wine-reviews

May 4, 2020

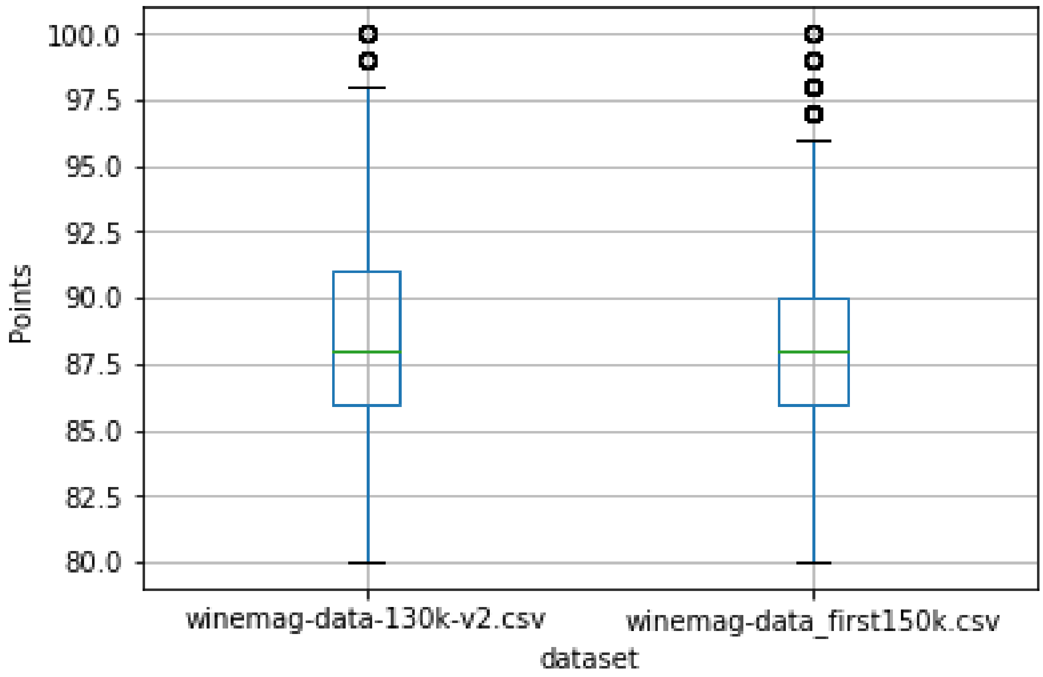
```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        from scipy.spatial.distance import pdist
        from math import ceil
        %matplotlib inline
        data1 = pd.read_csv("Wine/winemag-data_first150k.csv",encoding="utf-8")[
            ["country", "points", "price", "province", "variety", "winery"]
        ]
        data2 = pd.read_csv("Wine/winemag-data-130k-v2.csv",encoding="utf-8")[
            ["country", "points", "price", "province", "taster_name", "variety", "winery"]
        ]


        print("Wine/winemag-data_first150k.csvcountryprovincevarietywinery")
        print("Wine/winemag-data-130k-v2.csvcountryprovincetaster_namevarietywinery")

Wine/winemag-data_first150k.csvcountryprovincevarietywinery
Wine/winemag-data-130k-v2.csvcountryprovincetaster_namevarietywinery
```

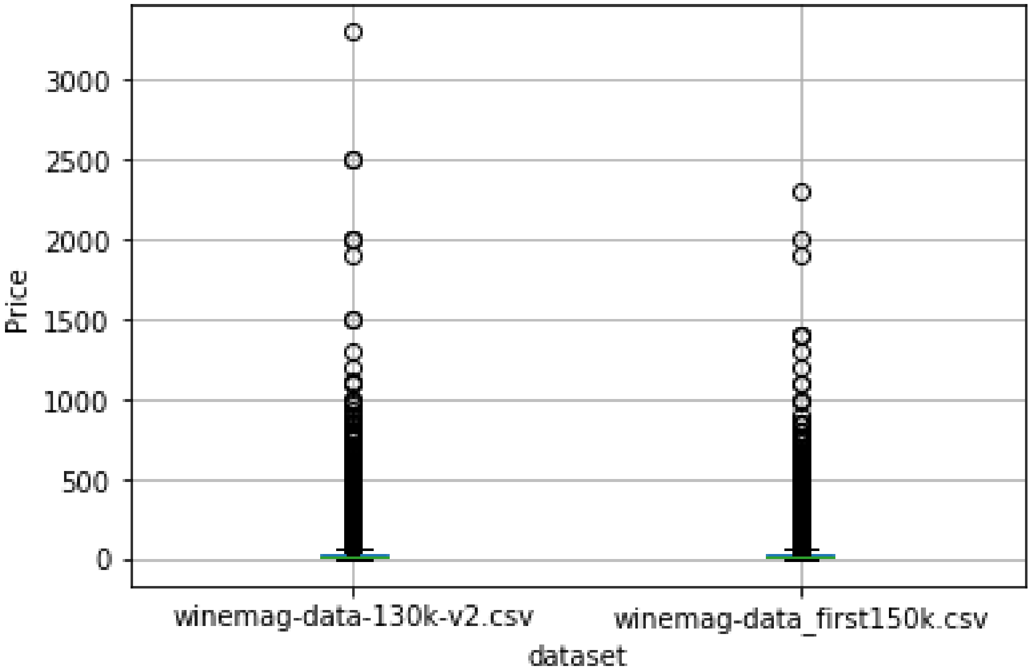
```
In [2]: def fiveNumber(nums):
            # MinimumQ1MedianQ3Maximum
            Minimum=min(nums)
            Maximum=max(nums)
            Q1=np.percentile(nums,25)
            Median=np.median(nums)
            Q3=np.percentile(nums,75)

            IQR=Q3-Q1
            lower_limit=Q1-1.5*IQR #
            upper_limit=Q3+1.5*IQR #

            return Minimum,Q1,Median,Q3,Maximum

        print("pointsprice")
        d = pd.DataFrame(data=data1[["price"]])#price,DataFrame
        d=d.dropna(axis=0, how='any')
        d=d.values
        d=d.flatten()
        m=fiveNumber(d)
        points_five1=fiveNumber(data1[["points"][0]])
        print("Wine/winemag-data_first150k.csv\npoints"+str(data1[["points"]].isnull().sum()[0
              ""+str(points_five1)+"\nprice"+str(data1[["price"]].isnull().sum()[0])+""+str(m)

        d2 = pd.DataFrame(data=data2[["price"]])
        d2=d2.dropna(axis=0, how='any')
        keep2=d2
        d2=d2.values
        d2=d2.flatten()
        m2=fiveNumber(d2)
        points_five2=fiveNumber(data2[["points"][0]])
        print("Wine/winemag-data-130k-v2.csv\npoints"+str(data2[["points"]].isnull().sum()[0])
              ""+str(points_five2)+"\nprice"+str(data2[["price"]].isnull().sum()[0])+""+str(m2


        #wineries = data1[["winery", "points", "price"]].groupby(by="winery").mean()
        #print( "Coeffitient of Pirson: "+str(wineries["points"].corr(wineries["price"]))
```

```
pointsprice
Wine/winemag-data_first150k.csv
points0(80, 86.0, 88.0, 90.0, 100)
price13695(4.0, 16.0, 24.0, 40.0, 2300.0)
Wine/winemag-data-130k-v2.csv
points0(80, 86.0, 88.0, 91.0, 100)
price8996(4.0, 17.0, 25.0, 42.0, 3300.0)
```

```
In [3]: #point

        point_box = pd.DataFrame({"winemag-data_first150k.csv":data1[["points"][0]],
                                  "winemag-data-130k-v2.csv":data2[["points"][0]]})
        point_box.boxplot()
        plt.ylabel("Points")
        plt.xlabel("dataset")
        plt.show()

        price_box = pd.DataFrame({"winemag-data_first150k.csv":data1[["price"][0]],
                                  "winemag-data-130k-v2.csv":data2[["price"][0]]})
        price_box.boxplot()
        plt.ylabel("Price")
        plt.xlabel("dataset")
        plt.show()
        point1_out=[]
        point2_out=[]
        #
```

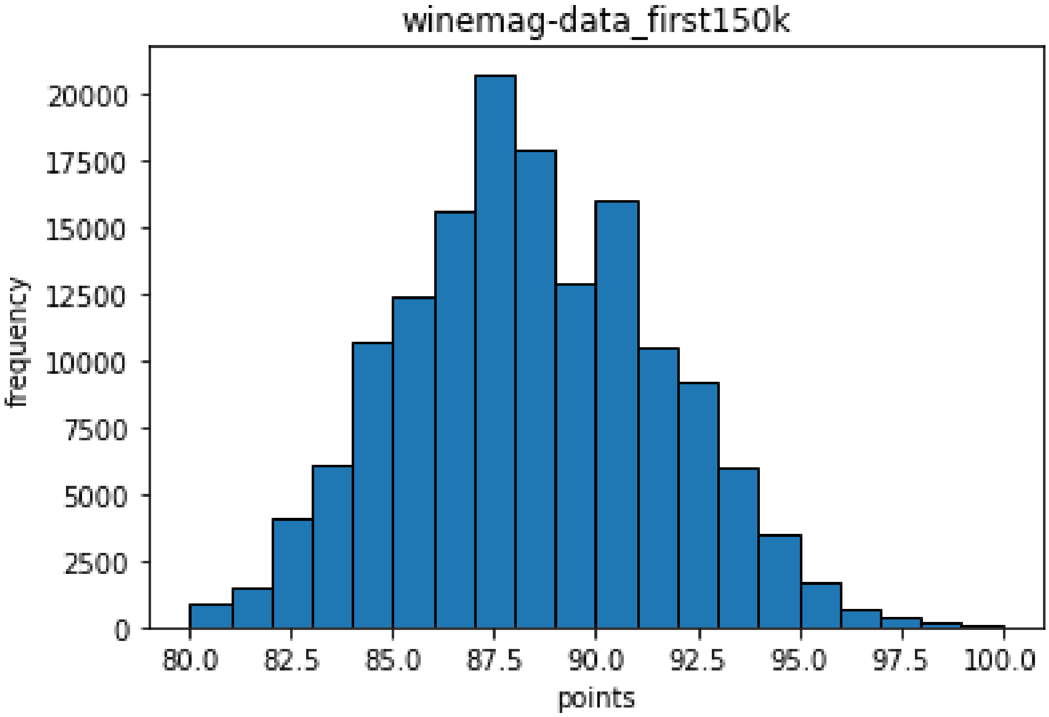
```
def outpoint(data,point3,point1):
    out=[]
    for i in range(len(data)):
        if (data[i]>(point3+1.5*(point3-point1))or\
           data[i]<(point1-1.5*(point3-point1))):
               out.append(data[i])
    return out
#point1_out=outpoint(data1[["points"][0]],points_five1[3],points_five1[1])
#print(point1_out)1point
#point2_out=outpoint(data2[["points"][0]],points_five2[3],points_five2[1])
#print(point2_out)2point
#price1_out=outpoint(data1[["price"][0]],points_five1[3],points_five1[1])
#print(price1_out)1price
#price2_out=outpoint(data2[["price"][0]],points_five2[3],points_five2[1])
#print(price2_out)2price
```

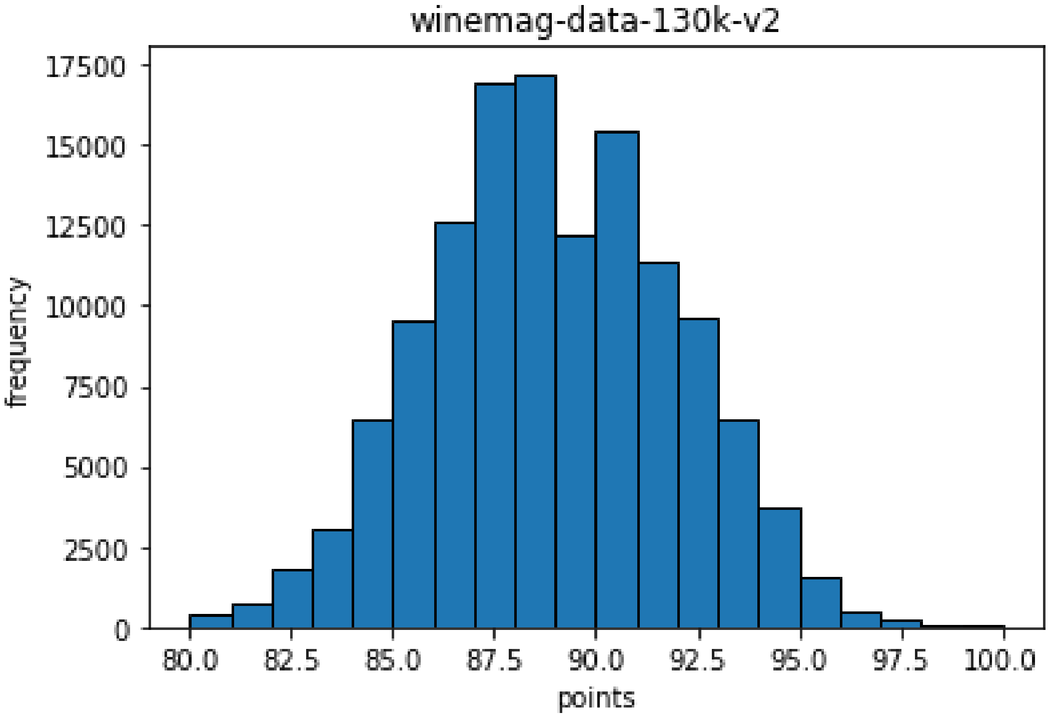




```
In [4]: print("Points' Histogram:")#
        plt.show()
        plt.hist(data1[["points"][0]], bins=20,  edgecolor = 'black',\
                 histtype='bar', align='mid', orientation='vertical')
        plt.xlabel('points')
        plt.ylabel('frequency')
        plt.title('winemag-data_first150k')
        plt.show()
        plt.hist(data2[["points"][0]], bins=20,  edgecolor = 'black',\
                 histtype='bar', align='mid', orientation='vertical')
        plt.xlabel('points')
        plt.ylabel('frequency')
        plt.title('winemag-data-130k-v2')
        plt.show()
```

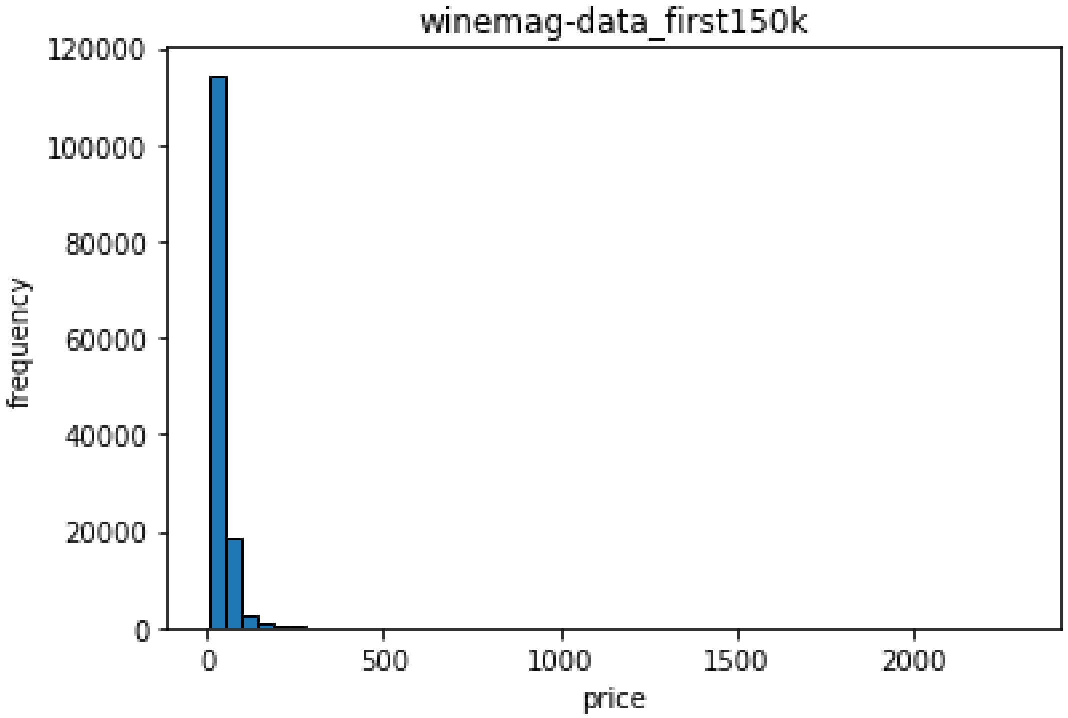
```
Points' Histogram:
```

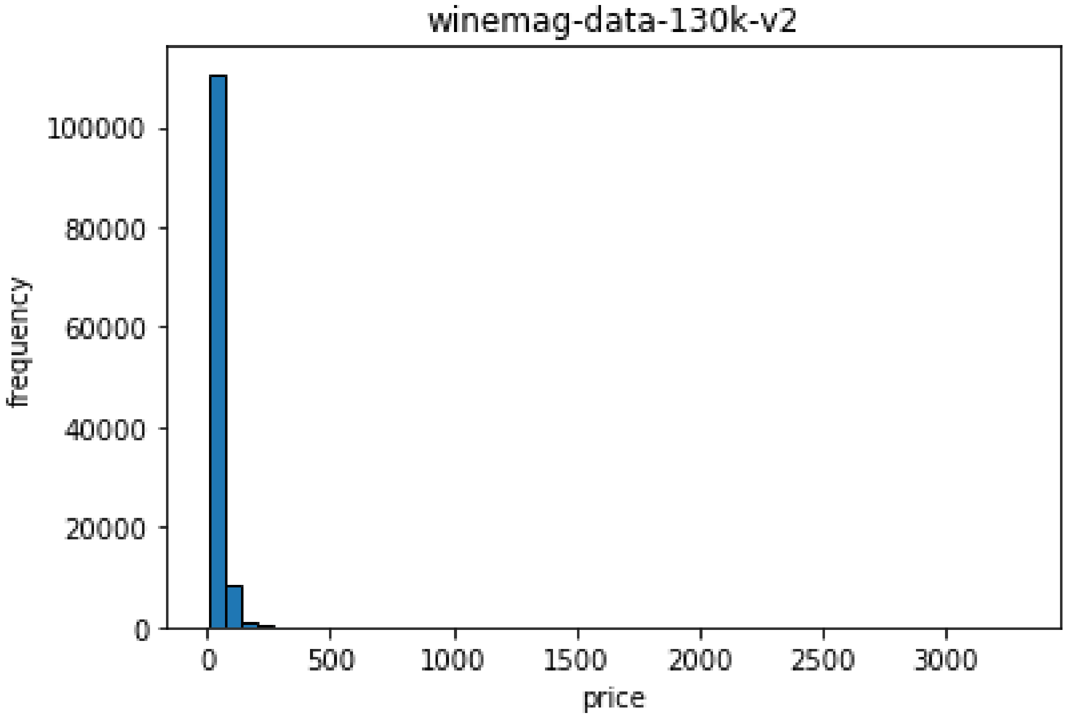

winemag-data_first150k
frequency
20000
17500
15000
12500
10000
7500
5000
2500
0
80.0
82.5
85.0
87.5
90.0
92.5
95.0
97.5
100.0
points


winemag-data-130k-v2
frequency
17500
15000
12500
10000
7500
5000
2500
0
80.0
82.5
85.0
87.5
90.0
92.5
95.0
97.5
100.0
points

```
In [5]: print("Points' Histogram:")#

        plt.hist(d, bins=50,  edgecolor = 'black',\
                 histtype='bar', align='mid', orientation='vertical')
        plt.xlabel('price')
        plt.ylabel('frequency')
        plt.title('winemag-data_first150k')
        plt.show()
        plt.hist(d2, bins=50,  edgecolor = 'black',\
                 histtype='bar', align='mid', orientation='vertical')
        plt.xlabel('price')
        plt.ylabel('frequency')
        plt.title('winemag-data-130k-v2')
        plt.show()
```

```
Points' Histogram:
```





```
In [6]: print("price")
        num_grapes = data1[ ["country","price", "variety"] ]
        num_grapes=num_grapes[num_grapes['price'].isnull()]
        num_country=num_grapes.groupby(by="country").count().sort_values(by="variety")[::-1]
        num_variety=num_grapes.groupby(by="variety").count().sort_values(by="country")[::-1]
        num_grapes2 = data2[ ["country","price", "variety"] ]
        num_grapes2=num_grapes2[num_grapes2['price'].isnull()]
        num_country2=num_grapes2.groupby(by="country").count().sort_values(by="variety")[::-1]
        num_variety2=num_grapes2.groupby(by="variety").count().sort_values(by="country")[::-1]
        print(num_country)
        print(num_variety)
        print("countryvariety,FranceItaly()")
```

```
price
              price  variety
country
France            0     6313
Italy             0     4694
Portugal          0     1146
Austria           0      574
US                0      258
New Zealand       0      250
Spain             0      108
Germany           0      105
```

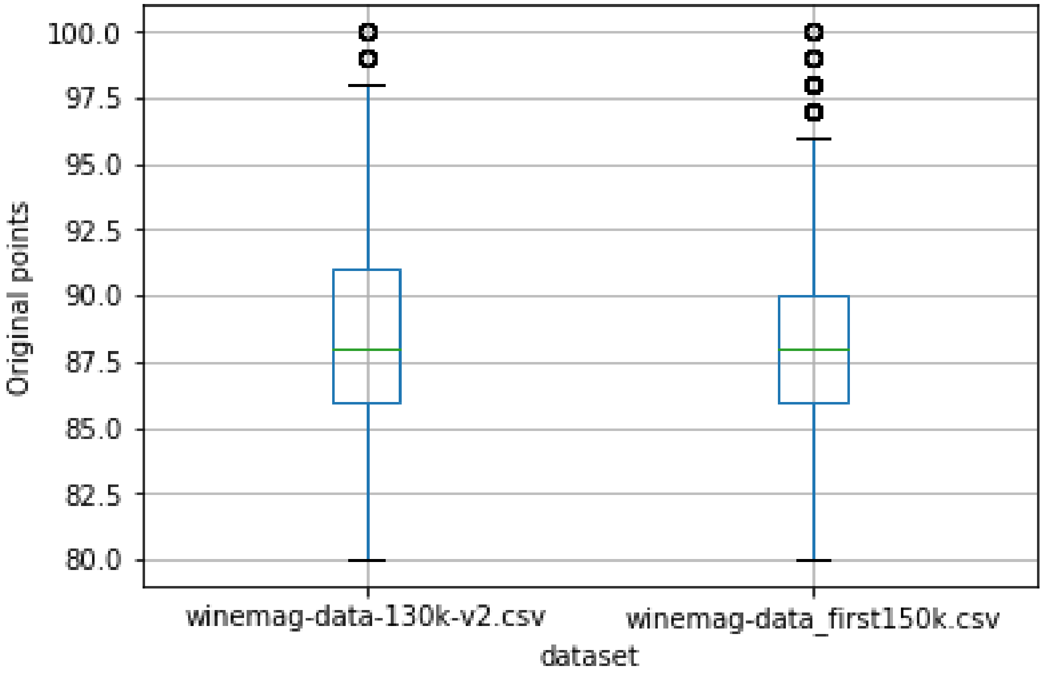
```
Australia         0       63
Chile             0       50
Argentina         0       44
South Africa      0       21
Israel            0       20
Slovenia          0       13
Greece            0       12
Uruguay           0        7
Croatia           0        6
Egypt             0        3
Canada            0        2
Tunisia           0        2
Turkey            0        2
England           0        1
Hungary           0        1
                               country  price
variety
Bordeaux-style Red Blend          2802      0
Nebbiolo                           712      0
Chardonnay                         707      0
Red Blend                          684      0
Bordeaux-style White Blend         681      0
Pinot Noir                         663      0
Sangiovese                         466      0
Sangiovese Grosso                  408      0
Portuguese Red                     404      0
Barbera                            398      0
Corvina, Rondinella, Molinara      390      0
Rosé                               356      0
Riesling                           312      0
White Blend                        270      0
Port                               267      0
Sauvignon Blanc                    266      0
Gamay                              236      0
Champagne Blend                    235      0
Portuguese White                   225      0
Grüner Veltliner                   200      0
Sparkling Blend                    184      0
Syrah                              158      0
Glera                              143      0
Cabernet Sauvignon                 129      0
Malbec                             123      0
Chenin Blanc                       107      0
Garganega                           95      0
Gewürztraminer                      91      0
Pinot Gris                          90      0
Nero d'Avola                        83      0
...                                ...    ...
```

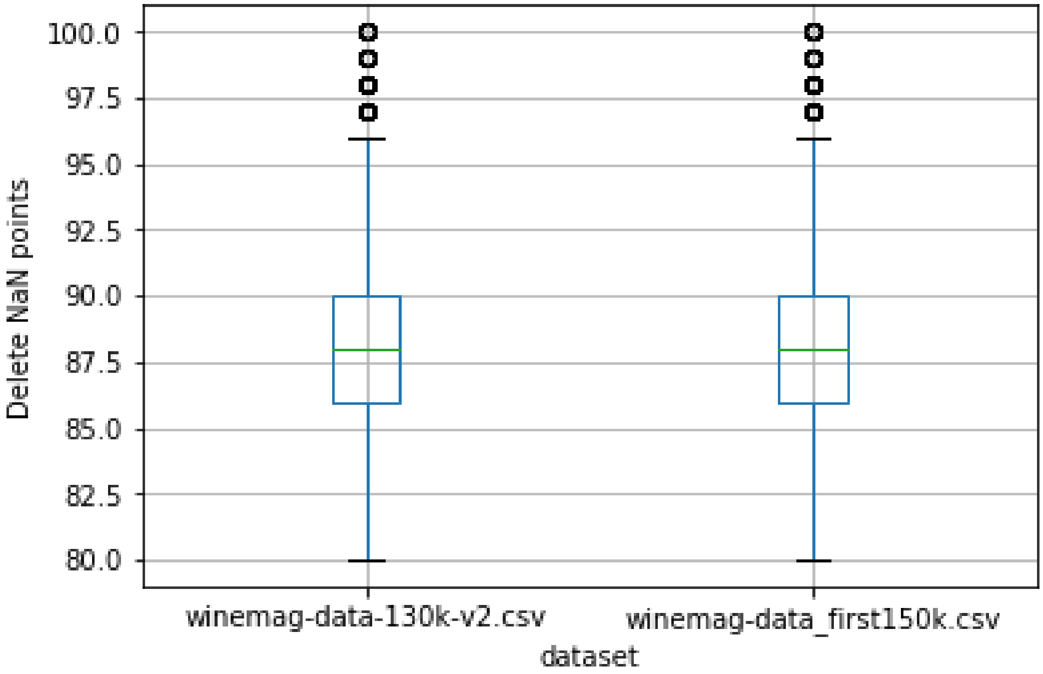
```
Pinot Auxerrois                       1      0
Petite Verdot                         1      0
Petit Meslier                         1      0
Johannisberg Riesling                 1      0
Loin de l'Oeil                        1      0
Pallagrello                           1      0
Greco Bianco                          1      0
Nasco                                 1      0
Grignolino                            1      0
Muskat Ottonel                        1      0
Muskat                                1      0
Friulano                              1      0
Roditis                               1      0
Magliocco                             1      0
Mansois                               1      0
Carricante                            1      0
Siria                                 1      0
Mondeuse                              1      0
Sercial                               1      0
Chardonnay-Pinot Blanc                1      0
Chardonnay-Sauvignon                  1      0
Sauvignon Gris                        1      0
Roscetto                              1      0
Sauvignon Blanc-Sauvignon Gris        1      0
Sacy                                  1      0
Corvina                               1      0
Roviello                              1      0
Merlot-Syrah                          1      0
Espadeiro                             1      0
Malvasia Bianca                       1      0

[256 rows x 2 columns]
countryvariety,FranceItaly()
```

1

```
In [7]: #
        def hist(d,bin,x,y,t):
            plt.hist(d, bins=bin,  edgecolor = 'black',\
                 histtype='bar', align='mid', orientation='vertical')
            plt.xlabel(x)
            plt.ylabel(y)
            plt.title(t)
            plt.show()
        def box(data1,data2,y):
            box = pd.DataFrame({"winemag-data_first150k.csv":data1,
                                 "winemag-data-130k-v2.csv":data2})
            box.boxplot()
            plt.ylabel(y)
            plt.xlabel("dataset")
            plt.show()
```

```
In [8]: #

        data1_1=data1
        data2_1=data2
        data1_1=data1_1.dropna(axis=0, how='any')
        data1_2=data1_1.dropna(axis=0, how='any')
        box(data1[["points"][0]],data2[["points"][0]],'Original points')
        box(data1_1[["points"][0]],data1_2[["points"][0]],'Delete NaN points')
        print("priceNANpricepointswinemag-data-130k-v2")
```

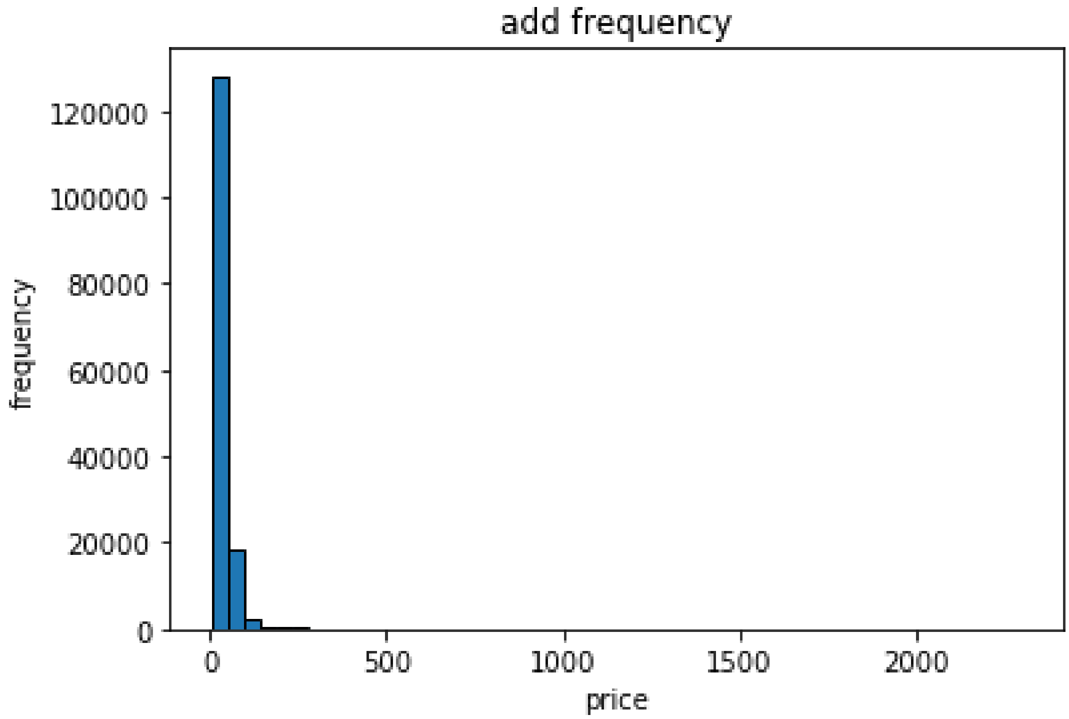


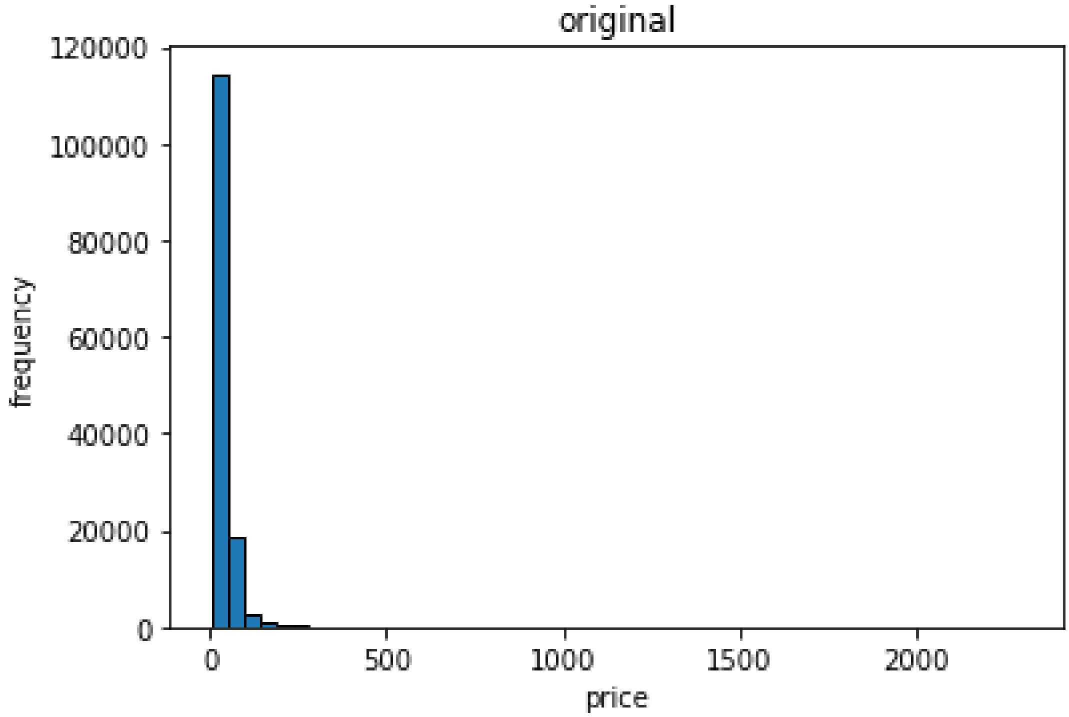


```
priceNANpricepointswinemag-data-130k-v2
```

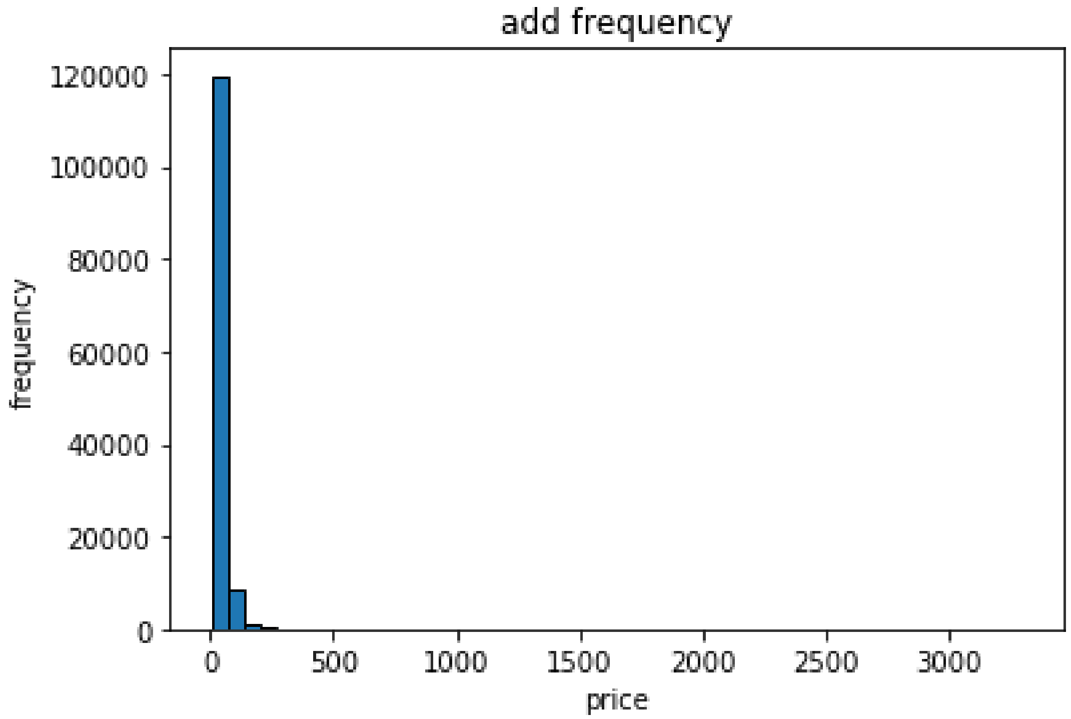
```
In [9]: #
        frequency1=data1[["price"][0]].mode()
        frequency2=data2[["price"][0]].mode()
        data2_1=data1[["price"][0]].fillna(frequency1[0])
        data2_2=data2[["price"][0]].fillna(frequency2[0])
        print("winemag-data_first150k.csv:"+str(frequency1[0]))
        hist(data2_1,50,"price","frequency","add frequency")
        hist(d,50,"price","frequency","original")
        print("0~60120000:\n")
        print(fiveNumber(data2_1))
        print("\nQ324402238")

        print("\n\nwinemag-data-130k-v2.csv:"+str(frequency2[0]))
        hist(data2_2,50,"price","frequency","add frequency")
        hist(d2,50,"price","frequency","original")
        print("0~601212:\n")
        print(fiveNumber(data2_2))
        print("\nQ1Q317421840")
```

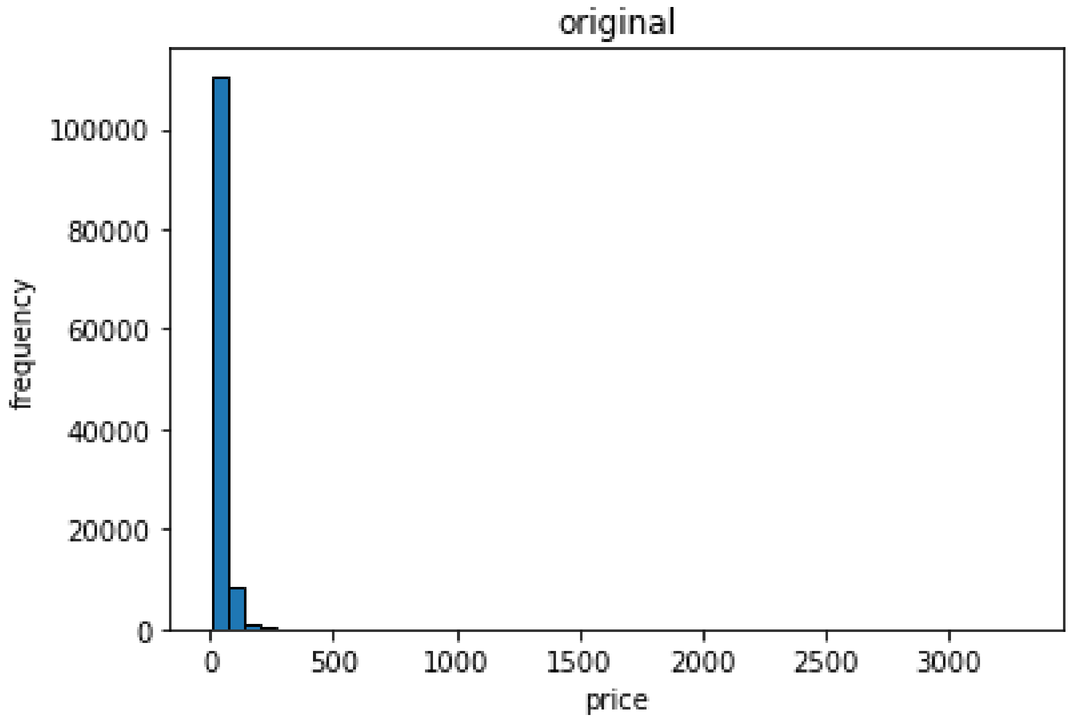
```
winemag-data_first150k.csv:20.0
```


add frequency
frequency
120000
100000
80000
60000
40000
20000
0
0
500
1000
1500
2000
price


original
frequency
120000
100000
80000
60000
40000
20000
0
0
500
1000
1500
2000
price

```
0~60120000:

(4.0, 16.0, 22.0, 38.0, 2300.0)

Q324402238


winemag-data-130k-v2.csv:20.0
```





```
0~601212:

(4.0, 18.0, 25.0, 40.0, 3300.0)

Q1Q317421840
```

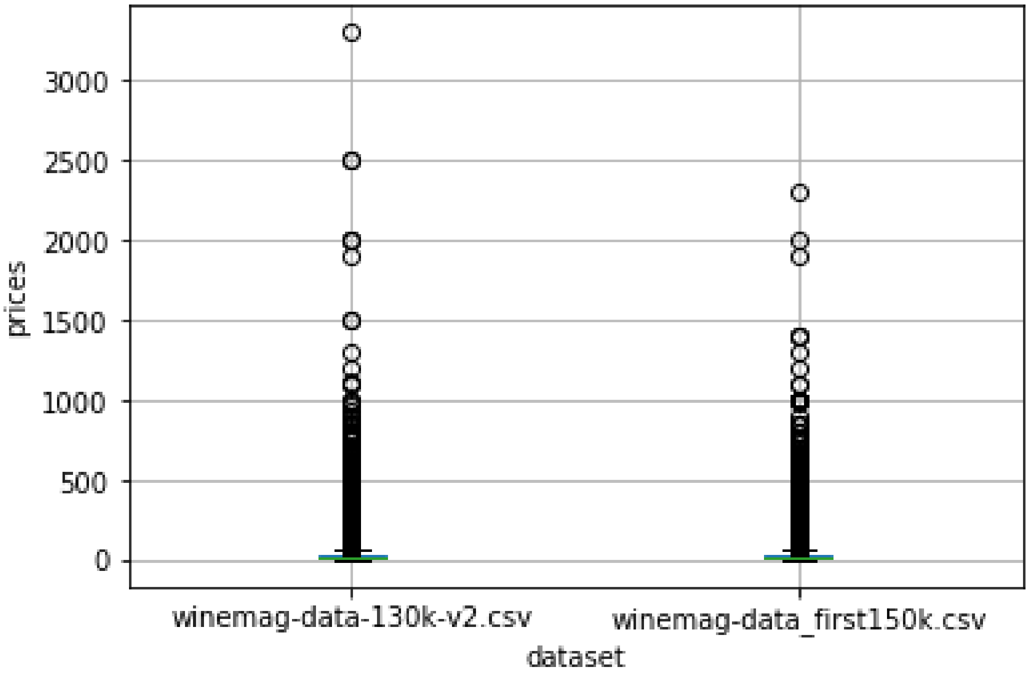
```
In [11]: #
         print("priceFranceItaly")
         data3_1=data1
         data3_2=data2
         data3_3=data3_1[data3_1["country"]=="France"]
         data3_4=data3_1[data3_1["country"]=="Italy"]
         data3_3=data3_3.dropna(axis=0, how='any')
         data3_4=data3_4.dropna(axis=0, how='any')

         data3_5=data3_2[data3_2["country"]=="France"]
         data3_6=data3_2[data3_2["country"]=="Italy"]
         data3_5=data3_5.dropna(axis=0, how='any')
         data3_6=data3_6.dropna(axis=0, how='any')
         #print(data3_3)
         frequency_France1=data3_3["price"].mode()
         frequency_Italy1=data3_4["price"].mode()
```

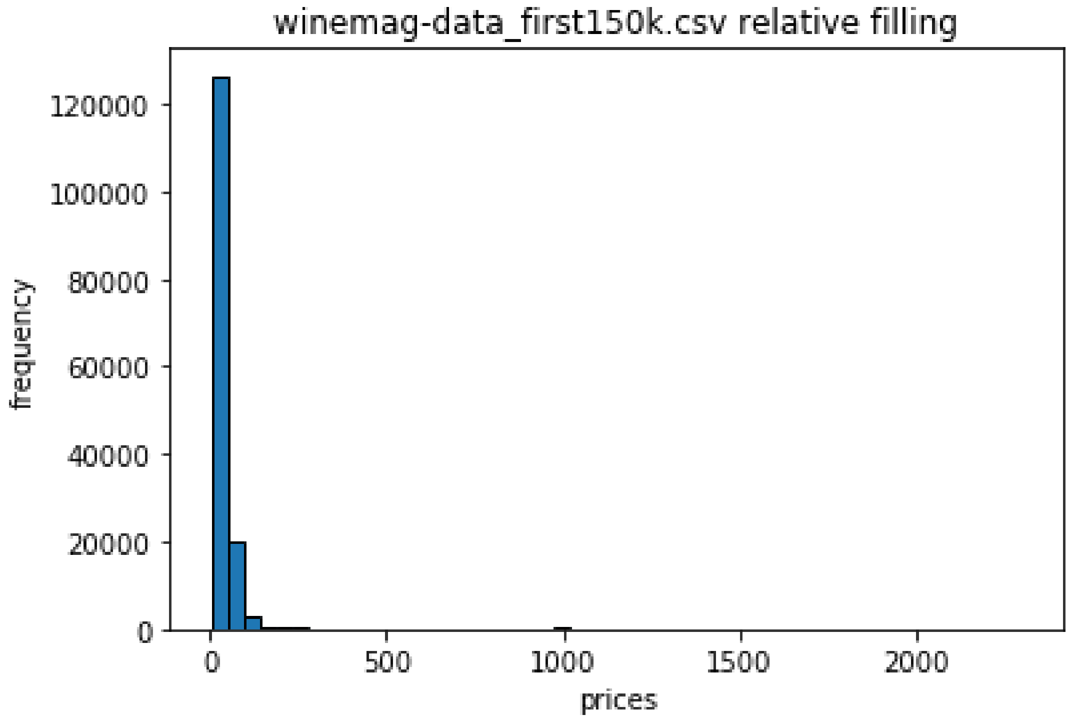
```
frequency_France2=data3_5["price"].mode()
frequency_Italy2=data3_6["price"].mode()
print("winemag-data_first150k.csv: France mode:"+str(frequency_France1[0])+"  Italy mo
print("winemag-data-130k-v2.csv: France mode:"+str(frequency_France2[0])+"  Italy mode
print("20")
```

```
priceFranceItaly
winemag-data_first150k.csv: France mode:20.0  Italy mode:20.0
winemag-data-130k-v2.csv: France mode:20.0  Italy mode:20.0
20
```

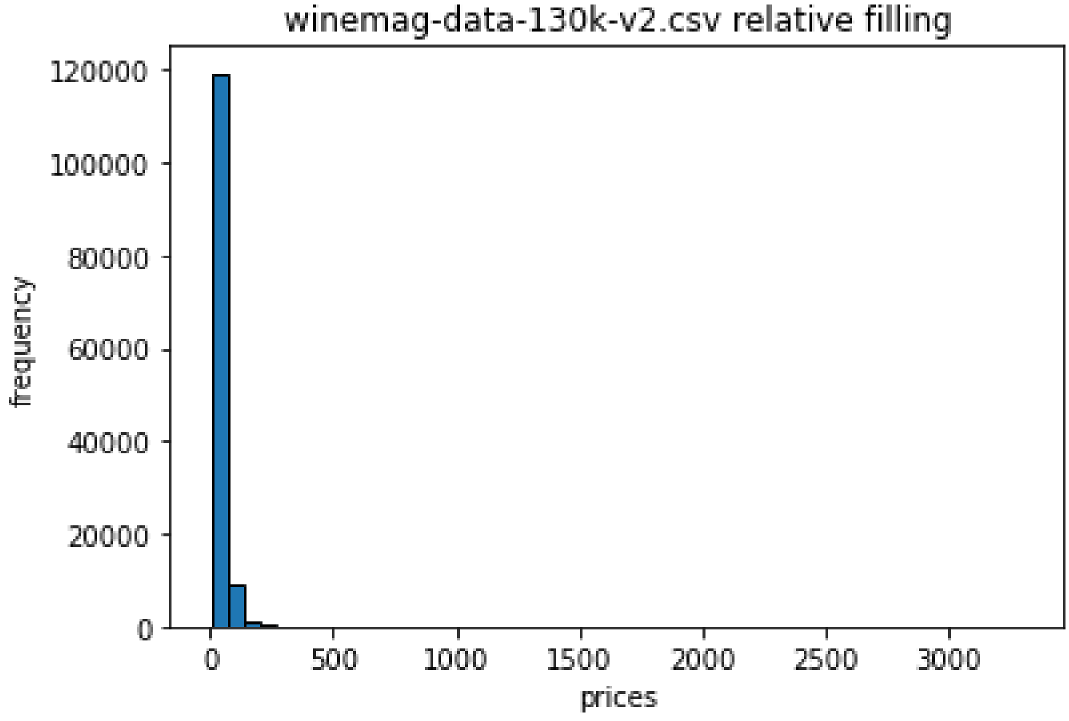
```
In [59]: #pointprice
         print("")
         data4_1=data1
         data4_2=data2
         data4_1=data4_1.dropna(axis=0, how='any')
         data4_2=data4_2.dropna(axis=0, how='any')
         points1=data4_1["points"]
         price1=data4_1["price"]
         points2=data4_2["points"]
         price2=data4_2["price"]
         cos1 = np.vstack([points1,price1])
         p1 = 1 - pdist(cos1,'cosine')
         cos2 = np.vstack([points2,price2])
         p2 = 1 - pdist(cos2,'cosine')
         print("winemag-data_first150k.csv: PLCC="+str(points1.corr(price1,method="pearson"))+"
         print("winemag-data-130k-v2.csv: PLCC="+str(points2.corr(price2,method="pearson"))+"
         print("")
         #
         xx = data1[[ "points", "price"]].groupby(by="points").median()
         xx=xx.values
         xx=xx.flatten()
         yy = data1[[ "points", "price"]].groupby(by="points").median()
         yy=yy.values
         yy=yy.flatten()

         data_add1=data1
         data_add2=data2
         dataadd_g1=pd.DataFrame()
         dataadd_g2=pd.DataFrame()
         for i in range(80,101):
             data_ad1=data_add1.loc[data_add1['points'].isin([i])].fillna(xx[i-80])
             data_ad2=data_add2.loc[data_add1['points'].isin([i])].fillna(yy[i-80])
             if(i==80):
                 data_add_g1=data_ad1
                 data_add_g2=data_ad2
             else:
                 data_add_g1=pd.concat([data_add_g1,data_ad1],axis=0)
                 data_add_g2=pd.concat([data_add_g2,data_ad2],axis=0)
         prices1=data_add_g1[["price"][0]]
         prices2=data_add_g2[["price"][0]]
         box(prices1,prices2,"prices")
         hist(prices1,50,"prices","frequency","winemag-data_first150k.csv relative filling")
         print("winemag-data_first150k.csv"+str(fiveNumber((prices1.values).flatten())))
         hist(prices2,50,"prices","frequency","winemag-data-130k-v2.csv relative filling")
         print("winemag-data-130k-v2.csv"+str(fiveNumber((prices2.values).flatten())))
         print("NaN;\
                 \n0~6011221211000\
                 \n2Q34241")
```

```
winemag-data_first150k.csv: PLCC=0.3422656666692296  Cosine similarity=[0.49131376]
winemag-data-130k-v2.csv: PLCC=0.4040017582872982  Cosine similarity=[0.64080925]
```





```
winemag-data_first150k.csv(4.0, 16.0, 24.0, 40.0, 2300.0)
```



```
winemag-data-130k-v2.csv(4.0, 17.0, 25.0, 41.0, 3300.0)
NaN;
0~6011221211000
2Q34241
```