数据挖掘互评作业二: 频繁模式与关联规则挖掘

3220211057 茅晓璐 计算机学院

Github Repository link

数据预处理

要求

对数据集进行处理，转换成适合进行关联规则挖掘的形式。

思路及代码分析

数据输入

首先，需要从文件 winemag-data_first150k.csv 中读入数据，因此需要调用读入 csv 格式数据的模块。选择 **pandas** 作为数据读入工具。数据中存在空值 **NAN**，首先将其转换成空串。

拟采用 **mlxtend** 进行频繁模式挖掘，因此需要将数据处理成 mlxtend 所需要的格式。

```
from pandas import read_csv

def ReadInputCSV(input_csv):
    f = open(input_csv, encoding="UTF-8")
    names = [
        "id", "country", "description", "designation", "points",
        "price", "province", "region_1", "region_2", "variety", "winery",
    ]
    data = read_csv(f, names=names)
    data = data.fillna("")
    return data
```

```
   data = read_csv(f, names=names)
              id  country  ...             variety                   winery
0            NaN  country  ...             variety                   winery
1            0.0       US  ...  Cabernet Sauvignon                    Heitz
2            1.0    Spain  ...       Tinta de Toro  Bodega Carmen Rodríguez
3            2.0       US  ...     Sauvignon Blanc                 Macauley
4            3.0       US  ...          Pinot Noir                    Ponzi
...          ...      ...  ...                 ...                      ...
150926  150925.0    Italy  ...         White Blend    Feudi di San Gregorio
150927  150926.0   France  ...     Champagne Blend               H.Germain
150928  150927.0    Italy  ...         White Blend               Terredora
150929  150928.0   France  ...     Champagne Blend                  Gosset
150930  150929.0    Italy  ...        Pinot Grigio           Alois Lageder

[150931 rows x 11 columns]
input end
```

读入数据展示

数据处理

由于我们要分析的数据是葡萄酒的数据，它包含产地、价格、评分等内容。因此我们需要根据数据的实际意义进行推测。由于价格和评分为 **数值** 型数据，而在挖掘中采用该类型的数据可能无法很好地反映一组价格相近的数据的情况。因此我们观察数据的总体情况并进行分类。

通过观察数据，我们发现，有 114403 的数据价格在 49.92 以下，占大多数。因此我们以 50 为分界线，认定价格在 50 及以上的数据为 **highPrice**。同理，我们认定 90 分及以上的数据为 **highPoints**。

```
# 定义 price >= 50 的数据为高价格
highPrice = []
for price in data["price"]:
    if price == "" or price < 50:
        highPrice.append("LowPrice")
    else:
        highPrice.append("HighPrice")
data["highPrice"] = highPrice

# 定义 points >= 90 的数据为高分数
highPoints = []
for points in data["points"]:
    if price == "" or points < 90:
        highPoints.append("LowPoints")
    else:
        highPoints.append("HighPoints")
data["highPoints"] = highPoints
```

频繁模式挖掘

要求

找出频繁模式。

分析

保留 "province", "highPrice", "highPoints" 三列，对其进行频繁模式挖掘。

```
data1 = data[["province", "highPrice", "highPoints"]]
```

首先要进行 encode。encode 的过程是将整个 dataFrame 重新改写，原本的 data[PropertyName] 代表一列属性，而改写后的 dataFrame 为一个只包含 True 和 False 的表格，代表某个属性值是否在当前 id 中存在。

```
te = TransactionEncoder()
input_data = data.values.tolist()
print("input data")
for i in range(len(input_data)):
    for j in range(len(input_data[i])):
        input_data[i][j] = str(input_data[i][j])
df_tf = te.fit_transform(input_data)

df = pd.DataFrame(df_tf, columns=te.columns_)
```

进行频繁模式挖掘

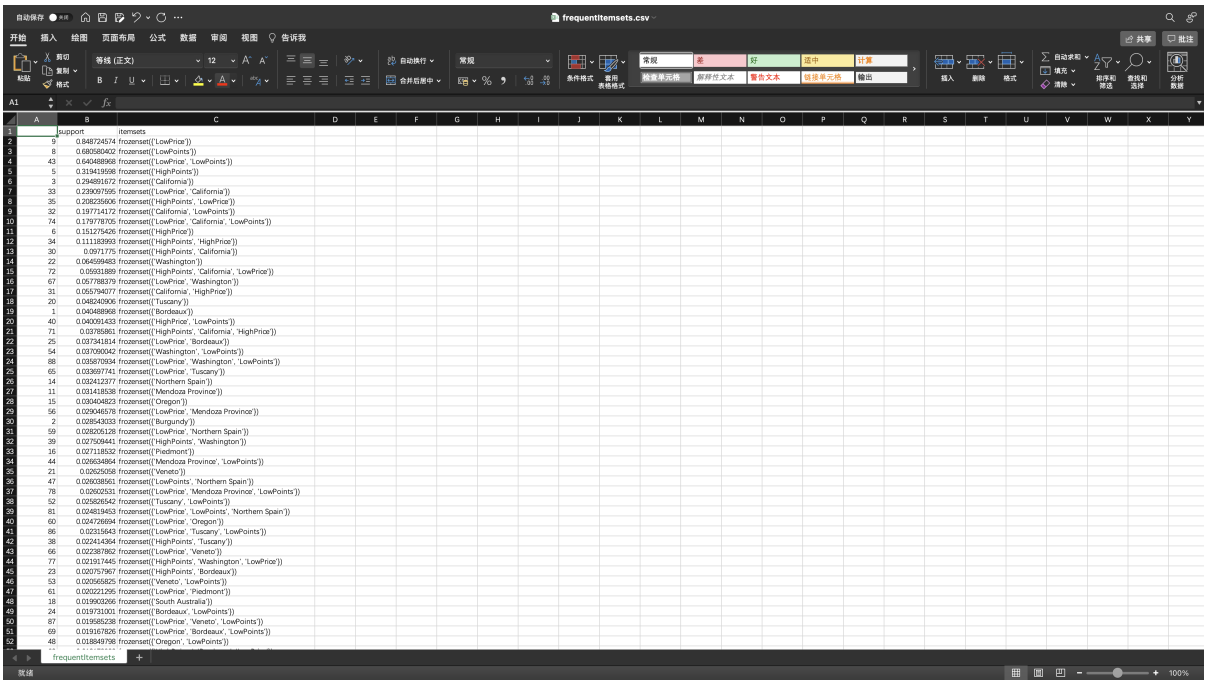
```
print("frequent pattern begin")
# use_colnames=True表示使用元素名字，默认的False使用列名代表元素，设置最小支持度min_support
```

```
frequent_itemsets = apriori(df, min_support=0.01, use_colnames=True)
print("frequent pattern end")
frequent_itemsets.sort_values(by="support", ascending=False, inplace=True)
```

筛选长度为2的频繁模式，并将结果输出到文件

```
# 选择2频繁项集
print(frequent_itemsets[frequent_itemsets.itemsets.apply(lambda x: len(x)) == 2])
print("frequent pattern end")
frequent_itemsets.to_csv("frequentItemsets.csv")
```

结果展示



频繁模式挖掘结果展示

关联规则挖掘

要求

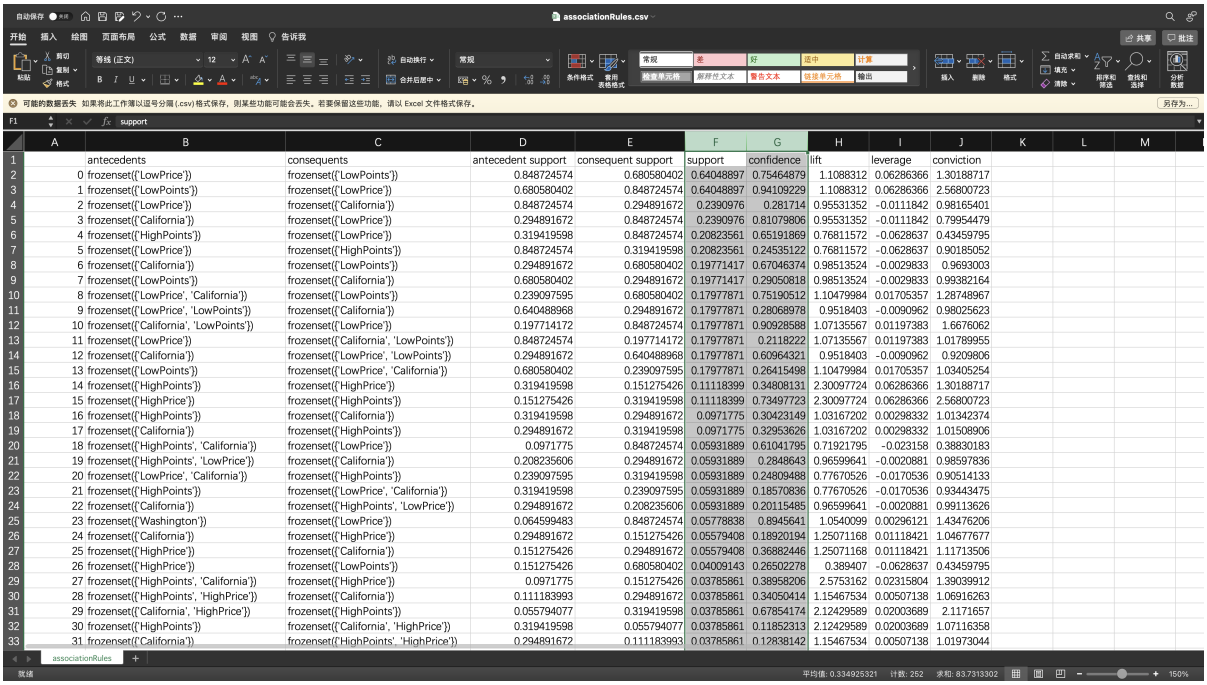
导出关联规则，计算其支持度和置信度。

思路及代码分析

保留 "province", "highPrice", "highPoints" 三列，对其进行关联规则挖掘。并将结果输出到文件。

```
association_rule = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.1)
association_rule.to_csv("associationRules.csv")
```

结果展示



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	frozenset({'LowPrice'})	frozenset({'LowPoints'})	0.848724574	0.680580402	0.64048897	0.75464879	1.1088312	0.06286366	1.30188717
1	frozenset({'LowPoints'})	frozenset({'LowPrice'})	0.680580402	0.848724574	0.64048897	0.94109229	1.1088312	0.06286366	2.56800723
2	frozenset({'LowPrice'})	frozenset({'California'})	0.848724574	0.294891672	0.2390976	0.281714	0.95531352	-0.0111842	0.98165401
3	frozenset({'California'})	frozenset({'LowPrice'})	0.294891672	0.848724574	0.2390976	0.81079806	0.95531352	-0.0111842	0.79954479
4	frozenset({'HighPoints'})	frozenset({'LowPrice'})	0.319419598	0.848724574	0.20823561	0.65191869	0.76811572	-0.0628637	0.43459795
5	frozenset({'LowPrice'})	frozenset({'HighPoints'})	0.848724574	0.319419598	0.20823561	0.24535122	0.76811572	-0.0628637	0.90185052
6	frozenset({'California'})	frozenset({'LowPoints'})	0.294891672	0.680580402	0.19771417	0.67046374	0.98513524	-0.0029833	0.9693003
7	frozenset({'LowPoints'})	frozenset({'California'})	0.680580402	0.294891672	0.19771417	0.29050818	0.98513524	-0.0029833	0.99382164
8	frozenset({'LowPrice', 'California'})	frozenset({'LowPoints'})	0.239097595	0.680580402	0.17977871	0.75190512	1.10479984	0.01705357	1.28748967
9	frozenset({'LowPrice', 'LowPoints'})	frozenset({'California'})	0.640488968	0.294891672	0.17977871	0.28068978	0.9518403	-0.0090962	0.98025623
10	frozenset({'California', 'LowPoints'})	frozenset({'LowPrice'})	0.197714172	0.848724574	0.17977871	0.90928588	1.07135567	0.01197383	1.6676062
11	frozenset({'LowPrice'})	frozenset({'California', 'LowPoints'})	0.848724574	0.197714172	0.17977871	0.2118222	1.07135567	0.01197383	1.01789955
12	frozenset({'California'})	frozenset({'LowPrice', 'LowPoints'})	0.294891672	0.640488968	0.17977871	0.60964321	0.9518403	-0.0090962	0.9209806
13	frozenset({'LowPoints'})	frozenset({'LowPrice', 'California'})	0.680580402	0.239097595	0.17977871	0.26415498	1.10479984	0.01705357	1.03405254
14	frozenset({'HighPoints'})	frozenset({'HighPrice'})	0.319419598	0.151275426	0.11118399	0.34808131	2.30097724	0.06286366	1.30188717
15	frozenset({'HighPrice'})	frozenset({'HighPoints'})	0.151275426	0.319419598	0.11118399	0.73497723	2.30097724	0.06286366	2.56800723
16	frozenset({'HighPoints'})	frozenset({'California'})	0.319419598	0.294891672	0.0971775	0.30423149	1.03167202	0.00298332	1.01342374
17	frozenset({'California'})	frozenset({'HighPoints'})	0.294891672	0.319419598	0.0971775	0.32953626	1.03167202	0.00298332	1.01508906
18	frozenset({'HighPoints', 'California'})	frozenset({'LowPrice'})	0.0971775	0.848724574	0.05931889	0.61041795	0.71921795	-0.023158	0.38830183
19	frozenset({'HighPoints', 'LowPrice'})	frozenset({'California'})	0.208235606	0.294891672	0.05931889	0.2848643	0.96599641	-0.0020881	0.98597836
20	frozenset({'LowPrice', 'California'})	frozenset({'HighPoints'})	0.239097595	0.319419598	0.05931889	0.24809488	0.77670526	-0.0170536	0.90514133
21	frozenset({'HighPoints'})	frozenset({'LowPrice', 'California'})	0.319419598	0.239097595	0.05931889	0.18570836	0.77670526	-0.0170536	0.93443475
22	frozenset({'California'})	frozenset({'HighPoints', 'LowPrice'})	0.294891672	0.208235606	0.05931889	0.20115485	0.96599641	-0.0020881	0.99113626
23	frozenset({'Washington'})	frozenset({'LowPrice'})	0.064599483	0.848724574	0.05778838	0.8945641	1.0540099	0.00296121	1.43476206
24	frozenset({'California'})	frozenset({'HighPrice'})	0.294891672	0.151275426	0.05579408	0.18920194	1.25071168	0.01118421	1.04677677
25	frozenset({'HighPrice'})	frozenset({'California'})	0.151275426	0.294891672	0.05579408	0.36882446	1.25071168	0.01118421	1.11713506
26	frozenset({'HighPrice'})	frozenset({'LowPoints'})	0.151275426	0.680580402	0.04009143	0.26502278	0.389407	-0.0628637	0.43459795
27	frozenset({'HighPoints', 'California'})	frozenset({'HighPrice'})	0.0971775	0.151275426	0.03785861	0.38958206	2.5753162	0.02315804	1.39039912
28	frozenset({'HighPoints', 'HighPrice'})	frozenset({'California'})	0.111183993	0.294891672	0.03785861	0.34050414	1.15467534	0.00507138	1.06916263
29	frozenset({'California', 'HighPrice'})	frozenset({'HighPoints'})	0.055794077	0.319419598	0.03785861	0.67854174	2.12429589	0.02003689	2.1171657
30	frozenset({'HighPoints'})	frozenset({'California', 'HighPrice'})	0.319419598	0.055794077	0.03785861	0.11852313	2.12429589	0.02003689	1.07116358
31	frozenset({'California'})	frozenset({'HighPoints', 'HighPrice'})	0.294891672	0.111183993	0.03785861	0.12838142	1.15467534	0.00507138	1.01973044

关联规则挖掘结果，及置信度/支持度

评价规则

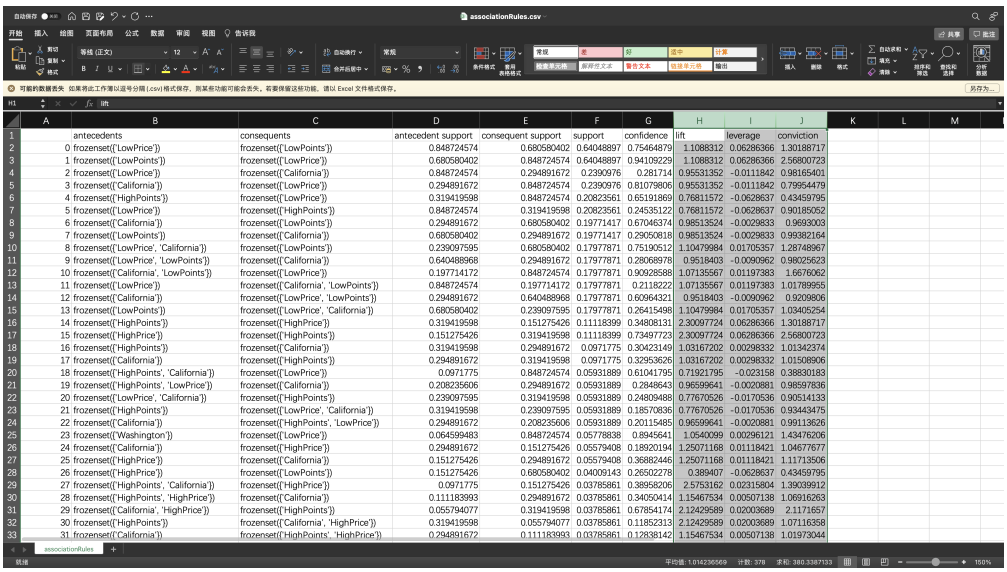
要求

对规则进行评价，可使用Lift、卡方和其它教材中提及的指标, 至少2种。

思路及代码分析

关联挖掘的结果中自带了一些评价指标，如 **antecedent support/ consequent support/ support/ confidence/ lift/ leverage/ conviction**

结果展示



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	frozenset({'LowPrice'})	frozenset({'LowPoints'})	0.848724574	0.680580402	0.64048897	0.75464879	1.1088312	0.06286366	1.30188717
1	frozenset({'LowPoints'})	frozenset({'LowPrice'})	0.680580402	0.848724574	0.64048897	0.94109229	1.1088312	0.06286366	2.56800723
2	frozenset({'LowPrice'})	frozenset({'California'})	0.848724574	0.294891672	0.2390976	0.281714	0.95531352	-0.0111842	0.98165401
3	frozenset({'California'})	frozenset({'LowPrice'})	0.294891672	0.848724574	0.2390976	0.81079806	0.95531352	-0.0111842	0.79954479
4	frozenset({'HighPoints'})	frozenset({'LowPrice'})	0.319419598	0.848724574	0.20823561	0.65191869	0.76811572	-0.0628637	0.43459795
5	frozenset({'LowPrice'})	frozenset({'HighPoints'})	0.848724574	0.319419598	0.20823561	0.24535122	0.76811572	-0.0628637	0.90185052
6	frozenset({'California'})	frozenset({'LowPoints'})	0.294891672	0.680580402	0.19771417	0.67046374	0.98513524	-0.0029833	0.9693003
7	frozenset({'LowPoints'})	frozenset({'California'})	0.680580402	0.294891672	0.19771417	0.29050818	0.98513524	-0.0029833	0.99382164
8	frozenset({'LowPrice', 'California'})	frozenset({'LowPoints'})	0.239097595	0.680580402	0.17977871	0.75190512	1.10479984	0.01705357	1.28748967
9	frozenset({'LowPrice', 'LowPoints'})	frozenset({'California'})	0.640488968	0.294891672	0.17977871	0.28068978	0.9518403	-0.0090962	0.98025623
10	frozenset({'California', 'LowPoints'})	frozenset({'LowPrice'})	0.197714172	0.848724574	0.17977871	0.90928588	1.07135567	0.01197383	1.6676062
11	frozenset({'LowPrice'})	frozenset({'California', 'LowPoints'})	0.848724574	0.197714172	0.17977871	0.2118222	1.07135567	0.01197383	1.01789955
12	frozenset({'California'})	frozenset({'LowPrice', 'LowPoints'})	0.294891672	0.640488968	0.17977871	0.60964321	0.9518403	-0.0090962	0.9209806
13	frozenset({'LowPoints'})	frozenset({'LowPrice', 'California'})	0.680580402	0.239097595	0.17977871	0.26415498	1.10479984	0.01705357	1.03405254
14	frozenset({'HighPoints'})	frozenset({'HighPrice'})	0.319419598	0.151275426	0.11118399	0.34808131	2.30097724	0.06286366	1.30188717
15	frozenset({'HighPrice'})	frozenset({'HighPoints'})	0.151275426	0.319419598	0.11118399	0.73497723	2.30097724	0.06286366	2.56800723
16	frozenset({'HighPoints'})	frozenset({'California'})	0.319419598	0.294891672	0.0971775	0.30423149	1.03167202	0.00298332	1.01342374
17	frozenset({'California'})	frozenset({'HighPoints'})	0.294891672	0.319419598	0.0971775	0.32953626	1.03167202	0.00298332	1.01508906
18	frozenset({'HighPoints', 'California'})	frozenset({'LowPrice'})	0.0971775	0.848724574	0.05931889	0.61041795	0.71921795	-0.023158	0.38830183
19	frozenset({'HighPoints', 'LowPrice'})	frozenset({'California'})	0.208235606	0.294891672	0.05931889	0.2848643	0.96599641	-0.0020881	0.98597836
20	frozenset({'LowPrice', 'California'})	frozenset({'HighPoints'})	0.239097595	0.319419598	0.05931889	0.24809488	0.77670526	-0.0170536	0.90514133
21	frozenset({'HighPoints'})	frozenset({'LowPrice', 'California'})	0.319419598	0.239097595	0.05931889	0.18570836	0.77670526	-0.0170536	0.93443475
22	frozenset({'California'})	frozenset({'HighPoints', 'LowPrice'})	0.294891672	0.208235606	0.05931889	0.20115485	0.96599641	-0.0020881	0.99113626
23	frozenset({'Washington'})	frozenset({'LowPrice'})	0.064599483	0.848724574	0.05778838	0.8945641	1.0540099	0.00296121	1.43476206
24	frozenset({'California'})	frozenset({'HighPrice'})	0.294891672	0.151275426	0.05579408	0.18920194	1.25071168	0.01118421	1.04677677
25	frozenset({'HighPrice'})	frozenset({'California'})	0.151275426	0.294891672	0.05579408	0.36882446	1.25071168	0.01118421	1.11713506
26	frozenset({'HighPrice'})	frozenset({'LowPoints'})	0.151275426	0.680580402	0.04009143	0.26502278	0.389407	-0.0628637	0.43459795
27	frozenset({'HighPoints', 'California'})	frozenset({'HighPrice'})	0.0971775	0.151275426	0.03785861	0.38958206	2.5753162	0.02315804	1.39039912
28	frozenset({'HighPoints', 'HighPrice'})	frozenset({'California'})	0.111183993	0.294891672	0.03785861	0.34050414	1.15467534	0.00507138	1.06916263
29	frozenset({'California', 'HighPrice'})	frozenset({'HighPoints'})	0.055794077	0.319419598	0.03785861	0.67854174	2.12429589	0.02003689	2.1171657
30	frozenset({'HighPoints'})	frozenset({'California', 'HighPrice'})	0.319419598	0.055794077	0.03785861	0.11852313	2.12429589	0.02003689	1.07116358
31	frozenset({'California'})	frozenset({'HighPoints', 'HighPrice'})	0.294891672	0.111183993	0.03785861	0.12838142	1.15467534	0.00507138	1.01973044

关联规则评价

挖掘结果分析

要求

对挖掘结果进行分析。

分析

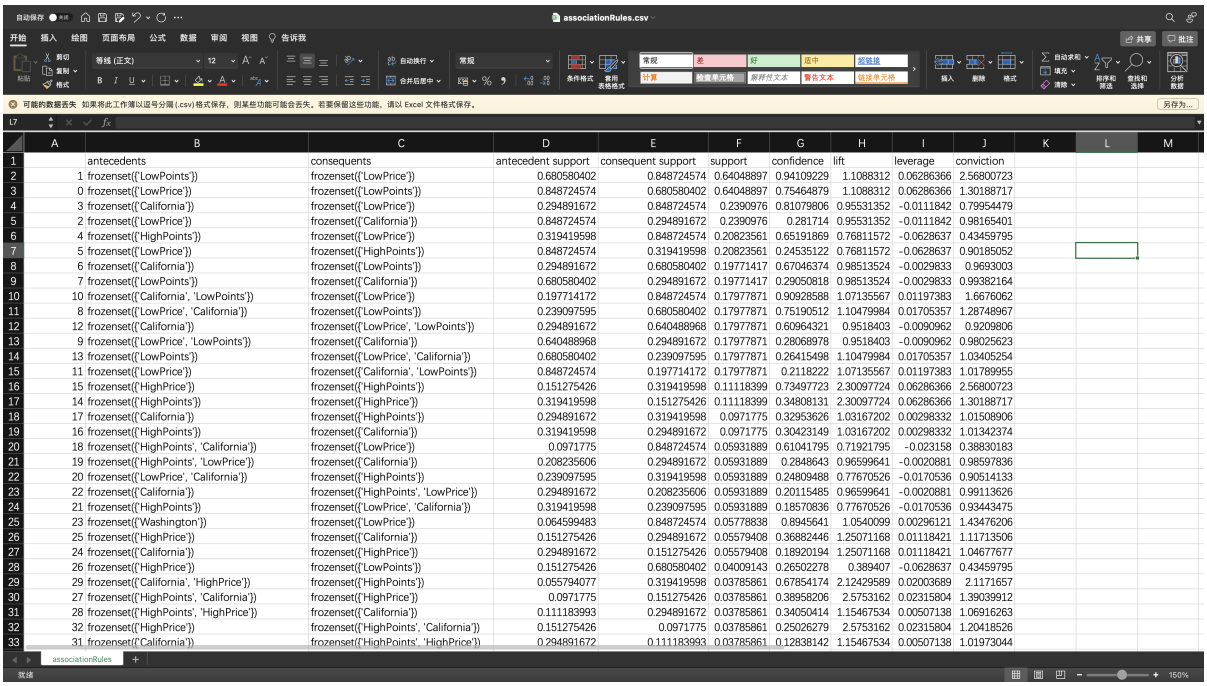
对关联规则按照 support 值进行排序，对得到的部分结果进行分析。

- 1. {LowPoints} -> {LowPrice}: 很显然，分值低的红酒其价格也低。
- xviii. {HighPoints, California} -> {LowPrice}: 加州产的高分红酒可能具有低价格：物美价廉。
- xxii. {Washington} -> {LowPrice}: 华盛顿的红酒可能价格低。

当然也会出现一些不合理的关联规则，例如：

- 4. {LowPoints} -> {HighPrice}: 低分的红酒高价格，这显然是不合理的。出现的原因可能是由于对 **高分数** 和 **高价格** 的划分不合理。

结果展示



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
1	frozenset({'LowPoints'})	frozenset({'LowPrice'})	0.680580402	0.848724574	0.64048897	0.94109229	1.1088312	0.06286366	2.56800723
0	frozenset({'LowPrice'})	frozenset({'LowPoints'})	0.848724574	0.680580402	0.64048897	0.75464879	1.1088312	0.06286366	1.30188717
3	frozenset({'California'})	frozenset({'LowPrice'})	0.294891672	0.848724574	0.2390976	0.81079806	0.95531352	-0.0111842	0.79954479
2	frozenset({'LowPrice'})	frozenset({'California'})	0.848724574	0.294891672	0.2390976	0.281714	0.95531352	-0.0111842	0.98165401
4	frozenset({'HighPoints'})	frozenset({'LowPrice'})	0.319419598	0.848724574	0.20823561	0.65191869	0.76811572	-0.0628637	0.43459795
5	frozenset({'LowPrice'})	frozenset({'HighPoints'})	0.848724574	0.319419598	0.20823561	0.24535122	0.76811572	-0.0628637	0.90185052
6	frozenset({'California'})	frozenset({'LowPoints'})	0.294891672	0.680580402	0.19771417	0.67046374	0.98513524	-0.0029833	0.9693003
7	frozenset({'LowPoints'})	frozenset({'California'})	0.680580402	0.294891672	0.19771417	0.29050818	0.98513524	-0.0029833	0.99382164
10	frozenset({'California', 'LowPoints'})	frozenset({'LowPrice'})	0.197714172	0.848724574	0.17977871	0.90928588	1.07135567	0.01197383	1.6676062
8	frozenset({'LowPrice', 'California'})	frozenset({'LowPoints'})	0.239097595	0.680580402	0.17977871	0.75190512	1.10479984	0.01705357	1.28748967
12	frozenset({'California'})	frozenset({'LowPrice', 'LowPoints'})	0.294891672	0.640488968	0.17977871	0.60964321	0.9518403	-0.0090962	0.9209806
9	frozenset({'LowPrice', 'LowPoints'})	frozenset({'California'})	0.640488968	0.294891672	0.17977871	0.28068978	0.9518403	-0.0090962	0.98025623
13	frozenset({'LowPoints'})	frozenset({'LowPrice', 'California'})	0.680580402	0.239097595	0.17977871	0.26415498	1.10479984	0.01705357	1.03405254
11	frozenset({'LowPrice'})	frozenset({'California', 'LowPoints'})	0.848724574	0.197714172	0.17977871	0.2118222	1.07135567	0.01197383	1.01789955
15	frozenset({'HighPrice'})	frozenset({'HighPoints'})	0.151275426	0.319419598	0.11118399	0.73497723	2.30097724	0.06286366	2.56800723
14	frozenset({'HighPoints'})	frozenset({'HighPrice'})	0.319419598	0.151275426	0.11118399	0.34808131	2.30097724	0.06286366	1.30188717
17	frozenset({'California'})	frozenset({'HighPoints'})	0.294891672	0.319419598	0.0971775	0.32953626	1.03167202	0.00298332	1.01508906
16	frozenset({'HighPoints'})	frozenset({'California'})	0.319419598	0.294891672	0.0971775	0.30423149	1.03167202	0.00298332	1.01342374
18	frozenset({'HighPoints', 'California'})	frozenset({'LowPrice'})	0.0971775	0.848724574	0.05931889	0.61041795	0.71921795	-0.023158	0.38830183
19	frozenset({'HighPoints', 'LowPrice'})	frozenset({'California'})	0.208235606	0.294891672	0.05931889	0.2848643	0.96599641	-0.0020881	0.98597836
20	frozenset({'LowPrice', 'California'})	frozenset({'HighPoints'})	0.239097595	0.319419598	0.05931889	0.24809488	0.77670526	-0.0170536	0.90514133
22	frozenset({'California'})	frozenset({'HighPoints', 'LowPrice'})	0.294891672	0.208235606	0.05931889	0.20115485	0.96599641	-0.0020881	0.99113626
21	frozenset({'HighPoints'})	frozenset({'LowPrice', 'California'})	0.319419598	0.239097595	0.05931889	0.18570836	0.77670526	-0.0170536	0.93443475
23	frozenset({'Washington'})	frozenset({'LowPrice'})	0.064599483	0.848724574	0.05778838	0.8945641	1.0540099	0.00296121	1.43476206
25	frozenset({'HighPrice'})	frozenset({'California'})	0.151275426	0.294891672	0.05579408	0.36882446	1.25071168	0.01118421	1.11713506
24	frozenset({'California'})	frozenset({'HighPrice'})	0.294891672	0.151275426	0.05579408	0.18920194	1.25071168	0.01118421	1.04677677
26	frozenset({'HighPrice'})	frozenset({'LowPoints'})	0.151275426	0.680580402	0.04009143	0.26502278	0.389407	-0.0628637	0.43459795
29	frozenset({'California', 'HighPrice'})	frozenset({'HighPoints'})	0.055794077	0.319419598	0.03785861	0.67854174	2.12429589	0.02003689	2.1171657
27	frozenset({'HighPoints', 'California'})	frozenset({'HighPrice'})	0.0971775	0.151275426	0.03785861	0.38958206	2.5753162	0.02315804	1.39039912
28	frozenset({'HighPoints', 'HighPrice'})	frozenset({'California'})	0.111183993	0.294891672	0.03785861	0.34050414	1.15467534	0.00507138	1.06916263
32	frozenset({'HighPrice'})	frozenset({'HighPoints', 'California'})	0.151275426	0.0971775	0.03785861	0.25026279	2.5753162	0.02315804	1.20418526
31	frozenset({'California'})	frozenset({'HighPoints', 'HighPrice'})	0.294891672	0.111183993	0.03785861	0.12838142	1.15467534	0.00507138	1.01973044

关联规则评价

可视化展示

要求

可视化展示。

思路及代码分析

例如：对关联规则按 lift 值排序

```
# 关联规则可以提升度排序
association_rule.sort_values(by="lift", ascending=False, inplace=True)
print(association_rule)
# 规则是：antecedents->consequents
```

结果展示

```
                 antecedents               consequents  ...  leverage  conviction
105     (HighPoints, Tuscany)               (HighPrice)  ...  0.008482    1.804673
32                (HighPrice)  (HighPoints, California)  ...  0.023158    1.204185
27   (HighPoints, California)               (HighPrice)  ...  0.023158    1.390399
107      (Tuscany, HighPrice)              (HighPoints)  ...  0.007228    3.706883
115                (Burgundy)               (HighPrice)  ...  0.006555    1.370943
..                        ...                       ...  ...       ...         ...
104                (Burgundy)     (LowPrice, LowPoints)  ... -0.006163    0.624757
77                (HighPrice)   (California, LowPoints)  ... -0.011974    0.910201
117     (HighPoints, Tuscany)                (LowPrice)  ... -0.008482    0.285583
75    (California, HighPrice)               (LowPoints)  ... -0.020037    0.470744
26                (HighPrice)               (LowPoints)  ... -0.062864    0.434598

[125 rows x 9 columns]
rule end
```

关联规则可视化，按 lift 值排序