互评作业1：数据探索性分析与数据预处理：wine-reviews

学号：1120183560 姓名：刘文楷

1. 数据集：wine-reviews

两个csv文件:

- winemag-data_first150k.csv 10列15万行评论
- winemag-data-130k-v2.csv 10列13万行评论

读取数据:

In [1]:

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

In [2]:

```
dirpath_150k = "wine-data/winemag-data_first
dirpath_130k = "wine-data/winemag-data-130k-
data_150k = pd.read_csv(dirpath_150k)
data_130k = pd.read_csv(dirpath_130k)
```

数据属性:

In [3]:

```
# 数据类型
data_150k.dtypes
```

```
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
```

In [4]:

```
data_150k.head(10)
```

	Unnamed: 0	country	description	desig
0	0	US	This tremendous 100% varietal wine hails from ...	Marth Vineya
1	1	Spain	Ripe aromas of fig, blackberry and cassis are ...	Carod Selecc Especi Reserv
2	2	US	Mac Watson honors the memory of a wine once ma...	Specia Select Harve:
3	3	US	This spent 20 months in 30% new French oak, an...	Reserv
4	4	France	This is the top wine from La Bégude, named aft...	La Brû
5	5	Spain	Deep, dense and pure from the opening bell, th...	Numa
6	6	Spain	Slightly gritty black-fruit aromas include a s...	San Rc

	Unnamed: 0	country	description	desig
7	7	Spain	Lush cedary black-fruit aromas are luxe and of...	Carod Único Crianz
8	8	US	This re-named vineyard was formerly bottled as...	Silice
9	9	US	The producer sources from two blocks	Gap's Vineya

2. 数据分析

2.1 数据可视化和摘要

2.1.1 country属性

标称属性，后面所有的属性都和country一样

In [5]:

```
attribute = "country"
d150kvc = data_150k[attribute].value_counts(
d150kvc
```

```
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
England                       9
Luxembourg                    9
Lithuania                     8
India                         8
```

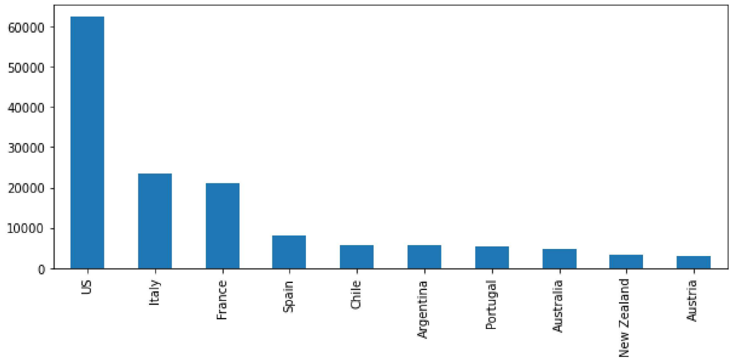
```
Czech Republic                6
NaN                           5
Ukraine                       5
Bosnia and Herzegovina        4
Switzerland                   4
South Korea                   4
Egypt                         3
Slovakia                      3
China                         3
Albania                       2
Tunisia                       2
Japan                         2
Montenegro                    2
US-France                     1
Name: country, dtype: int64
```

可视化

In [6]:

```
# 仅显示前10个，数据太大画不了
d150kvc[:10].plot(kind = "bar", figsize = (1
```

```
<AxesSubplot:>
```



2.1.2 designation属性

同country

In [8]:

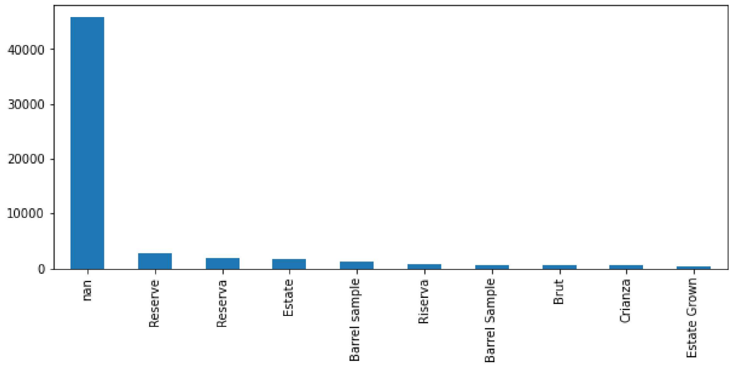
```
attribute = "designation"
d150kvc = data_150k[attribute].value_counts(
d150kvc
```

```
NaN                             45735
Reserve                          2752
Reserva                          1810
Estate                           1571
Barrel sample                    1326
                                 ...
Mostly                              1
Clos des Rocs Monopole              1
Eternity Sparkling Cuvée            1
Gueta-Lupia                         1
Ramona Pinot Noir                   1
Name: designation, Length: 30622, dtyp
e: int64
```

In [9]:

```
# 仅显示前10个，数据太大画不了
d150kvc[:10].plot(kind = "bar", figsize = (1
```

```
<AxesSubplot:>
```



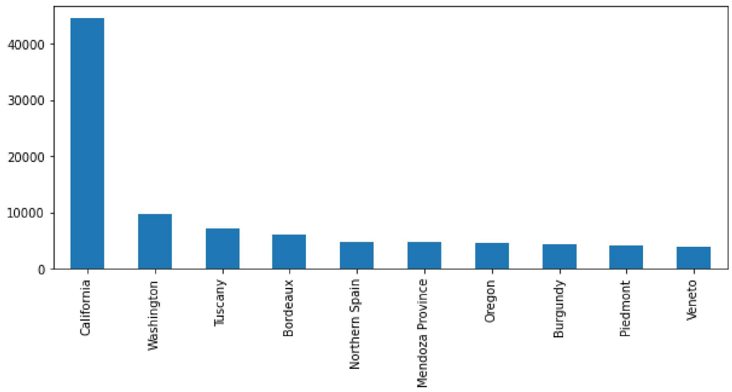
2.1.3 province属性

In [10]:

```
attribute = "province"
d150kvc = data_150k[attribute].value_counts(
print(d150kvc)
# 仅显示前10个，数据太大画不了
d150kvc[:10].plot(kind = "bar", figsize = (1
```

```
California       44508
Washington        9750
Tuscany           7281
Bordeaux          6111
Northern Spain    4892
                  ...
Pannon               1
Beni M'Tir           1
Stirling             1
Nevada               1
Rose Valley          1
Name: province, Length: 456, dtype: in
t64
```

```
<AxesSubplot:>
```



2.1.4 region_1

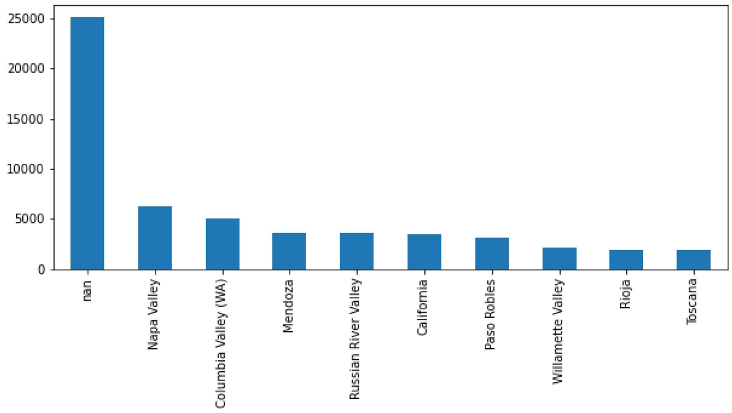
In [11]:

```
attribute = "region_1"
d150kvc = data_150k[attribute].value_counts(
print(d150kvc)
# 仅显示前10个，数据太大画不了
d150kvc[:10].plot(kind = "bar", figsize = (1
```

```
NaN
25060
Napa Valley
6209
Columbia Valley (WA)
4975
Mendoza
3586
Russian River Valley
3571

...
Vin de Pays des Coteaux de Murviel
1
Sonoma County-Monterey County
1
Fara
1
Central Valley
1
Geelong
1
Name: region_1, Length: 1237, dtype: i
nt64
```

```
<AxesSubplot:>
```


25000
20000
15000
10000
5000
0
nan
Napa Valley
Columbia Valley (WA)
Mendoza
Russian River Valley
California
Paso Robles
Willamette Valley
Rioja
Toscana

2.1.4 region_2

In [12]:

```
attribute = "region_2"
d150kvc = data_150k[attribute].value_counts(
print(d150kvc)
# 仅显示前10个，数据太大画不了
d150kvc[:10].plot(kind = "bar", figsize = (1
```

```
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
```

```
<AxesSubplot:>
```

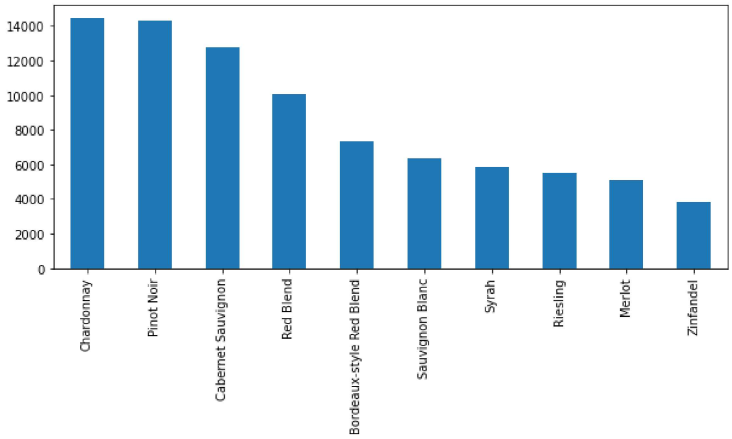
2.1.5 variety

In [13]:

```
attribute = "variety"
d150kvc = data_150k[attribute].value_counts(
print(d150kvc)
# 仅显示前10个，数据太大画不了
d150kvc[:10].plot(kind = "bar", figsize = (1
```

```
Chardonnay                  14482
Pinot Noir                  14291
Cabernet Sauvignon          12800
Red Blend                   10062
Bordeaux-style Red Blend     7347
                            ...
Pinela                          1
Silvaner-Traminer               1
Sideritis                       1
Merlot-Petite Verdot            1
Morava                          1
Name: variety, Length: 632, dtype: int
64
```

```
<AxesSubplot:>
```



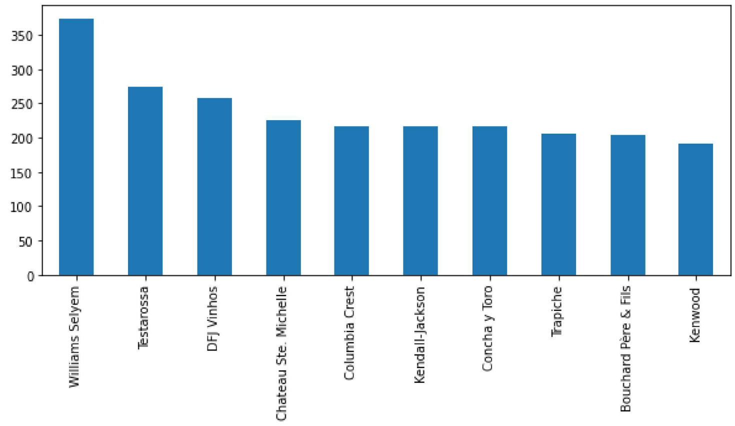
2.1.6 winery

In [14]:

```
attribute = "winery"
d150kvc = data_150k[attribute].value_counts(
print(d150kvc)
# 仅显示前10个，数据太大画不了
d150kvc[:10].plot(kind = "bar", figsize = (1
```

```
Williams Selyem         374
Testarossa              274
DFJ Vinhos              258
Chateau Ste. Michelle   225
Columbia Crest          217
                       ... 
Kilroy Was Here!          1
Bjorn                     1
La Greña                  1
Donna Anita               1
Au Contraire              1
Name: winery, Length: 14810, dtype: in
t64
```

```
<AxesSubplot:>
```



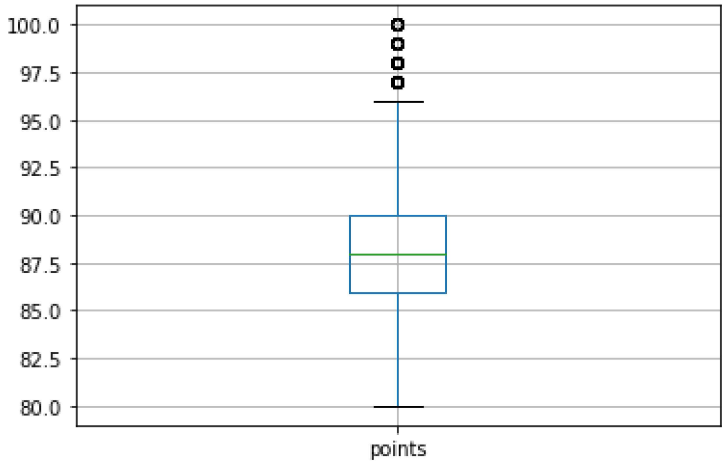
2.1.7 points

由于points为数值属性，所以给出5组数据概括，并绘制盒图

In [19]:

```
attribute = "points"
for i in range(0,5):
    print("Q:%d %.2f"%(i, data_150k[attribut
p = data_150k.boxplot([attribute],return_typ
```

```
Q:0 80.00
Q:1 86.00
Q:2 88.00
Q:3 90.00
Q:4 100.00
```



检查离散群点

In [20]:

```
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 9
7 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 9
7 97 97 97 97 97 100 98 98 97 97 97
 97 97 97 97 97 97 97 97 97 97 97 97 9
7 97 97 99 97 97 99 99 99 98 98 98
 98 98 97 97 97 97 97 97 99 97 97 97 9
7 97 97 97 97 97 100 99 99 98 98 98
 98 98 98 97 97 97 97 97 99 99 98 97 9
7 97 97 97 98 98 97 97 97 97 97 100
 99 98 97 97 97 97 97 97 98 97 97 97 9
7 97 97 97 99 99 99 98 98 98 98 97
 97 97 97 97 97 97 97 97 97 97 97 97 9
8 97 97 97 98 98 97 98 99 98 98 97
 97 97 97 97 99 98 97 97 97 97 98 97 9
7 97 97 97 97 97 97 97 97 97 100 98
 98 97 97 97 97 97 97 97 97 97 97 97 9
7 99 99 99 98 98 97 97 97 97 97 97
 100 99 98 97 97 97 97 97 97 100 100 9
9 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 9
7 97 97 98 97 97 97 100 98 98 97 97
 97 97 97 98 97 97 97 99 98 97 97 98 9
8 98 98 97 97 98 98 97 97 97 97 98
 97 97 97 97 99 99 99 98 98 98 98 97 9
7 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 9
7 100 100 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 9
9 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
```

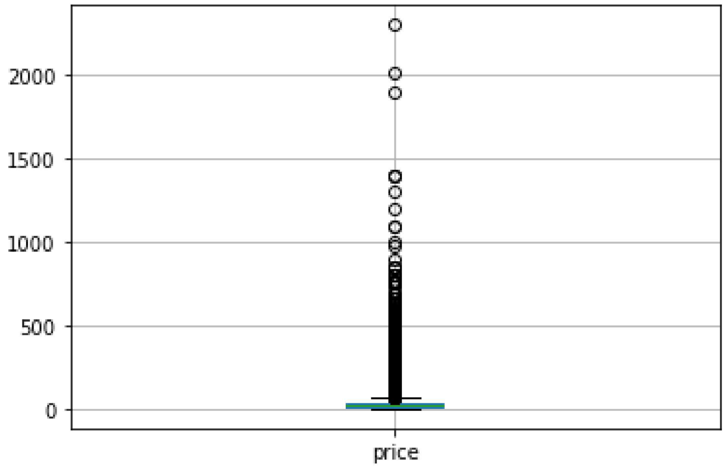
2.1.8 price

同points

In [21]:

```
attribute = "price"
for i in range(0,5):
    print("Q:%d %.2f"%(i, data_150k[attribut
p = data_150k.boxplot([attribute],return_typ
```

```
Q:0 4.00
Q:1 16.00
Q:2 24.00
Q:3 40.00
Q:4 2300.00
```



In [22]:

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

```
[235.0 110.0 90.0 ... 83.0 100.0 87.0]
MIN: 77.0
```

综上，price中大于等于77的项被识别为离群点。

2.2 处理数据缺失

统计所有数据的缺失值

In [23]:

```
print(data_150k.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
```

2.2.1 处理country属性缺失

原因：可能为人为因素，我们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 [24]:

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

In [25]:

```
data_150k_new = data_150k.iloc[:,:]
for i in range(0,len(data_150k_new)):
    tmp = data_150k_new.iloc[i,1]
    if pd.isnull(tmp):
        designation = data_150k_new.iloc[i,3
        data_150k_new.iloc[i,1] = designatio
data_150k_new[attribute].value_counts(dropna
```

```
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
```

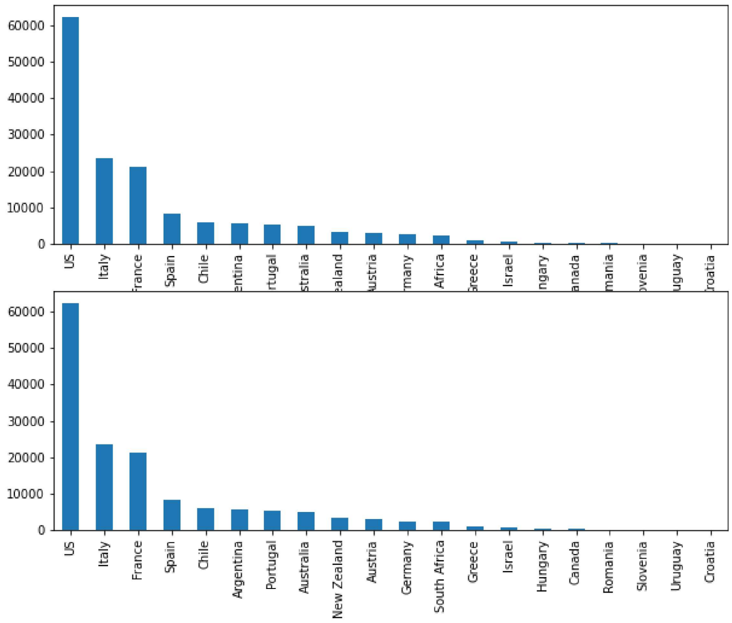
```
Macedonia                 16
Serbia                    14
Morocco                   12
England                    9
Luxembourg                 9
India                      8
Lithuania                  8
Czech Republic             6
Ukraine                    5
South Korea                4
Bosnia and Herzegovina     4
Switzerland                4
Egypt                      3
Slovakia                   3
China                      3
Albania                    2
Tunisia                    2
Japan                      2
Montenegro                 2
US-France                  1
Name: country, dtype: int64
```

可视化对比

In [26]:

```
# 考虑到数据太大，我们这里只取前20列
attribute = "country"
matplotlib.pyplot.subplot(2,1,1)
data_150k[attribute].value_counts(dropna = F
matplotlib.pyplot.subplot(2,1,2)
data_150k_new[attribute].value_counts(dropna
```

```
<AxesSubplot:>
```



2.2.2 designation

原因：同country，可能是人为因素

处理方法：此处我们选择将缺失部分剔除

In [27]:

```
attribute = "designation"
data_150k.dropna(subset=[attribute])
```

	Unnamed: 0	country	description	de
0	0	US	This tremendous 100% varietal wine hails from ...	Ma Vir
1	1	Spain	Ripe aromas of fig, blackberry and cassis are ...	Ca Sel Esp Re
2	2	US	Mac Watson honors the memory of a wine once ma...	Sp Sel Ha
3	3	US	This spent 20 months in 30% new French oak, an...	Re
4	4	France	This is the top wine from La Bégude, named aft...	La
...	...	...	...	...
150923	150923	France	Rich and toasty, with tiny bubbles. The bouque...	De

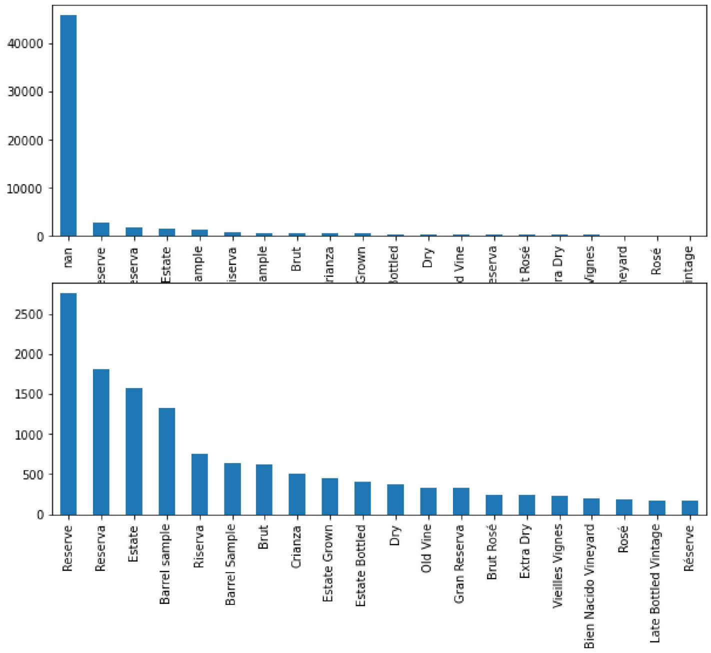
	Unnamed: 0	country	description	de
150924	150924	France	Really fine for a low-acid vintage, there's an...	Dia
150926	150926	France	Offers an intriguing nose with ginger, lime an...	Cu Pre
150927	150927	Italy	This classic example comes from a cru vineyard...	Ter
150928	150928	France	A perfect salmon shade, with scents of peaches...	Gra Ros

```
105195 rows × 11 columns
```

In [31]:

```
matplotlib.pyplot.subplot(2,1,1)
data_150k[attribute].value_counts(dropna = F
matplotlib.pyplot.subplot(2,1,2)
d150 = data_150k.dropna(subset=[attribute])
d150[attribute].value_counts(dropna = False)
```

```
<AxesSubplot:>
```



2.2.3 处理price

原因：葡萄酒价格没法获取

处理：用最高频率高来填补缺失值

In [32]:

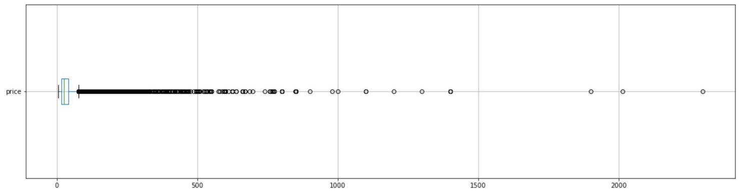
```
attribute = "price"
mode = data_150k[attribute].mode()
d150f = data_150k[attribute].fillna(int(mode
d150f
```

```
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: fl
oat64
```

In [33]:

```
data_150k.boxplot([attribute],vert=False,fig
```

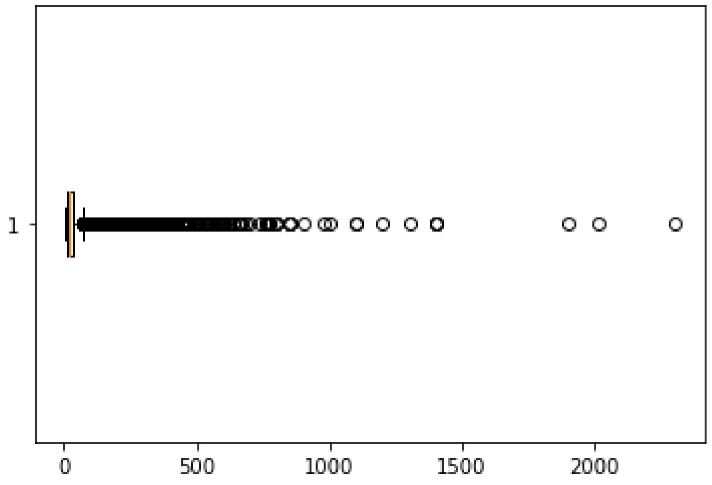
```
<AxesSubplot:>
```



In [36]:

```
matplotlib.pyplot.boxplot(d150f,vert=False)
```

```
{'whiskers': [<matplotlib.lines.Line2D
at 0x2d4227f8160>,
  <matplotlib.lines.Line2D at 0x2d4227
f8ac0>],
 'caps': [<matplotlib.lines.Line2D at
0x2d4227f88e0>,
  <matplotlib.lines.Line2D at 0x2d425b
b5f70>],
 'boxes': [<matplotlib.lines.Line2D at
0x2d4227c8370>],
 'medians': [<matplotlib.lines.Line2D
at 0x2d425bb5670>],
 'fliers': [<matplotlib.lines.Line2D a
t 0x2d425bbc550>],
 'means': []}
```



2.2.4 处理region_1

原因：同price，region_1无法获取

处理：用最高频率值来填补缺失值

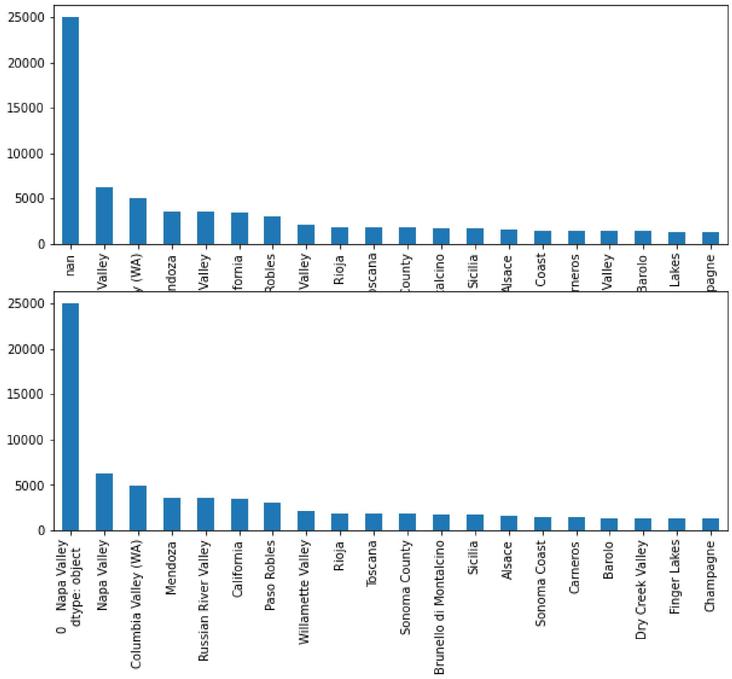
In [38]:

```
attribute = "region_1"
mode = data_150k[attribute].mode()
d150f = data_150k[attribute].fillna(str(mode
d150f
```

```
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
```

```
In [39]:
matplotlib.pyplot.subplot(2,1,1)
data_150k[attribute].value_counts(dropna = F
matplotlib.pyplot.subplot(2,1,2)
d150f.value_counts(dropna = False)[:20].plot
```

```
<AxesSubplot:>
```



2.4.5 处理region_2

原因：这部分根本就不存在region_2数据

处理：将这部分剔除

In [41]:

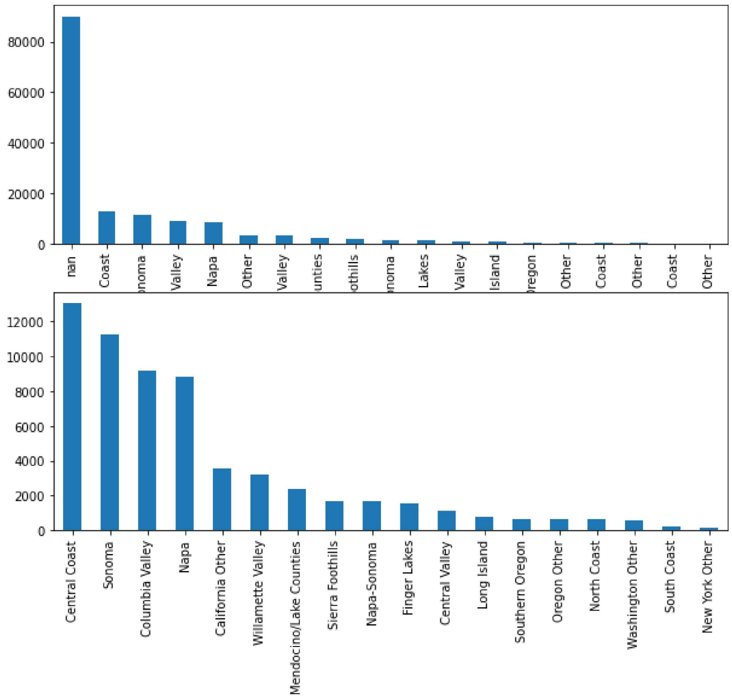
```
attribute = "region_2"
new_region_2 = data_150k.dropna(subset=[attr
new_region_2[attribute].value_counts(dropna
```

```
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
```

In [43]:

```
matplotlib.pyplot.subplot(2,1,1)
data_150k[attribute].value_counts(dropna = F
matplotlib.pyplot.subplot(2,1,2)
new_region_2[attribute].value_counts(dropna
```

```
<AxesSubplot:>
```



In [ ]: