- 需求：
  - 导入文件，查看原始数据
  - 将人口数据和各州简称数据进行合并
  - 将合并的数据中重复的abbreviation列进行删除
  - 查看存在缺失数据的列
  - 找到有哪些state/region使得state的值为NaN，进行去重操作
  - 为找到的这些state/region的state项补上正确的值，从而去除掉state这一列的所有NaN
  - 合并各州面积数据areas
  - 我们会发现area(sq.mi)这一列有缺失数据，找出是哪些行
  - 去除含有缺失数据的行
  - 找出2010年的全民人口数据
  - 计算各州的人口密度
  - 排序，并找出人口密度最高的五个州 df.sort_values()

In [1]:
```
import numpy as np
import pandas as pd
from pandas import Series,DataFrame
```

In [2]:
```
abb = pd.read_csv('./data/state-abbrevs.csv')
pop = pd.read_csv('./data/state-population.csv')
area = pd.read_csv('./data/state-areas.csv')
```

In [3]:
```
abb.head() # abb各州简称
```

Out[3]:

	state	abbreviation
0	Alabama	AL
1	Alaska	AK
2	Arizona	AZ
3	Arkansas	AR
4	California	CA

In [4]:
```
pop.head() # 人口
```

Out[4]:

	state/region	ages	year	population
0	AL	under18	2012	1117489.0
1	AL	total	2012	4817528.0
2	AL	under18	2010	1130966.0
3	AL	total	2010	4785570.0
4	AL	under18	2011	1125763.0

In [5]:
```
area.head() # 面积
```

Out[5]:

	state	area (sq. mi)
0	Alabama	52423
1	Alaska	656425
2	Arizona	114006
3	Arkansas	53182
4	California	163707

In [6]:
```
# 将人口数据和各州简称数据进行合并
abb_pop = pd.merge(abb,pop,how='outer',left_on='abbreviation',right_on='state/region')
abb_pop.head()
```

Out[6]:

	state	abbreviation	state/region	ages	year	population
0	Alabama	AL	AL	under18	2012	1117489.0
1	Alabama	AL	AL	total	2012	4817528.0
2	Alabama	AL	AL	under18	2010	1130966.0
3	Alabama	AL	AL	total	2010	4785570.0
4	Alabama	AL	AL	under18	2011	1125763.0

In [7]:
```
# 将合并的数据中重复的abbreviation列进行删除
# (合并后  删除abbreviation这一列)
abb_pop.drop(labels='abbreviation',axis=1,inplace=True)
abb_pop.head()
```

Out[7]:

	state	state/region	ages	year	population
0	Alabama	AL	under18	2012	1117489.0
1	Alabama	AL	total	2012	4817528.0
2	Alabama	AL	under18	2010	1130966.0
3	Alabama	AL	total	2010	4785570.0
4	Alabama	AL	under18	2011	1125763.0

In [8]:
```
# 查看存在缺失数据的列
abb_pop.isnull().any(axis=0)
```

Out[8]:
```
state            True
state/region    False
ages            False
year            False
population       True
dtype: bool
```

In [9]:
```
# 找到有哪些state/region使得state的值为NaN，进行去重操作
# (找到state为空值的行,对 state/region 进行去重)
con = abb_pop['state'].isnull()
abb_pop[con]['state/region'].unique()
```

Out[9]:
```
array(['PR', 'USA'], dtype=object)
```

In [10]:
```
# 为找到的这些state/region的state项补上正确的值，从而去除掉state这一列的所有NaN

# 找出 state/region 列中 为PR的 行索引
indexs = abb_pop.loc[abb_pop['state/region'] == 'PR'].index

abb_pop.loc[indexs,'state'] = 'PPPP_RRRR' # 空值 填为 PPPP_RRRR
```

In [11]:
```
abb_pop.loc[abb_pop['state/region'] == 'PR'].head()# 查看是否赋值
```

Out[11]:

	state	state/region	ages	year	population
2448	PPPP_RRRR	PR	under18	1990	NaN
2449	PPPP_RRRR	PR	total	1990	NaN
2450	PPPP_RRRR	PR	total	1991	NaN
2451	PPPP_RRRR	PR	under18	1991	NaN
2452	PPPP_RRRR	PR	total	1993	NaN

In [12]:
```
# USA 同理
indexs = abb_pop.loc[abb_pop['state/region']=='USA'].index
abb_pop.loc[indexs,'state'] = 'UUUUSA'
abb_pop.loc[abb_pop['state/region']=='USA'].head()
```

Out[12]:

	state	state/region	ages	year	population
2496	UUUUSA	USA	under18	1990	64218512.0
2497	UUUUSA	USA	total	1990	249622814.0
2498	UUUUSA	USA	total	1991	252980942.0
2499	UUUUSA	USA	under18	1991	65313018.0
2500	UUUUSA	USA	under18	1992	66509177.0

In [13]:
```
# 合并各州面积数据areas （三个合并）
abb_pop.head()
```

Out[13]:

	state	state/region	ages	year	population
0	Alabama	AL	under18	2012	1117489.0
1	Alabama	AL	total	2012	4817528.0
2	Alabama	AL	under18	2010	1130966.0
3	Alabama	AL	total	2010	4785570.0
4	Alabama	AL	under18	2011	1125763.0

In [14]:
```
area.head()
```

Out[14]:

	state	area (sq. mi)
0	Alabama	52423
1	Alaska	656425
2	Arizona	114006
3	Arkansas	53182
4	California	163707

In [15]:
```
# 合并各州面积数据areas （三个合并）
abb_pop_area = pd.merge(abb_pop,area,how='outer')
abb_pop_area.head()
```

Out[15]:

	state	state/region	ages	year	population	area (sq. mi)
0	Alabama	AL	under18	2012.0	1117489.0	52423.0
1	Alabama	AL	total	2012.0	4817528.0	52423.0
2	Alabama	AL	under18	2010.0	1130966.0	52423.0
3	Alabama	AL	total	2010.0	4785570.0	52423.0
4	Alabama	AL	under18	2011.0	1125763.0	52423.0

In [16]:
```
# - 我们会发现area(sq.mi)这一列有缺失数据，找出是哪些行
indexs= abb_pop_area.loc[abb_pop_area['area (sq. mi)'].isnull()].index
indexs
```

Out[16]:
```
Int64Index([2448, 2449, 2450, 2451, 2452, 2453, 2454, 2455, 2456, 2457, 2458,
            2459, 2460, 2461, 2462, 2463, 2464, 2465, 2466, 2467, 2468, 2469,
            2470, 2471, 2472, 2473, 2474, 2475, 2476, 2477, 2478, 2479, 2480,
            2481, 2482, 2483, 2484, 2485, 2486, 2487, 2488, 2489, 2490, 2491,
            2492, 2493, 2494, 2495, 2496, 2497, 2498, 2499, 2500, 2501, 2502,
            2503, 2504, 2505, 2506, 2507, 2508, 2509, 2510, 2511, 2512, 2513,
            2514, 2515, 2516, 2517, 2518, 2519, 2520, 2521, 2522, 2523, 2524,
            2525, 2526, 2527, 2528, 2529, 2530, 2531, 2532, 2533, 2534, 2535,
            2536, 2537, 2538, 2539, 2540, 2541, 2542, 2543],
           dtype='int64')
```

In [17]:
```
#  - 去除含有缺失数据的行
# abb_pop_area.dropna(axis=0)
# 删除 上面取出的行
abb_pop_area.drop(labels=indexs,axis=0,inplace=True)
abb_pop_area.head()
```

Out[17]:

	state	state/region	ages	year	population	area (sq. mi)
0	Alabama	AL	under18	2012.0	1117489.0	52423.0
1	Alabama	AL	total	2012.0	4817528.0	52423.0
2	Alabama	AL	under18	2010.0	1130966.0	52423.0
3	Alabama	AL	total	2010.0	4785570.0	52423.0
4	Alabama	AL	under18	2011.0	1125763.0	52423.0

In [18]:
```
#- 找出2010年的全民人口数据
abb_pop_area.query('year == 2010 & ages == "total"').head()
```

Out[18]:

	state	state/region	ages	year	population	area (sq. mi)
3	Alabama	AL	total	2010.0	4785570.0	52423.0
91	Alaska	AK	total	2010.0	713868.0	656425.0
101	Arizona	AZ	total	2010.0	6408790.0	114006.0
189	Arkansas	AR	total	2010.0	2922280.0	53182.0
197	California	CA	total	2010.0	37333601.0	163707.0

In [19]:
```
# - 计算各州的人口密度
abb_pop_area['midu'] = abb_pop_area['population']/abb_pop_area['area (sq. mi)']
abb_pop_area.head()
```

Out[19]:

	state	state/region	ages	year	population	area (sq. mi)	midu
0	Alabama	AL	under18	2012.0	1117489.0	52423.0	21.316769
1	Alabama	AL	total	2012.0	4817528.0	52423.0	91.897221
2	Alabama	AL	under18	2010.0	1130966.0	52423.0	21.573851
3	Alabama	AL	total	2010.0	4785570.0	52423.0	91.287603
4	Alabama	AL	under18	2011.0	1125763.0	52423.0	21.474601

In [20]:
```
# - 排序，并找出人口密度最高的五个州   df.sort_values()
abb_pop_area.sort_values(by='midu',axis=0,ascending=False).head()
```

Out[20]:

	state	state/region	ages	year	population	area (sq. mi)	midu
391	District of Columbia	DC	total	2013.0	646449.0	68.0	9506.602941
385	District of Columbia	DC	total	2012.0	633427.0	68.0	9315.102941
387	District of Columbia	DC	total	2011.0	619624.0	68.0	9112.117647
431	District of Columbia	DC	total	1990.0	605321.0	68.0	8901.779412
389	District of Columbia	DC	total	2010.0	605125.0	68.0	8898.897059

In [ ]: