Dora	91	20.0	Southern Italy	Fiano di Avellino

	Unnamed: 0	country	description	designation	points	price	province	region_1
150928	150928	France	A perfect salmon shade, with scents of peaches...	Grand Brut Rosé	90	52.0	Champagne	Champagne

105195 rows × 11 columns

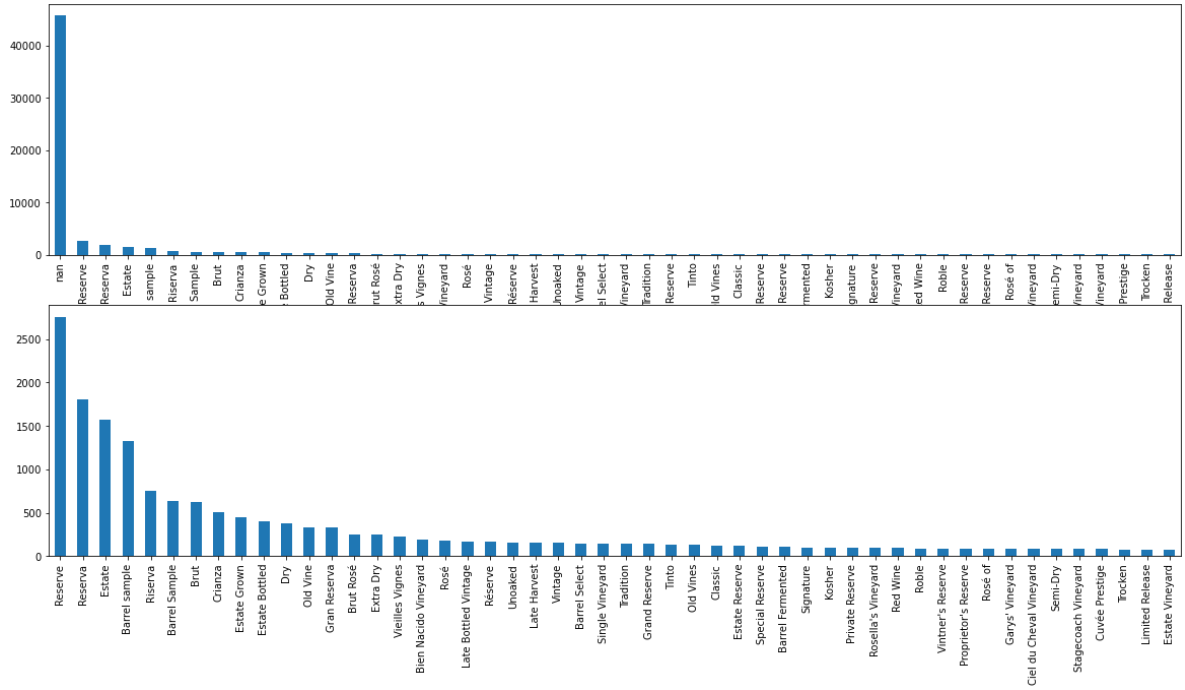
可视化对比，直方图中只显示前50项

In [30]:

```
attri = "designation"
matplotlib.pyplot.subplot(2,1,1)
data_15k[attri].value_counts(dropna = False)[:50].plot(kind='bar',figsize=(20,10))
matplotlib.pyplot.subplot(2,1,2)
d[attri].value_counts(dropna = False)[:50].plot(kind='bar',figsize=(20,10))
```

Out[30]:

<matplotlib.axes._subplots.AxesSubplot at 0x13c65ca0>



3.2.3 处理price属性缺失

可能原因是该种葡萄酒的价格无法获取,用最高频率值来填补缺失值

In [31]:

```
attri = "price"
mode = data_15k[attri].mode()
f = data_15k[attri].fillna(int(mode))
print(f)
```

```
0         235.0
1         110.0
2          90.0
3          65.0
4          66.0
          ...
150925     20.0
150926     27.0
150927     20.0
150928     52.0
150929     15.0
Name: price, Length: 150930, dtype: float64
```

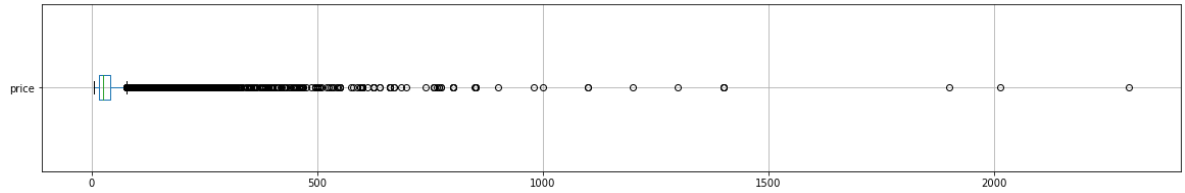
可视化对比

In [32]:

```
data_15k.boxplot([attri],vert=False,figsize=(20,3))
```

Out[32]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x13fa3298>
```

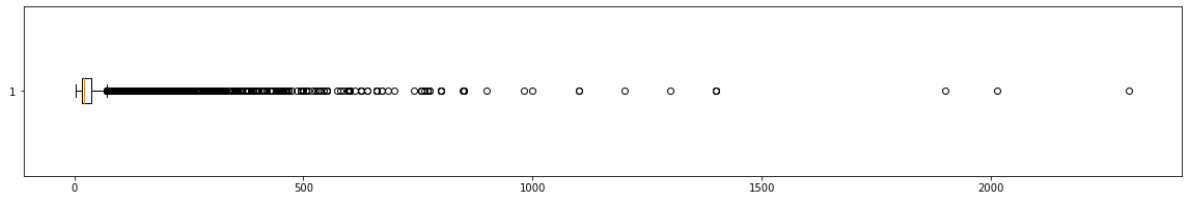


In [33]:

```
matplotlib.pyplot.figure(figsize=(20,3))
matplotlib.pyplot.boxplot(f,vert=False)
```

Out[33]:

```
{'whiskers': [<matplotlib.lines.Line2D at 0x1506c0b8>,
  <matplotlib.lines.Line2D at 0x1506c268>],
 'caps': [<matplotlib.lines.Line2D at 0x1506c418>,
  <matplotlib.lines.Line2D at 0x1506c5c8>],
 'boxes': [<matplotlib.lines.Line2D at 0x1505fee0>],
 'medians': [<matplotlib.lines.Line2D at 0x1506c778>],
 'fliers': [<matplotlib.lines.Line2D at 0x1506c8f8>],
 'means': []}
```



3.2.4 处理region_1属性缺失

可能原因是在数据收集阶段无法获取到region_1,用最高频率值来填补缺失值

In [34]:

```
attri = "region_1"
mode = data_15k[attri].mode()
f = data_15k[attri].fillna(str(mode))
print(f)
```

```
0               Napa Valley
1                     Toro
2            Knights Valley
3         Willamette Valley
4                   Bandol
                 ...
150925    Fiano di Avellino
150926            Champagne
150927    Fiano di Avellino
150928            Champagne
150929           Alto Adige
Name: region_1, Length: 150930, dtype: object
```

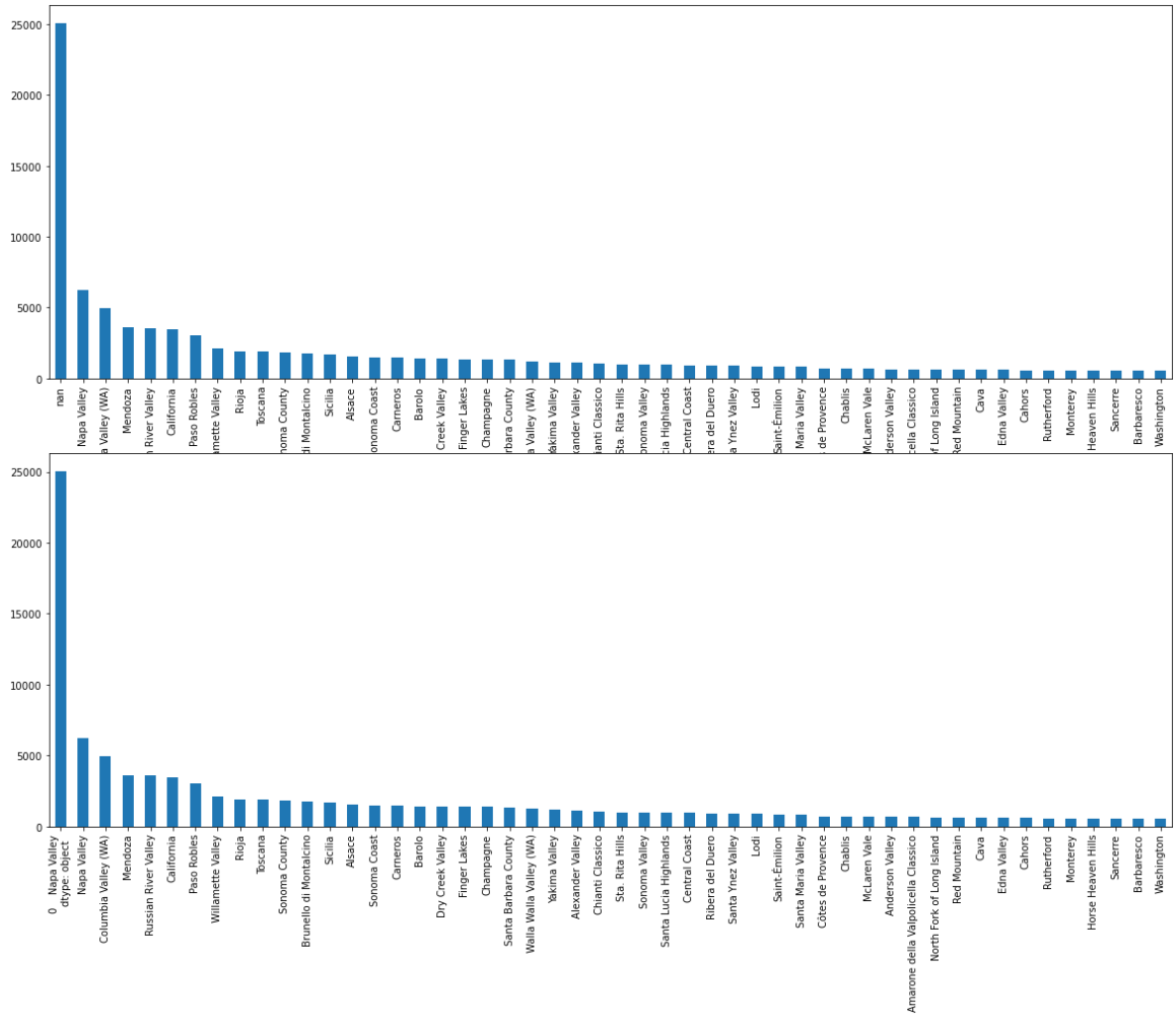
可视化对比，直方图只显示前50项

In [35]:

```
attri = "region_1"
matplotlib.pyplot.subplot(2,1,1)
data_15k[attri].value_counts(dropna = False)[:50].plot(kind='bar',figsize=(20,15))
matplotlib.pyplot.subplot(2,1,2)
f.value_counts(dropna = False)[:50].plot(kind='bar',figsize=(20,15))
```

Out[35]:

<matplotlib.axes._subplots.AxesSubplot at 0x1509f190>



3.2.5 处理region_2属性缺失

可能原因是部分数据不存在region_2的属性

这里，将缺失部分剔除

In [36]:

```
attri = "region_2"
new_region_2 = data_15k.dropna(subset=[attri])
print(new_region_2[attri].value_counts(dropna = False))
new_region_2
```

```
Central Coast              13057
Sonoma                     11258
Columbia Valley             9157
Napa                        8801
California Other            3516
Willamette Valley           3181
Mendocino/Lake Counties     2389
Sierra Foothills            1660
Napa-Sonoma                 1645
Finger Lakes                1510
Central Valley              1115
Long Island                  771
Southern Oregon              662
Oregon Other                 661
North Coast                  632
Washington Other             593
South Coast                  198
New York Other               147
Name: region_2, dtype: int64
```

Out[36]:

	Unnamed: 0	country	description	designation	points	price	province	region_1
0	0	US	This tremendous 100% varietal wine hails from ...	Martha's Vineyard	96	235.0	California	Napa Valley
2	2	US	Mac Watson honors the memory of a wine once ma...	Special Selected Late Harvest	96	90.0	California	Knights Valley
3	3	US	This spent 20 months in 30% new French oak, an...	Reserve	96	65.0	Oregon	Willamette Valley
8	8	US	This re-named vineyard was formerly bottled as...	Silice	95	65.0	Oregon	Chehalem Mountains
9	9	US	The producer sources from two blocks of the vi...	Gap's Crown Vineyard	95	60.0	California	Sonoma Coast
...	...	...	...	...	...	...	...	...

	Unnamed: 0	country	description	designation	points	price	province	region_1	
150892	150892	US	A light, earthy wine, with violet, berry and t...	Coastal	82	10.0	California	California	Calif
150896	150896	US	Some raspberry fruit in the aroma, but things ...	NaN	82	10.0	California	California	Calif
150914	150914	US	Old-gold in color, and thick and syrupy. The a...	Late Harvest Cluster Select	94	25.0	California	Anderson Valley	Mend
150915	150915	US	Decades ago, Beringer’s then-winemaker Myron N...	Nightingale	93	30.0	California	North Coast	N
150916	150916	US	An impressive wine that presents a full bouque...	J. Schram	93	65.0	California	Napa Valley	

60953 rows × 11 columns

可视化对比

In [37]:

```
attri = "region_2"
matplotlib.pyplot.subplot(2,1,1)
data_15k[attri].value_counts(dropna = False).plot(kind='bar',figsize=(20,10))
matplotlib.pyplot.subplot(2,1,2)
new_region_2[attri].value_counts(dropna = False).plot(kind='bar',figsize=(20,10))
```

Out[37]:

<matplotlib.axes._subplots.AxesSubplot at 0x137a7d30>