

DataPreprocess-Wine Reviews

April 7, 2021

1 Wine Reviews

1.1 Wine Reviews
winemag-data_first150k10 countrydescriptiondesignationprovinceregionregion_2varietywinery pointsprice 7

```
In [22]: import csv
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import json
from math import sqrt
from statsmodels.formula.api import ols
plt.rcParams['font.sans-serif'] = ['SimHei'] #
# import statsmodels.api as sm
#
df = pd.read_csv('C:/Users/ZL/Desktop/winemag-data_first150k.csv')
#
label_num = ['points', 'price']
label_nom = ['country', 'description', 'designation',
            'province', 'region_1', 'region_2', 'variety', 'winery']
df_num = df[label_num]#
df_nom = df[label_nom]#

# -----
#
def frequency(label):
    f = df[label].value_counts()
    f = pd.DataFrame(f)
    return f

#
def OutNominal():
    for item in label_nom:
        print("{}:\n{}".format(item, frequency(item)))
OutNominal()
```

country:

country

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
Lithuania	8
India	8
Czech Republic	6
Ukraine	5
South Korea	4
Bosnia and Herzegovina	4
Switzerland	4
China	3
Slovakia	3
Egypt	3
Albania	2
Japan	2
Montenegro	2
Tunisia	2
US-France	1

description:

	description
Powerful in Zinny character, this blend of Dry ...	6
A little bit funky and unsettled when you pop t...	6
86-88 This could work as a rich wine, because t...	6
92-94 Barrel sample. A rounded wine, its tannin...	6
Sweet cherry and baking vanilla aromas are foll...	5
Gibilmo, a pure expression of Nero d'Avola, s...	5
This is the kind of inexpensive Cabernet you ca...	4
The bubbles really intensify the varietal chara...	4
Tageto is a Bordeaux blend from Coastal Tuscany...	4
Made from hand-picked grapes and unfiltered, th...	4
Dark and fully oaked Rioja with vanilla, mocha,...	4
This oak-aged blend of Trebbiano, Colombana and...	4
A blend from two well-regarded vineyards, this ...	4
A bright, berry-flavored wine, acidity showing ...	4
Ripe pear fruit flavors go with green plums in ...	4
Sweetly ripe and oaky, with pineapple jam, kiwi...	4
Fresh and fruity, with attractive pink coloring...	4
This Bolgheri blend of Cabernet Sauvignon (60%)...	4
Baia al Vento (windy bay) is a mostly Merlot-ba...	4
One-dimensional in fruit, but clean and zesty, ...	4
Well integrated in terms of barrel use (listed ...	4
From a fourth-generation farming family, this w...	4
This Chardonnay and Pinot Noir blend smells swe...	4
Lime zest and freshly cut grass lend a particul...	4
Tastes soft and almost sweet, with melted raspb...	4
A good everyday wine. It's dry, full-bodied and...	4
A finely structured wine, its tannin fitting we...	4
Soft and full in the mouth, with spice adding a...	4
A spicy, berry-dominated wine, rustic edging to...	4
Light on the nose, with dilute red-berry aromas...	4
...	...
Mint and milk chocolate are the lead aromas, fo...	1
For a five-spot you get syrupy but fairly deep ...	1
This is a very herbal wine, with a green medici...	1
Klopp is a mix of California and Burgundy clone...	1
Rich in blackberry, blueberry, plum, cocoa and ...	1
The nose begins with wheat grass and chopped ch...	1
A change in style from the soft, round version ...	1
Certified French organic with this vintage, thi...	1
Smooth, chewy tannins on the palate with ripe r...	1
This gorgeous Riserva Chardonnay shows modern o...	1
Full of red fruits, it also has a strong struct...	1
Founded in 1915, Donelli produces some of centr...	1
Intense, concentrated, this blend seems to brin...	1
A lovely wine to drink now. It's soft, dry, bal...	1
Vegetal, unripe, with gluey-sweet flavors.	1

Aged only in stainless steel, this is a food-fr...	1
Has an odd green aroma, like capers. On the pal...	1
Smells like a chocolate brownie, tastes dry and...	1
Smoky wood aromas, with pure black fruits are a...	1
This is tastes a little soft and sweet, with ra...	1
Mauritson has been building up a pretty good tr...	1
There's an intriguing tension to this wine, shi...	1
This steely white offers up very little aromas...	1
Gonzague Lurton, current president of the Marga...	1
Soft and rounded with ripe apple and pear flavo...	1
This ultrafresh wine is in a very different sty...	1
This 55% Cabernet Sauvignon, 45% Merlot wine is...	1
Very soft and ripe in black cherries, chocolate...	1
Father-and-son vintners Joe and Sam Miller cons...	1
Cloudy in the glass, this appellation blend sho...	1

[97821 rows x 1 columns]

designation:

	designation
Reserve	2752
Reserva	1810
Estate	1571
Barrel sample	1326
Riserva	754
Barrel Sample	639
Brut	624
Crianza	503
Estate Grown	449
Estate Bottled	396
Dry	374
Old Vine	331
Gran Reserva	330
Brut Rosé	248
Extra Dry	244
Vieilles Vignes	225
Bien Nacido Vineyard	195
Rosé	180
Late Bottled Vintage	171
Réserve	166
Late Harvest	161
Unoaked	161
Vintage	152
Barrel Select	145
Single Vineyard	144
Tradition	141
Grand Reserve	139
Tinto	128
Old Vines	127

Classic	123
...	...
Estate Damiana Vineyard	1
Egérie Extra Brut	1
Small Lot Collection Barrel-Fermented	1
Cuvée C.M.	1
Miser	1
Vivanco Viura-Malvasía	1
Rich Demi-Sec	1
Auld Alliance	1
T Crianza	1
Dedicación Personal	1
Serrocielo	1
Clos de Clos Genet	1
Premium Vecchia Modena	1
Betsy Vineyard	1
Cuvée César à Sumeire	1
Galí Evreshe	1
Gesture	1
1762	1
Tagus Creek Cabernet Sauvignon and Aragones	1
Vigna Schiena d'Asino	1
Dogtown Vineyard	1
Cava Chrisohoou	1
Semmonay	1
Serra del Conte	1
Lone Acre	1
Sol Duc	1
Reserva Vendimia Seleccionada	1
Tapteil Red Wine	1
Piesporter Michelsberg Auslese	1
Ca' L'Inverno	1

[30621 rows x 1 columns]

province:

	province
California	44508
Washington	9750
Tuscany	7281
Bordeaux	6111
Northern Spain	4892
Mendoza Province	4742
Oregon	4589
Burgundy	4308
Piedmont	4093
Veneto	3962
South Australia	3004
Sicily & Sardinia	2545

New York	2428
Northeastern Italy	1982
Loire Valley	1786
Alsace	1680
Marlborough	1655
Southwest France	1601
Central Italy	1530
Southern Italy	1439
Champagne	1370
Catalonia	1352
Rhône Valley	1318
Colchagua Valley	1201
Languedoc-Roussillon	1082
Douro	1075
Provence	1021
Port	903
Maipo Valley	895
Other	889
...	...
Serra do Sudeste	1
Pannon	1
Colchagua Costa	1
Waitaki Valley	1
San Clemente	1
Dalmatian Coast	1
Lemnos	1
Vino da Tavola della Svizzera Italiana	1
Stereia Ellada	1
Pafos	1
Casablanca-Curicó Valley	1
Viile Timis	1
Central Otago-Marlborough	1
Terasele Dunarii	1
Central Greece	1
Pocerina	1
Limnos	1
Dolenjska	1
Langenlois	1
Stirling	1
Viile Carasului	1
Morocco	1
Valais	1
Zitsa	1
Cape South Coast	1
Beni M'Tir	1
Malgas	1
Martinborough Terrace	1
Vale Trentino	1

Neuchâtel 1

[455 rows x 1 columns]

region_1:

	region_1
Napa Valley	6209
Columbia Valley (WA)	4975
Mendoza	3586
Russian River Valley	3571
California	3462
Paso Robles	3053
Willamette Valley	2096
Rioja	1893
Toscana	1885
Sonoma County	1853
Brunello di Montalcino	1746
Sicilia	1701
Alsace	1574
Sonoma Coast	1473
Carneros	1458
Barolo	1398
Dry Creek Valley	1398
Finger Lakes	1372
Champagne	1369
Santa Barbara County	1319
Walla Walla Valley (WA)	1225
Yakima Valley	1162
Alexander Valley	1139
Chianti Classico	1029
Sta. Rita Hills	983
Sonoma Valley	971
Santa Lucia Highlands	962
Central Coast	950
Ribera del Duero	899
Santa Ynez Valley	898
...	...
Cour-Cheverny	1
Chignin-Bergeron	1
Hautes Cotes de Beaune	1
Saint-Georges-Saint-Émilion	1
Alpilles	1
Monterey County-Napa County-Sonoma County	1
Galatina	1
Isle St. George	1
Malibu Coast	1
Vin de Pays de Côtes du Tarn	1
Côtes du Roussillon Villages Tautavel	1
Outer Coastal Plain	1

Vin de Pays de Caux	1
Asprinio di Aversa	1
Vin de Pays de Hauterive	1
Cérons	1
Grand Roussillon	1
Côtes de Montravel	1
Santa Barbara-Monterey	1
Clos de Lambrays	1
Vin de Pays de Sainte-Marie la Blanche	1
Mitterberg	1
Vino de Calidad de Valtiendas	1
Catamarca	1
Coteaux du Languedoc-Pézenas	1
Central Valley	1
McLaren Vale-Padthaway	1
El Pomar District	1
Lugana Superiore	1
Arribes del Duero	1

[1236 rows x 1 columns]

region_2:

	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

variety:

	variety
Chardonnay	14482
Pinot Noir	14291
Cabernet Sauvignon	12800
Red Blend	10062
Bordeaux-style Red Blend	7347
Sauvignon Blanc	6320

Syrah	5825
Riesling	5524
Merlot	5070
Zinfandel	3799
Sangiovese	3345
Malbec	3208
White Blend	2824
Rosé	2817
Tempranillo	2556
Nebbiolo	2241
Portuguese Red	2216
Sparkling Blend	2004
Shiraz	1970
Corvina, Rondinella, Molinara	1682
Rhône-style Red Blend	1505
Barbera	1365
Pinot Gris	1365
Cabernet Franc	1363
Sangiovese Grosso	1346
Pinot Grigio	1305
Viognier	1263
Bordeaux-style White Blend	1261
Champagne Blend	1238
Port	1058
...	...
Syrah-Bonarda	1
Altesse	1
Premsal	1
Aidani	1
Grenache Gris	1
Sacy	1
Chardonel	1
Macabeo-Gewürztraminer	1
Trousseau Gris	1
Azal	1
Cabernet Pfeffer	1
Vidadillo	1
Cabernet Franc-Malbec	1
Forcallà	1
Xynisteri	1
Groppello	1
Sangiovese Cabernet	1
Dafni	1
Catalanesca	1
Pinot Grigio-Chardonnay	1
Tinta Francisca	1
Blauburgunder	1
Pinela	1

Romorantin	1
Pinotage-Merlot	1
Mandilaria	1
Espadeiro	1
Garnacha Tintorera	1
Sarba	1
Terret Blanc	1

[632 rows x 1 columns]
winery:

	winery
Williams Selyem	374
Testarossa	274
DFJ Vinhos	258
Chateau Ste. Michelle	225
Columbia Crest	217
Kendall-Jackson	216
Concha y Toro	216
Trapiche	205
Bouchard Père & Fils	203
Kenwood	191
De Loach	189
Joseph Drouhin	189
Georges Duboeuf	188
Cameron Hughes	172
Wines & Winemakers	169
Albert Bichot	167
Robert Mondavi	166
Louis Latour	154
D'Arenberg	153
Morgan	153
Dry Creek Vineyard	153
Concannon	151
Martin Ray	149
Errazuriz	148
L'Ecole No. 41	144
Gary Farrell	144
Olivier Leflaive	143
Montes	142
Waterbrook	142
Iron Horse	142
...	...
Château la Caderie	1
Alpha Zeta	1
Doña Isadora	1
Gauthier	1
Domaine Gerald Talmard	1
Boyer	1

James Estate	1
Mark Moretti	1
Piliota	1
Château Reynats	1
Garcia Schwaderer	1
Insania	1
Fontana d'Italia	1
Limb	1
Costanza Malfatti	1
Bodega Rolland	1
Viberti Giovanni	1
Château la Fortune	1
Cataregia	1
Gridley	1
Château Vieux Peyrouquet	1
Vina Moda	1
Gran Cruz	1
Château Beaulieu Comtes de Tastes	1
Bodega Cecchin	1
Magpie Estate	1
Château Maréchaux	1
Ocaso	1
Paul Buisse	1
Barbanera	1

[14810 rows x 1 columns]

.max(),.min(),.mean(),.median(),.quantile().isnull().sum()

In [23]: #5

```
def Num5():
    for item in label_num:
        Minimum = df[item].min()
        Maximum = df[item].max()
        Q1 = df[item].quantile(0.25)
        Median = df[item].mean()
        Q3 = df[item].quantile(0.75)
        print("{}{}{}{}{}{}".format(item,Minimum,Q1,Median,Q3,Maximum))

#
def lostdata(nums,item):
    nulltotal = nums[item].isnull().sum()
    print("{}{}".format(item,nulltotal))

Num5()
for item in label_num :
    lostdata(df,item)
```

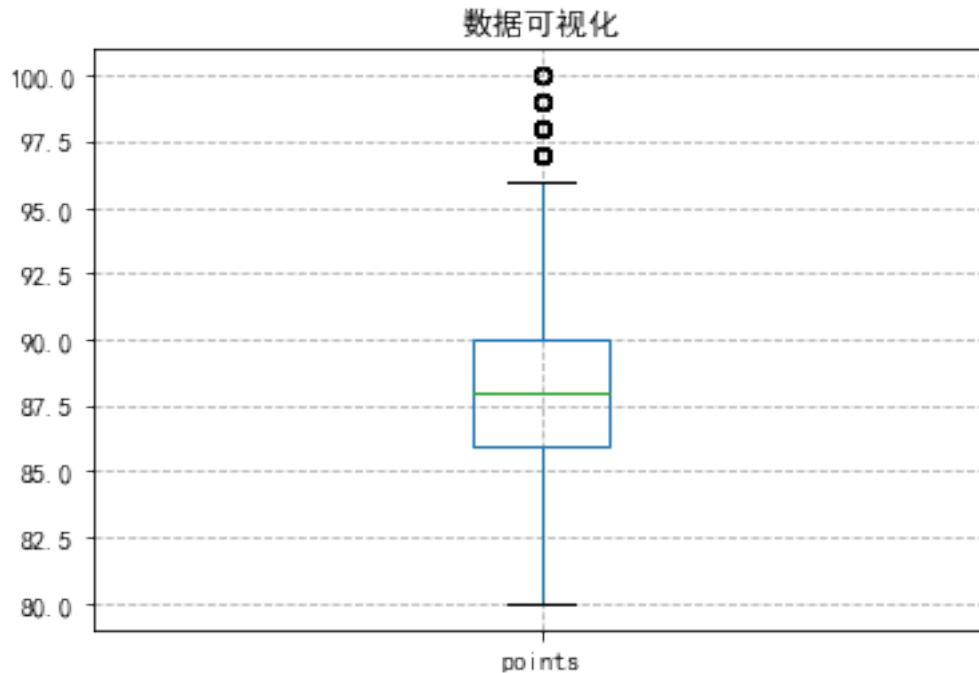
points8086.087.888418472139490.0100

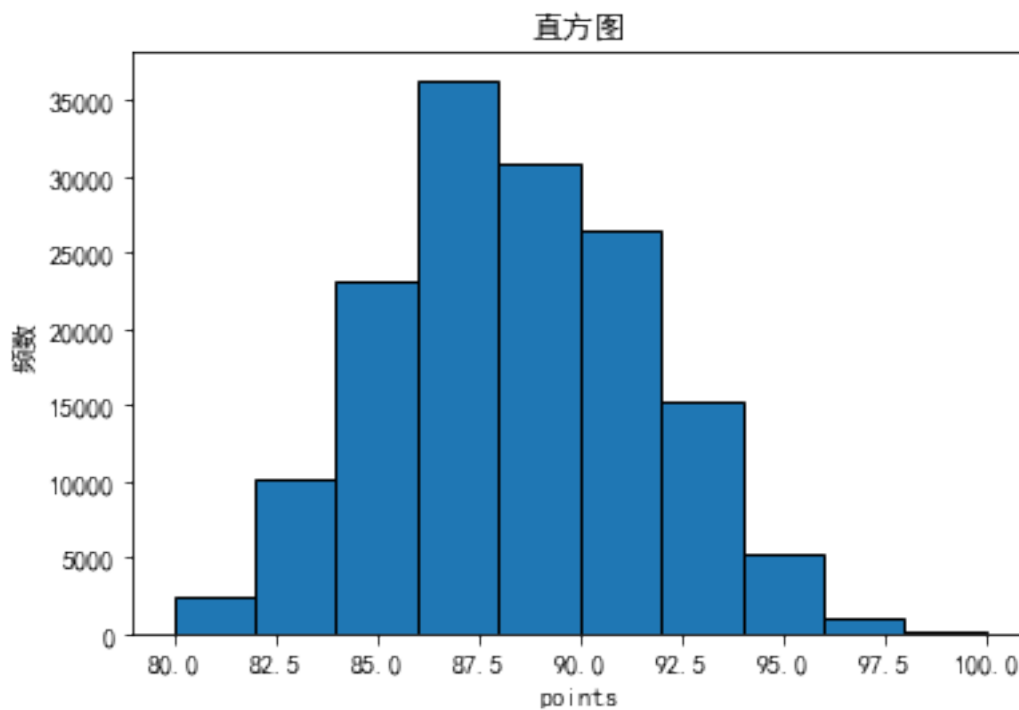
price4.016.033.1314824935329940.02300.0

```
points0  
price13695
```

pointsprice points80~90

```
In [14]: import pandas as pd  
import numpy as np  
from collections import Counter  
import matplotlib.pyplot as plt  
plt.rcParams['font.sans-serif'] = ['SimHei']  
plt.rcParams['axes.unicode_minus'] = False  
  
path = 'C:/Users/ZL/Desktop/winemag-data_first150k.csv'  
wine_data = pd.read_csv(path, header=0, index_col=0, engine='python', encoding='utf-8')  
wine_data['points'].plot.box(title="")  
plt.grid(linestyle="--")  
plt.show()  
plt.hist(x=wine_data['points'], bins=10, edgecolor='black')  
plt.xlabel('points')  
plt.ylabel('')  
plt.title('')  
plt.show()
```

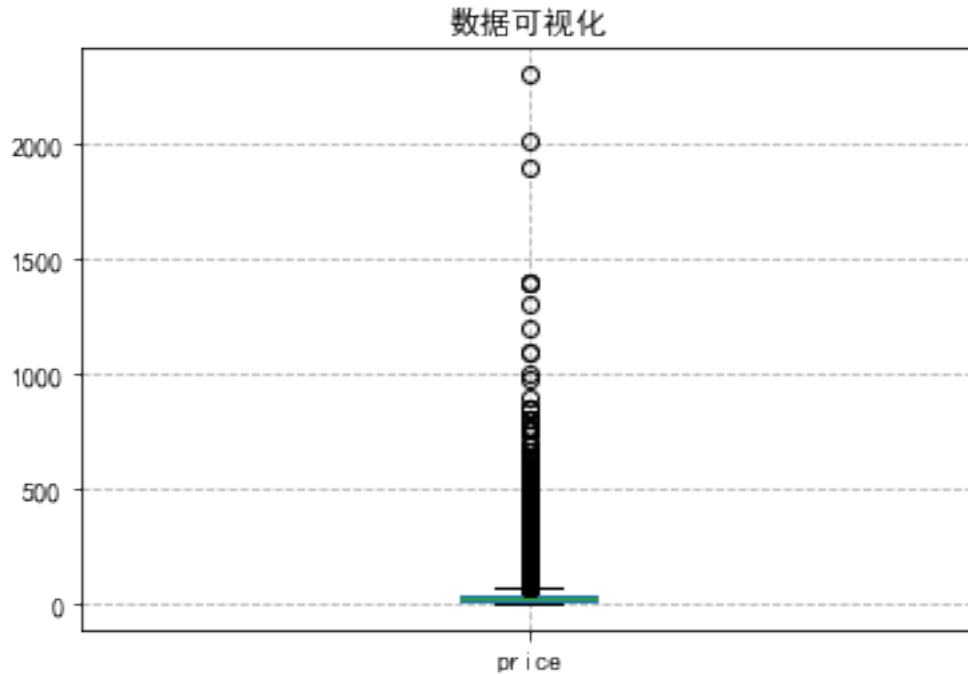




price

```
In [5]: import pandas as pd
import numpy as np
from collections import Counter
import matplotlib.pyplot as plt
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.rcParams['axes.unicode_minus'] = False

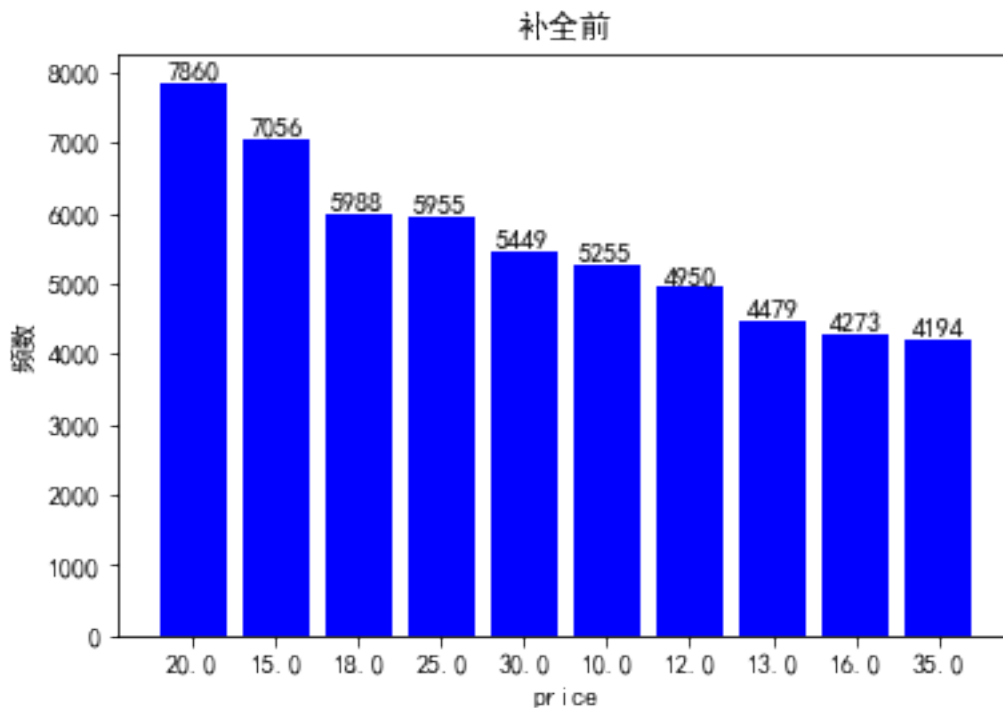
path = 'C:/Users/ZL/Desktop/winemag-data_first150k.csv'
wine_data = pd.read_csv(path, header=0, index_col=0, engine='python', encoding='utf-8')
wine_data['price'].plot.box(title="")
plt.grid(linestyle="--")
plt.show()
```



priceprice10

```
In [12]: # -*- coding: utf-8 -*-
import matplotlib.pyplot as plt

plt.rcParams['font.sans-serif'] = ['SimHei'] #
plt.rcParams['axes.unicode_minus'] = False #
fig, ax = plt.subplots()
num_list = [7860, 7056, 5988, 5955, 5449, 5255, 4950, 4479, 4273, 4194]
name_list = ['20.0', '15.0', '18.0', '25.0', '30.0', '10.0', '12.0', '13.0', '16.0',
b = ax.bar(name_list, num_list)
plt.bar(range(len(num_list)), num_list, color='blue', tick_label=name_list)
for a, b in zip(name_list, num_list):
    ax.text(a, b + 1, b, ha='center', va='bottom')
plt.title('')
plt.xlabel('price')
plt.ylabel('')
plt.show()
```



#2 #2.1 39241 price10

```
In [37]: import pandas as pd
import numpy as np
from collections import Counter
import matplotlib.pyplot as plt
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.rcParams['axes.unicode_minus'] = False
path = 'C:/Users/ZL/Desktop/winemag-data_first150k.csv'

def draw(data):
    num = 10
    wine_data = data.values
    counter = Counter(wine_data[:, 4])
    frequency = counter.most_common() # n
    num_list = []
    name_list = []
    for i in range(num):
        num_list.append(int(frequency[i][1]))
        name_list.append(str(frequency[i][0]))
    fig, ax = plt.subplots()
    b = ax.bar(name_list, num_list)
    plt.bar(range(len(num_list)), num_list, color='blue', tick_label=name_list)
    for a, b in zip(name_list, num_list):
        ax.text(a, b + 1, b, ha='center', va='bottom')
```

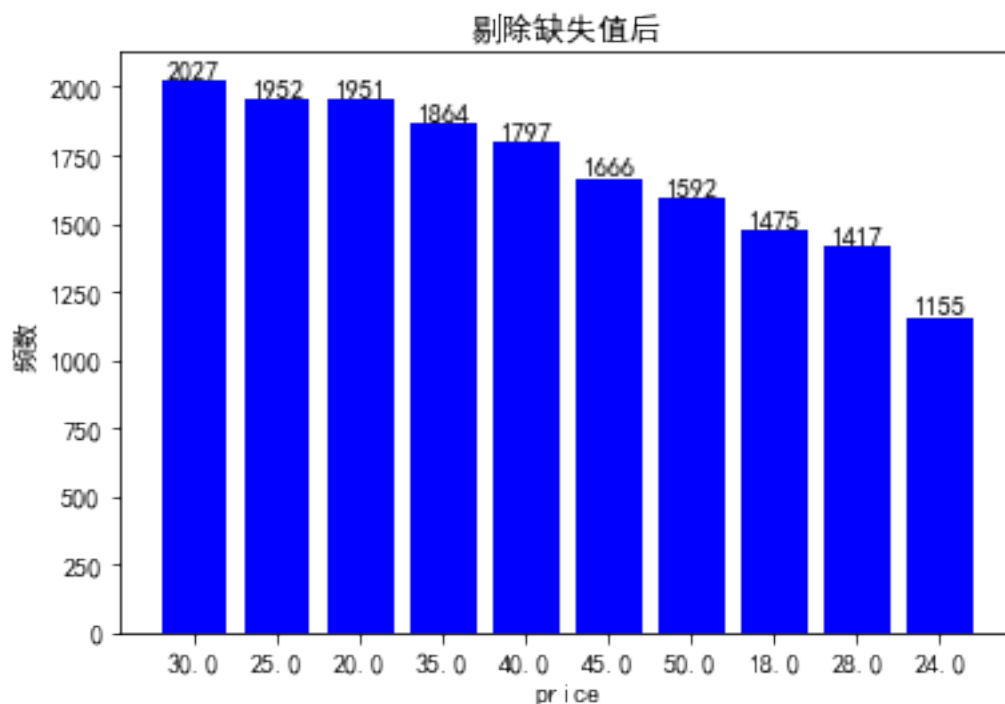
```

plt.title('')
plt.xlabel('price')
plt.ylabel('')
plt.show()

wine_data = pd.read_csv(path, header=0, index_col=0, engine='python', encoding='utf-8')
wine_drop = wine_data.dropna() #
print(''+str(wine_drop.shape[0]))
draw(wine_drop)

```

39241



##2.2 1010 pointsprice150930 price10price20

```

In [18]: import pandas as pd
import numpy as np
from collections import Counter
import matplotlib.pyplot as plt
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.rcParams['axes.unicode_minus'] = False
path = 'C:/Users/ZL/Desktop/winemag-data_first150k.csv'

def draw(data):
    num = 10

```



```

wine_data = data.values
counter = Counter(wine_data[:, 4])
frequency = counter.most_common() # n
num_list = []
name_list = []
for i in range(num):
    num_list.append(int(frequency[i][1]))
    name_list.append(str(frequency[i][0]))
fig, ax = plt.subplots()
b = ax.bar(name_list, num_list)
plt.bar(range(len(num_list)), num_list, color='blue', tick_label=name_list)
for a, b in zip(name_list, num_list):
    ax.text(a, b + 1, b, ha='center', va='bottom')
plt.title('')
plt.xlabel('price')
plt.ylabel('')
plt.show()

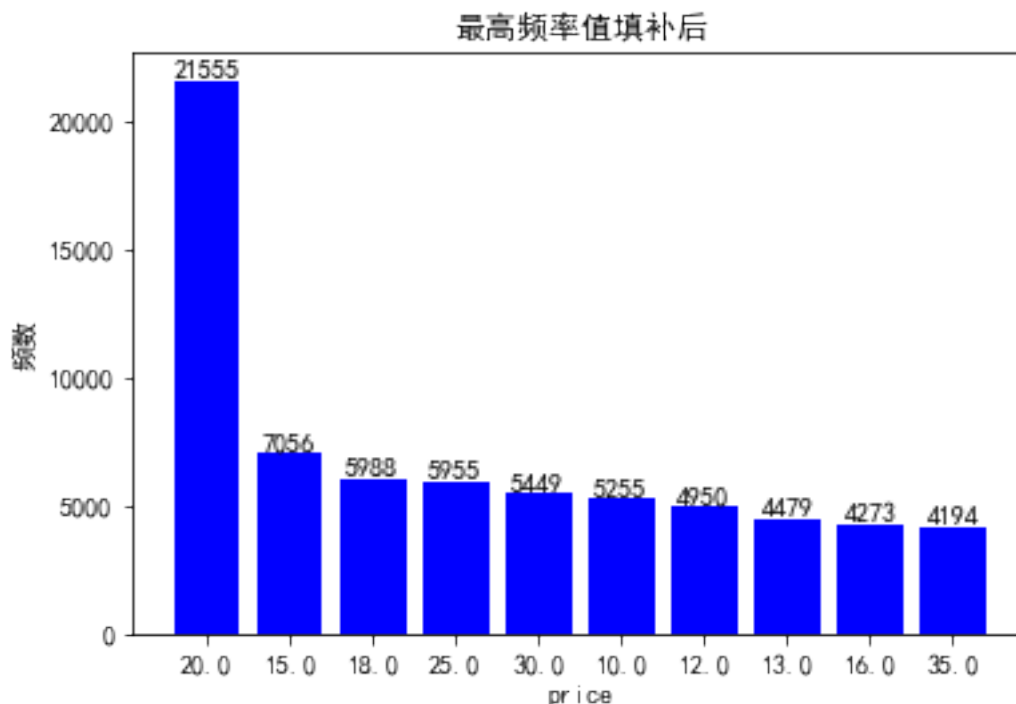
```

```

wine_data = pd.read_csv(path, header=0, index_col=0, engine='python', encoding='utf-8')
wine_data = wine_data.values
max_time = [] #
#
for cl in range(wine_data.shape[1]):
    counter = Counter(wine_data[:, cl])
    counter = counter.most_common() # listlist
    if counter[0][0] == counter[0][0]: #
        max_time.append(counter[0][0])
    else: #
        max_time.append(counter[1][0])
#
wine_max = pd.DataFrame(wine_data)
for cl in range(wine_data.shape[1]):
    wine_max[cl] = wine_max[cl].fillna(max_time[cl])
print(wine_max.describe())
draw(wine_max)

```

	3	4
count	150930.000000	150930.000000
mean	87.888418	31.939966
std	3.222392	34.840211
min	80.000000	4.000000
25%	86.000000	16.000000
50%	88.000000	22.000000
75%	90.000000	38.000000
max	100.000000	2300.000000



##2.3 price1033.13

```
In [46]: import pandas as pd
import numpy as np
from collections import Counter
import matplotlib.pyplot as plt
from sklearn.preprocessing import Imputer

plt.rcParams['font.sans-serif'] = ['SimHei']
plt.rcParams['axes.unicode_minus'] = False
path = 'C:/Users/ZL/Desktop/winemag-data_first150k.csv'

def draw(data, cl):
    num = 10
    wine_data = data.values
    counter = Counter(wine_data[:, cl])
    frequency = counter.most_common() # n
    num_list = []
    name_list = []
    for i in range(num):
        num_list.append(int(frequency[i][1]))
        name_list.append(str(frequency[i][0]))
    fig, ax = plt.subplots()
    b = ax.bar(name_list, num_list)
    plt.bar(range(len(num_list)), num_list, color='blue', tick_label=name_list)
```

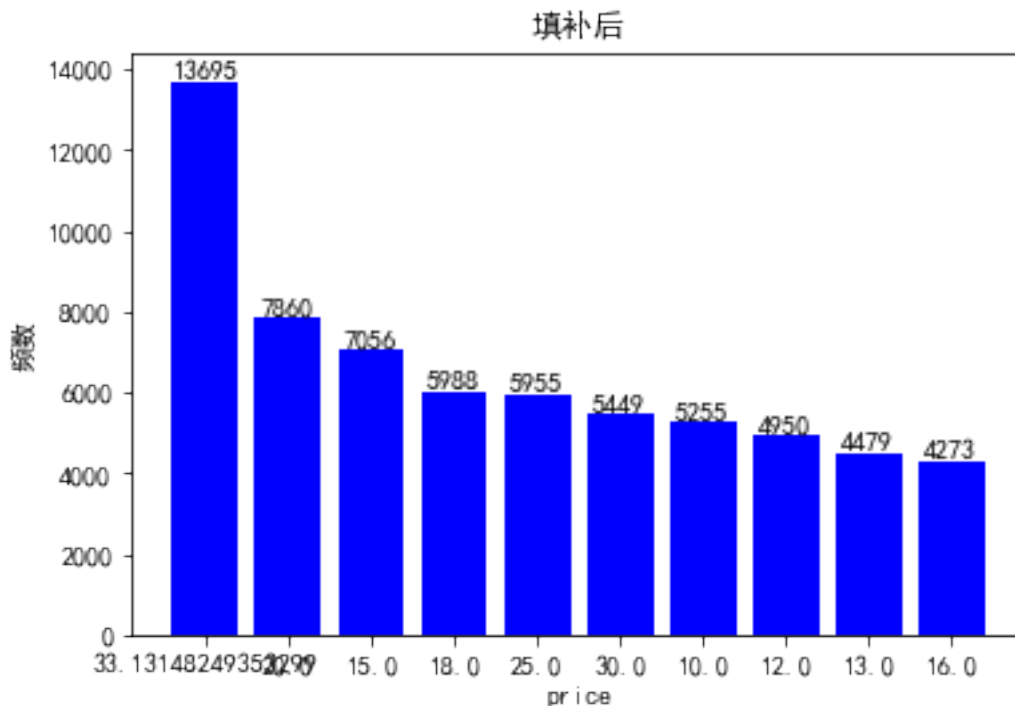
```

for a, b in zip(name_list, num_list):
    ax.text(a, b + 1, b, ha='center', va='bottom')
plt.title('')
plt.xlabel('price')
plt.ylabel('')
plt.show()

#
def att_rel(path):
    wine_data = pd.read_csv(path, header=0, index_col=0, engine='python', encoding='u
    wine_data = wine_data.values
    im = Imputer(missing_values='NaN', strategy='mean', axis=0)
    att_data = im.fit_transform(wine_data[:, 4].reshape(-1, 1))
    draw(pd.DataFrame(att_data), 0)

att_rel(path)

```



##2.4 priceIterativeImputer10

```

In [2]: import pandas as pd
import numpy as np
from collections import Counter
import matplotlib.pyplot as plt
from fancyimpute import IterativeImputer

```

```

plt.rcParams['font.sans-serif'] = ['SimHei']
plt.rcParams['axes.unicode_minus'] = False
path = 'C:/Users/ZL/Desktop/winemag-data_first150k.csv'

def draw(data, cl):
    num = 10
    wine_data = data.values
    counter = Counter(wine_data[:, cl])
    frequency = counter.most_common() # n
    num_list = []
    name_list = []
    for i in range(num):
        num_list.append(int(frequency[i][1]))
        name_list.append(str(frequency[i][0]))
    fig, ax = plt.subplots()
    b = ax.bar(name_list, num_list)
    plt.bar(range(len(num_list)), num_list, color='blue', tick_label=name_list)
    for a, b in zip(name_list, num_list):
        ax.text(a, b + 1, b, ha='center', va='bottom')
    plt.title('')
    plt.xlabel('price')
    plt.ylabel('')
    plt.show()

#
def obj_sim(path):
    wine_data = pd.read_csv(path, header=0, index_col=0, engine='python', encoding='utf-8')
    wine_data = wine_data.values
    # t = BiScaler().fit_transform(wine_data[:, 4].reshape(-1, 1))
    # obj_data = SoftImpute().fit_transform(t)
    obj_data = IterativeImputer().fit_transform(wine_data[:, 4].reshape(-1, 1))
    draw(pd.DataFrame(obj_data), 0)

obj_sim(path)

```

