作业二：关联规则挖掘

1.作业要求：

1.对数据集进行处理，转换成适合关联规则挖掘的形式；
2.找出频繁项集；
3.导出关联规则，计算其支持度和置信度;
4.对规则进行评价，可使用 Lift 及其它指标, 要求至少 2 种；
5.对挖掘结果进行可视化展示。

2.作业内容：

2.1 数据预处理

数据集选用作业一中 Wine Reviews 的 winemag-data-130k-v2.csv。

数据集中的属性分为标称属性和数值属性，其中数值属性不适合进行关联规则挖掘，因此只对标称属性进行处理，即对 country，designation，province，variety，winery 这五项属性进行处理。

因此，需要填补这部分的缺失值，利用 null 代替空值。

2.2 找出频繁项集

本次作业是使用 Apriori 算法来寻找频繁集，即支持度大于最小支持度的项集。本次数据挖掘将最小支持度设置为 0.06。具体频繁项集寻找过程如下：

```
Generating itemsets.
 Counting itemsets of length 1.
  Found 54882 candidate itemsets of length 1.
  Found 10 large itemsets of length 1.
 Counting itemsets of length 2.
  Found 45 candidate itemsets of length 2.
  Found 5 large itemsets of length 2.
 Counting itemsets of length 3.
  Found 1 candidate itemsets of length 3.
  Found 1 large itemsets of length 3.
 Counting itemsets of length 4.
  Found 0 candidate itemsets of length 4.
Itemset generation terminated.
```

可以看出项集大小为 1 时，符合条件的数据项有 54882 条；项集大小为 2，符合条件的数据项有 45 条；项集大小为 3，符合条件的数据项有 1 条；其中具体的频繁集内容如下:

```
1 itemsets:
('Cabernet Sauvignon',)
('California',)
('Chardonnay',)
('France',)
('Italy',)
('Pinot Noir',)
('Red Blend',)
('US',)
('Washington',)
('null',)

2 itemsets:
('California', 'US')
('California', 'null')
('Pinot Noir', 'US')
('US', 'Washington')
('US', 'null')

3 itemsets:
('California', 'US', 'null')
```

2.3 导出关联规则

导出的基本关联规则如下:

	Rules	Left	Right	Support	Confidence
0	{US} -> {California}	(US,)	(California,)	0.278882	0.665021
1	{California} -> {US}	(California,)	(US,)	0.278882	1.000000
2	{US} -> {Pinot Noir}	(US,)	(Pinot Noir,)	0.076221	0.181756
3	{Pinot Noir} -> {US}	(Pinot Noir,)	(US,)	0.076221	0.744905
4	{Washington} -> {US}	(Washington,)	(US,)	0.066456	1.000000
5	{US} -> {Washington}	(US,)	(Washington,)	0.066456	0.158470

增加 Lift 指标后的关联规则为：

	Rules	Left	...	Confidence	Lift
0	{US} -> {California}	(US,)	...	0.665021	2.384597
1	{California} -> {US}	(California,)	...	1.000000	2.384597
2	{US} -> {Pinot Noir}	(US,)	...	0.181756	1.776298
3	{Pinot Noir} -> {US}	(Pinot Noir,)	...	0.744905	1.776298
4	{Washington} -> {US}	(Washington,)	...	1.000000	2.384597
5	{US} -> {Washington}	(US,)	...	0.158470	2.384597

增加 Conviction 指标后的关联规则为：

	Rules	Left	...	Lift	Conviction
0	{US} -> {California}	(US,)	...	2.384597	2.152728e+00
1	{California} -> {US}	(California,)	...	2.384597	5.806419e+08
2	{US} -> {Pinot Noir}	(US,)	...	1.776298	1.097078e+00
3	{Pinot Noir} -> {US}	(Pinot Noir,)	...	1.776298	2.276178e+00
4	{Washington} -> {US}	(Washington,)	...	2.384597	5.806419e+08
5	{US} -> {Washington}	(US,)	...	2.384597	1.109342e+00