

```
import numpy as np #用来调用包
import pandas as pd
```

```
!pip install pandasql #安装sql函数
```

```
Collecting pandasql
  Downloading pandasql-0.7.3.tar.gz (26 kB)
  Preparing metadata (setup.py) ... done
Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist-packages (from pandasql) (1.25.2)
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages (from pandasql) (1.5.3)
Requirement already satisfied: sqlalchemy in /usr/local/lib/python3.10/dist-packages (from pandasql) (2.0.28)
Requirement already satisfied: python-dateutil>=2.8.1 in /usr/local/lib/python3.10/dist-packages (from pandas->pandasql) (2.
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas->pandasql) (2023.4)
Requirement already satisfied: typing-extensions>=4.6.0 in /usr/local/lib/python3.10/dist-packages (from sqlalchemy->pandasq
Requirement already satisfied: greenlet!=0.4.17 in /usr/local/lib/python3.10/dist-packages (from sqlalchemy->pandasql) (3.0.
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.1->pandas->pan
Building wheels for collected packages: pandasql
  Building wheel for pandasql (setup.py) ... done
  Created wheel for pandasql: filename=pandasql-0.7.3-py3-none-any.whl size=26771 sha256=5c233116dd4078b19449523246aaf4095ce
  Stored in directory: /root/.cache/pip/wheels/e9/bc/3a/8434bdccc5f779e72894a9b24fecbdca9f97940607eaf4bcd9
Successfully built pandasql
Installing collected packages: pandasql
Successfully installed pandasql-0.7.3
```

```
a=np.array
b=np.arange
transpose是指转置
axis=0就是指的一列, axis=1就是指的一行
```

```
import numpy as np #产生一个随机数
print(np.random.rand(2,2))
print(np.random.randint(0,10)) #在0到10之间产生一个随机数
```

```
[[0.36633964 0.5075148 ]
 [0.27094587 0.49434128]]
4
```

```
import numpy as np
print(np.random.choice(['a','e','i','o','u'],size=5)) #还可以用p给随机一个权重

['e' 'o' 'o' 'o' 'o']
```

```
url='https://www2.deloitte.com/us/en.html'
iris =pd.read_html(url)
print(iris.head())
```

```
-----
ValueError                                Traceback (most recent call last)
<ipython-input-21-e28041f40f45> in <cell line: 2>()
    1 url='https://www2.deloitte.com/us/en.html'
----> 2 iris =pd.read_html(url)
    3 print(iris.head())
```

```
----- 5 frames -----
/usr/local/lib/python3.10/dist-packages/pandas/io/html.py in _parse_tables(self, doc, match, attrs)
    616
    617         if not tables:
--> 618             raise ValueError("No tables found")
    619
    620         result = []
```

```
ValueError: No tables found
```

```
url="https://gist.githubusercontent.com/curran/a08a1080b88344b0c8a7/raw/d546eae765268bf2f487608c537c05e22e4b221/iris.csv"
iris = pd.read_csv(url)#改用别的文件试试能不能跑出来
print(iris.shape[0])
iris_idx = np.arange(iris.shape[0])
num_training = int(np.floor(0.75*iris.shape[0]))
print(num_training)
iris_train_idx = np.random.choice(iris_idx,size = num_training,replace = False)
iris_test_idx = np.delete(iris_idx,iris_train_idx)

150
112
```

```
url = 'spotify_top_songs_audio_features.csv'
spotify = pd.read_csv(url)
print(spotify.shape[0])
print(spotify.shape[1])
spotify_idx = np.arange(spotify.shape[0])
num_training = int(np.floor(0.75*spotify.shape[0])) #这一步是为了选择训练数据
print(num_training)

6513
19
4884
```

```
url = '/content/stocks.csv'
stock = pd.read_csv(url)
print(stock.shape[0])
print(stock.shape[1])
stock_idx = np.arange(stock.shape[0])
num_training = int(np.floor(0.75*stock.shape[0]))
print(num_training)

248
8
186
```

```
iris.head()
```

| | sepal_length | sepal_width | petal_length | petal_width | species |
|---|--------------|-------------|--------------|-------------|---------|
| 0 | 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 1 | 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| 2 | 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 3 | 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 4 | 5.0 | 3.6 | 1.4 | 0.2 | setosa |

```
spotify.head()
```

| | id | artist_names | track_name | source | key | mode | time_signature | danceability | energy | sp |
|---|-------------------------|-------------------------------|--|-------------------------|-------|-------|----------------|--------------|--------|----|
| 0 | 000xQL6tZNLJzIrtlgxqSI | ZAYN, PARTYNEXTDOOR | Still Got Time (feat. PARTYNEXTDOOR) | RCA Records Label | G | Major | 4 beats | 0.748 | 0.627 | |
| 1 | 003eolwxETJujVWmNFMoZy | Alessia Cara | Growing Pains | Def Jam Recordings | C#/Db | Minor | 4 beats | 0.353 | 0.755 | |
| 2 | 003vvx7Niy0yvvhvHt4a68B | The Killers | Mr. Brightside | Island Records | C#/Db | Major | 4 beats | 0.352 | 0.911 | |
| 3 | 00B7TZ0Xawar6NZ00JFomN | Cardi B, Chance the Rapper | Best Life (feat. Chance The Rapper) | Atlantic/KSR | A | Major | 4 beats | 0.620 | 0.625 | |
| 4 | 00Blm7zeNqgYLPtW6zg8cj | Post Malone, The Weeknd | One Right Now (with The Weeknd) | Republic Records | C#/Db | Major | 4 beats | 0.687 | 0.781 | |

```
stock.head()
```

| | Ticker | Date | Open | High | Low | Close | Adj Close | Volume |
|---|--------|------------|------------|------------|------------|------------|------------|----------|
| 0 | AAPL | 2023-02-07 | 150.639999 | 155.229996 | 150.639999 | 154.649994 | 154.414230 | 83322600 |
| 1 | AAPL | 2023-02-08 | 153.880005 | 154.580002 | 151.169998 | 151.919998 | 151.688400 | 64120100 |
| 2 | AAPL | 2023-02-09 | 153.779999 | 154.330002 | 150.419998 | 150.869995 | 150.639999 | 56007100 |
| 3 | AAPL | 2023-02-10 | 149.460007 | 151.339996 | 149.220001 | 151.009995 | 151.009995 | 57450700 |
| 4 | AAPL | 2023-02-13 | 150.949997 | 154.259995 | 150.919998 | 153.850006 | 153.850006 | 62199000 |

```
iris.shape[0]
```

150

```
spotify.shape[0]
```

6513

```
stock.shape[0]
```

248

```
import numpy as np
a = np.random.RandomState(60)
print(type(a))
a.randint(1000)

<class 'numpy.random.mtrand.RandomState'>
205
```

```
a.randint(100)#直接出100以内的随机数
```

99

```
s=pd.Series([1,3,5,6,8])
print(type(s))
s

<class 'pandas.core.series.Series'>
0    1
1    3
2    5
3    6
4    8
dtype: int64
```

```
d = pd.DataFrame({'coll':[1,2,3,4,5,6], 'col2':['1','2','3','4','5','6']})
print(d)
```

| | coll | col2 |
|---|------|------|
| 0 | 1 | 1 |
| 1 | 2 | 2 |
| 2 | 3 | 3 |
| 3 | 4 | 4 |
| 4 | 5 | 5 |
| 5 | 6 | 6 |

```
d.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 2 columns):
#   Column  Non-Null Count  Dtype
---  ---
0    coll    6 non-null      int64
1    col2     6 non-null      object
dtypes: int64(1), object(1)
memory usage: 224.0+ bytes
```

