声明：

hello，小伙伴们~

感谢大家对小鲸鱼的大力支持，目前我在搜集大量经典书籍和资料，其中涉及：

1. 🔥**数学统计**：打好理论基础！
2. 🔥**数据分析**：包括数据分析、商业数据分析等
3. 🔥**数据科学**：数据建模、机器学习、AI框架等
4. 🔥**Python/R cookbook**

12分

还有关于数据结构、算法、数据库等等

以及面经、更多的视频资料整理！

总之，这个仓库会一直更新，分享不会停止！

不收费，但只分享给有需要的人，曾经没人帮我指路，但我不希望大家走太多弯路！

公众号：鲸析

Pandas 数据分析功能一览（第二部分）

基本知识第二弹，包括附加题哦！

本次bonus满分：9分！

注意：代码的输出，是我用标准答案run好的，不要run题目的chunk，输出会消失！另开一个代码块输入你的答案！！！

抽样

In [1]:

```python
# 创建随机数df

import pandas as pd
import numpy as np

df = pd.read_csv('./Iris.csv')
```

In [2]:

```python
df
```

Out[2]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...	...
145	146	6.7	3.0	5.2	2.3	Iris-virginica
146	147	6.3	2.5	5.0	1.9	Iris-virginica
147	148	6.5	3.0	5.2	2.0	Iris-virginica
148	149	6.2	3.4	5.4	2.3	Iris-virginica
149	150	5.9	3.0	5.1	1.8	Iris-virginica

150 rows × 6 columns

简单抽样

In [3]:

```python
# 随机抽取五行
df.sample(n=5)
```

Out[3]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
95	96	5.7	3.0	4.2	1.2	Iris-versicolor
101	102	5.8	2.7	5.1	1.9	Iris-virginica
11	12	4.8	3.4	1.6	0.2	Iris-setosa
42	43	4.4	3.2	1.3	0.2	Iris-setosa
51	52	6.4	3.2	4.5	1.5	Iris-versicolor

In [4]:

```python
# 每一次run这一chunk，都不会变
df.sample(n=5,random_state = 1)
```

Out[4]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
14	15	5.8	4.0	1.2	0.2	Iris-setosa
98	99	5.1	2.5	3.0	1.1	Iris-versicolor
75	76	6.6	3.0	4.4	1.4	Iris-versicolor
16	17	5.4	3.9	1.3	0.4	Iris-setosa
131	132	7.9	3.8	6.4	2.0	Iris-virginica

In [5]:

```python
df.Species.unique()
```

Out[5]:

```
array(['Iris-setosa', 'Iris-versicolor', 'Iris-virginica'], dtype=object)
```

In [7]:

```python
# 加权抽样
sample = df.sample(n = 20,weights = [20]*50+[50]*50+[30]*50,random_state=1024)
sample
```

Out[7]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
94	95	5.6	2.7	4.2	1.3	Iris-versicolor
149	150	5.9	3.0	5.1	1.8	Iris-virginica
81	82	5.5	2.4	3.7	1.0	Iris-versicolor
95	96	5.7	3.0	4.2	1.2	Iris-versicolor
89	90	5.5	2.5	4.0	1.3	Iris-versicolor
108	109	6.7	2.5	5.8	1.8	Iris-virginica
34	35	4.9	3.1	1.5	0.1	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
37	38	4.9	3.1	1.5	0.1	Iris-setosa
99	100	5.7	2.8	4.1	1.3	Iris-versicolor
133	134	6.3	2.8	5.1	1.5	Iris-virginica
74	75	6.4	2.9	4.3	1.3	Iris-versicolor
61	62	5.9	3.0	4.2	1.5	Iris-versicolor
137	138	6.4	3.1	5.5	1.8	Iris-virginica
76	77	6.8	2.8	4.8	1.4	Iris-versicolor
116	117	6.5	3.0	5.5	1.8	Iris-virginica
49	50	5.0	3.3	1.4	0.2	Iris-setosa
119	120	6.0	2.2	5.0	1.5	Iris-virginica
35	36	5.0	3.2	1.2	0.2	Iris-setosa
98	99	5.1	2.5	3.0	1.1	Iris-versicolor

In [8]:

```python
sample.Species.value_counts()
```

Out[8]:

```
Iris-versicolor    9
Iris-virginica     6
Iris-setosa        5
Name: Species, dtype: int64
```

`value_counts()` 可以统计categorical数据的对应类别的size 常用指数：☆☆☆

In [9]:

```python
sample.Species.value_counts()/sample.shape[0] # 求百分比
```

Out[9]:

```
Iris-versicolor    0.45
Iris-virginica     0.30
Iris-setosa        0.25
Name: Species, dtype: float64
```

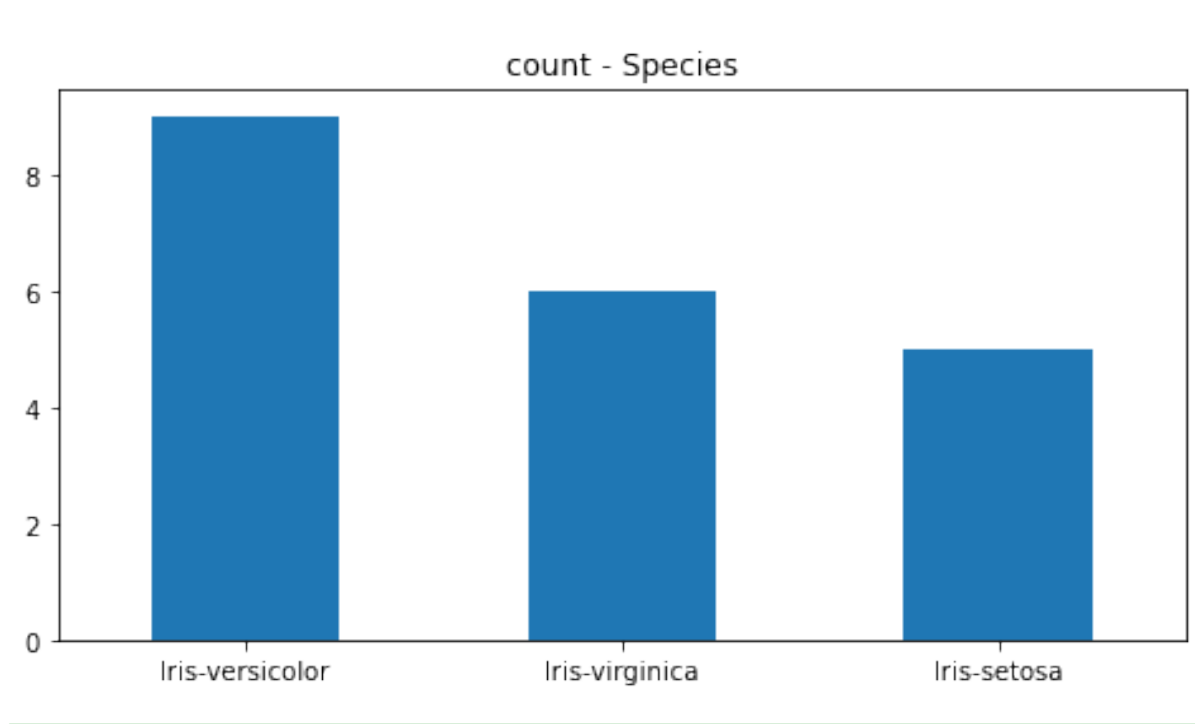
如何更直观的看它们的分布呢？

In [10]:

```python
sample.Species.value_counts().plot(kind = 'bar',
                                   title = 'count - Species',
                                   rot = 0,
                                   figsize = [8,4])
```

Out[10]:

```
<AxesSubplot:title={'center':'count - Species'}>
```



Step 1: 第一题：df这个数据行太多了，我们用不到，所以在每一类别里，随机抽选5条样本吧！seed = 1 1. df新增一列，命名为$Species_{e}ncod \in g$，当Species为setosa，则值为1，versicolor，则值为2，virginica，值为3（不能用if elif）【2分】 2. 根据每个$Species$进行抽样，在每一类别里，随机抽选5条样本，命名为$iris_{s}amp \leq$ 。【3分】

In [11]:

```python
# 不要run，在下面的代码块写答案
```

Out[11]:

	index	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species	Species_encoding
0	27	28	5.2	3.5	1.5	0.2	Iris-setosa	1
1	35	36	5.0	3.2	1.2	0.2	Iris-setosa	1
2	40	41	5.0	3.5	1.3	0.3	Iris-setosa	1
3	38	39	4.4	3.0	1.3	0.2	Iris-setosa	1
4	2	3	4.7	3.2	1.3	0.2	Iris-setosa	1
5	77	78	6.7	3.0	5.0	1.7	Iris-versicolor	2
6	85	86	6.0	3.4	4.5	1.6	Iris-versicolor	2
7	90	91	5.5	2.6	4.4	1.2	Iris-versicolor	2
8	88	89	5.6	3.0	4.1	1.3	Iris-versicolor	2
9	52	53	6.9	3.1	4.9	1.5	Iris-versicolor	2
10	127	128	6.1	3.0	4.9	1.8	Iris-virginica	3
11	135	136	7.7	3.0	6.1	2.3	Iris-virginica	3
12	140	141	6.7	3.1	5.6	2.4	Iris-virginica	3
13	138	139	6.0	3.0	4.8	1.8	Iris-virginica	3
14	102	103	7.1	3.0	5.9	2.1	Iris-virginica	3

In [2]:

```python
#### 统计encoding之前每个类别有多少
df.Species.value_counts()
```

Out[2]:

```
Iris-setosa        50
Iris-versicolor    50
Iris-virginica     50
Name: Species, dtype: int64
```

In [11]:

```python
# 1: apply()
df["Species_encoding"] = df.Species.apply(lambda x: 1 if 'setosa' in x else(2 if 'versicolor' in x else 3))
df["Species_encoding"].value_counts()
```

Out[11]:

```
1    50
2    50
3    50
Name: Species_encoding, dtype: int64
```

In [12]:

```python
# 1: map() - JuD 2/25
species_dict = {'Iris-setosa' : 1,
                'Iris-versicolor' : 2,
                'Iris-virginica' : 3}
df["Species_encoding"] = df['Species'].map(species_dict)
df["Species_encoding"].value_counts()
```

Out[12]:

```
1    50
2    50
3    50
Name: Species_encoding, dtype: int64
```

In [17]:

```python
# 1: loc() - 十二 2/26
df.loc[df['Species'].str.contains("setosa"), "Species_encoding"]= 1
df.loc[df['Species'].str.contains("versicolor"), "Species_encoding"] = 2
df.loc[df['Species'].str.contains("virginica"), "Species_encoding"] = 3
df["Species_encoding"].value_counts()
```

Out[17]:

```
1    50
2    50
3    50
Name: Species_encoding, dtype: int64
```

In [15]:

```python
# 1: np.select - 石刘气泡水 2/26
import numpy as np
condlist = [df['Species'] == 'Iris-setosa',
            df['Species'] == 'Iris-versicolor',
            df['Species'] == 'Iris-virginica']
choicelist = [1,2,3]
df['Species_encoding'] = np.select(condlist, choicelist)
df["Species_encoding"].value_counts()
```

学习了

Out[15]:

```
1    50
2    50
3    50
Name: Species_encoding, dtype: int64
```

In [19]:

```python
# 1: replace - Leeyi 2/26
df['Species_encoding'] = df['Species']
df['Species_encoding'].replace(['Iris-setosa', 'Iris-versicolor', 'Iris-virginica'],
                               [1, 2, 3], inplace=True)
df["Species_encoding"].value_counts()
```

学习了

Out[19]:

```
1    50
2    50
3    50
Name: Species_encoding, dtype: int64
```

In [ ]:

In [23]:

```python
# 2
iris_sample = df.groupby('Species', as_index=False).apply(
    lambda x: x.sample(n=5,random_state=1)).reset_index(level=0, drop=True).reset_index()
iris_sample
```

Out[23]:

	index	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species	Species_encoding
0	27	28	5.2	3.5	1.5	0.2	Iris-setosa	1
1	35	36	5.0	3.2	1.2	0.2	Iris-setosa	1
2	40	41	5.0	3.5	1.3	0.3	Iris-setosa	1
3	38	39	4.4	3.0	1.3	0.2	Iris-setosa	1
4	2	3	4.7	3.2	1.3	0.2	Iris-setosa	1
5	77	78	6.7	3.0	5.0	1.7	Iris-versicolor	2
6	85	86	6.0	3.4	4.5	1.6	Iris-versicolor	2
7	90	91	5.5	2.6	4.4	1.2	Iris-versicolor	2
8	88	89	5.6	3.0	4.1	1.3	Iris-versicolor	2
9	52	53	6.9	3.1	4.9	1.5	Iris-versicolor	2
10	127	128	6.1	3.0	4.9	1.8	Iris-virginica	3
11	135	136	7.7	3.0	6.1	2.3	Iris-virginica	3
12	140	141	6.7	3.1	5.6	2.4	Iris-virginica	3
13	138	139	6.0	3.0	4.8	1.8	Iris-virginica	3
14	102	103	7.1	3.0	5.9	2.1	Iris-virginica	3

In [ ]:

分组【groupby】上

那么如何根据类别对数据分组？

groupby函数可以帮我们解决问题！

什么是groupby? 可以总结为3个步骤：

1. Splitting

根据分组标准，把数据拆分。

1. Applying

分组之后，目的是通过

- 聚合函数
- 转换函数
- 筛选函数

对组内数据进行操作（有实例）

1. Combining

今天这一期，我们只讲【Splitting】的部分！

注意： 没有完成上一题目，无法向下进行哦，请务必完成上题，得到 iris_sample

In [24]:

```python
iris_sample.groupby('Species')  # 会得到一个DataFrameGroupBy对象
```

Out[24]:

```
<pandas.core.groupby.generic.DataFrameGroupBy object at 0x7ff99e9fa640>
```

In [25]:

```python
sample_group = iris_sample.groupby('Species')
```

In [14]:

```python
list(sample_group) # 用list查看
```

Out[14]:

```
[('Iris-setosa',
     index  Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm  \
  0     27  28            5.2           3.5            1.5           0.2
  1     35  36            5.0           3.2            1.2           0.2
  2     40  41            5.0           3.5            1.3           0.3
  3     38  39            4.4           3.0            1.3           0.2
  4      2   3            4.7           3.2            1.3           0.2

         Species  Species_encoding
  0  Iris-setosa                 1
  1  Iris-setosa                 1
  2  Iris-setosa                 1
  3  Iris-setosa                 1
  4  Iris-setosa                 1  ),
 ('Iris-versicolor',
     index  Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm  \
  5     77  78            6.7           3.0            5.0           1.7
  6     85  86            6.0           3.4            4.5           1.6
  7     90  91            5.5           2.6            4.4           1.2
  8     88  89            5.6           3.0            4.1           1.3
  9     52  53            6.9           3.1            4.9           1.5

             Species  Species_encoding
  5  Iris-versicolor                 2
  6  Iris-versicolor                 2
  7  Iris-versicolor                 2
  8  Iris-versicolor                 2
  9  Iris-versicolor                 2  ),
 ('Iris-virginica',
      index   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm  \
  10    127  128            6.1           3.0            4.9           1.8
  11    135  136            7.7           3.0            6.1           2.3
  12    140  141            6.7           3.1            5.6           2.4
  13    138  139            6.0           3.0            4.8           1.8
  14    102  103            7.1           3.0            5.9           2.1

             Species  Species_encoding
  10  Iris-virginica                 3
  11  Iris-virginica                 3
  12  Iris-virginica                 3
  13  Iris-virginica                 3
  14  Iris-virginica                 3  )]
```

In [15]:

```python
# .ngroups 查看有几组
sample_group.ngroups
```

Out[15]:

```
3
```

In [16]:

```python
# .size() 查看组内容量
sample_group.size()
```

Out[16]:

```
Species
Iris-setosa        5
Iris-versicolor    5
Iris-virginica     5
dtype: int64
```

Step 2: 第二题：基于$samp \leq_{g} roup$ 1. 返回一个df，叫first，为每个group的第一行数据合并【1分】 2. 返回一个df，叫last，为每个group的最后一行数据合并 【1分】

In [17]:

```python
# first
```

Out[17]:

	index	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species_encoding
Species							
Iris-setosa	27	28	5.2	3.5	1.5	0.2	1
Iris-versicolor	77	78	6.7	3.0	5.0	1.7	2
Iris-virginica	127	128	6.1	3.0	4.9	1.8	3

In [9]:

```python
first = sample_group.first()
first
```

Out[9]:

	index	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species_encoding
Species							
Iris-setosa	27	28	5.2	3.5	1.5	0.2	1
Iris-versicolor	77	78	6.7	3.0	5.0	1.7	2
Iris-virginica	127	128	6.1	3.0	4.9	1.8	3

In [18]:

```python
# last
```

Out[18]:

	index	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species_encoding
Species							
Iris-setosa	2	3	4.7	3.2	1.3	0.2	1
Iris-versicolor	52	53	6.9	3.1	4.9	1.5	2
Iris-virginica	102	103	7.1	3.0	5.9	2.1	3

In [10]:

```python
last = sample_group.last()
last
```

Out[10]:

	index	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species_encoding
Species							
Iris-setosa	2	3	4.7	3.2	1.3	0.2	1
Iris-versicolor	52	53	6.9	3.1	4.9	1.5	2
Iris-virginica	102	103	7.1	3.0	5.9	2.1	3

In [19]:

```python
# 根据分组的名称取出对应数据
sample_group.get_group('Iris-setosa')
```

Out[19]:

	index	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species	Species_encoding
0	27	28	5.2	3.5	1.5	0.2	Iris-setosa	1
1	35	36	5.0	3.2	1.2	0.2	Iris-setosa	1
2	40	41	5.0	3.5	1.3	0.3	Iris-setosa	1
3	38	39	4.4	3.0	1.3	0.2	Iris-setosa	1
4	2	3	4.7	3.2	1.3	0.2	Iris-setosa	1

Step 3: 第三题：基于$samp \leq_{g} roup$【提示：filter】 请问那组数据的$SepalWhCm$的最大值大于3.2，并且$P\eta lWhCm$的最小值小于1【2分】

In [20]:

Out[20]:

	index	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species	Species_encoding
0	27	28	5.2	3.5	1.5	0.2	Iris-setosa	1
1	35	36	5.0	3.2	1.2	0.2	Iris-setosa	1
2	40	41	5.0	3.5	1.3	0.3	Iris-setosa	1
3	38	39	4.4	3.0	1.3	0.2	Iris-setosa	1
4	2	3	4.7	3.2	1.3	0.2	Iris-setosa	1

In [26]:

```python
sample_group.filter(lambda x: (x['SepalWidthCm'].max() > 3.2)
                    & (x['PetalWidthCm'].min() < 1))
```

Out[26]:

	index	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species	Species_encoding
0	27	28	5.2	3.5	1.5	0.2	Iris-setosa	1
1	35	36	5.0	3.2	1.2	0.2	Iris-setosa	1
2	40	41	5.0	3.5	1.3	0.3	Iris-setosa	1
3	38	39	4.4	3.0	1.3	0.2	Iris-setosa	1
4	2	3	4.7	3.2	1.3	0.2	Iris-setosa	1

小结

今天就分享到这里，总结一下！

1. 抽样以及具体参数用法
2. groupby的Splitting以及group的基本manipulation


缺氧的小鲸鱼🐳

小红书号：1821360719

🎓美国乔治城大学理学硕士

✈️中国商飞数据管理中心数据分析师...

扫描二维码
在小红书找到我