美国 1880-2010 年的新出生婴儿情况之 python 分析

➢ 数据理解

本次分析的数据来自美国社会保障局（SSA）

http://www.ssa.gov/oact/babynames/limits.html 。美国社会保障局每年提供一份数据文件，现有从 1880 年到 2018 年（共 139 份 TEXT 文件）出生的美国婴儿姓名及性别数据。其中，每份文档的命名格式为’yob’+年份，包含的数据信息为婴儿姓名，婴儿性别，出生数量。

➢ 分析目的

✧ 比较男孩和女孩出生数随时间的变化。

✧ 比较男孩和女孩名字随时间的变化。

✧ 统计随时间的变化，名字开头字母的变化。

✧ 查看每年都在用的名字。

➢ 分析步骤

✧ 数据导入

代码如下：

Input：TEXT 文件存储路径。

```
def load2(path):
    years = range(1880, 2019)
    result = []
    for year in years:
        frame = pd.read_csv(path + '/' + f'yob{year}.txt',
                            names=['name', 'sex', 'births'])
        frame['year'] = year
        result.append(frame)
    df = pd.concat(result, ignore_index=True)
    return df
```

Out：

```
              name sex  births  year
0             Mary   F    7065  1880
1             Anna   F    2604  1880
2             Emma   F    2003  1880
3        Elizabeth   F    1939  1880
4           Minnie   F    1746  1880
...            ...  ..     ...   ...
1957041      Zylas   M       5  2018
1957042      Zyran   M       5  2018
1957043      Zyrie   M       5  2018
1957044      Zyron   M       5  2018
1957045      Zzyzx   M       5  2018

[1957046 rows x 4 columns]
```

结论：
从导入的结果来看，该数据一共包含四个字段：name，sex，births，year。

✧ 数据清洗

代码如下：

```
print(raw_data.info())
print()
print(raw_data.count())
```

Out：

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1957046 entries, 0 to 1957045
Data columns (total 4 columns):
name      object
sex       object
births    int64
year      int64
dtypes: int64(2), object(2)
memory usage: 44.8+ MB
None

name      1957046
sex       1957046
births    1957046
year      1957046
dtype: int64
```

结论：从输出结果可以看出，该数据中无无关数据、缺失数据以及脏数据，可以进行后续的分析。

✧ 数据分析

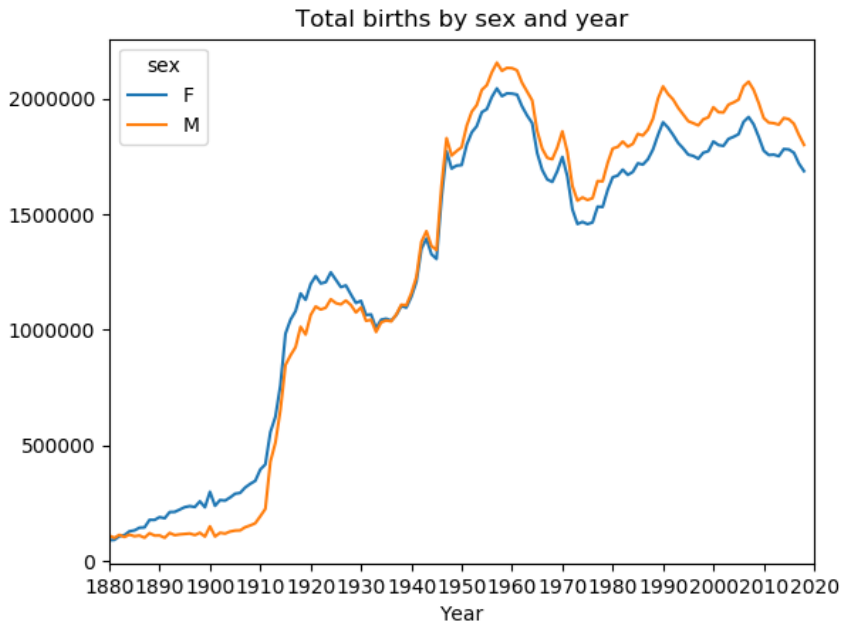
- 比较男孩和女孩出生数随时间的变化

代码如下：

Input：清洗后的数据。

```
def sex_births(rawdata):
    sex_birth_num = pd.pivot_table(rawdata, index='year', columns='sex',
                                   values='births', aggfunc='sum')
    sex_birth_num.plot(title='Total births by sex and year',
                       xticks=range(1880, 2025, 10))
    plt.xlabel('Year')
    plt.ylabel('Births')
    plt.savefig('0 Total births by sex and year.png',
                )
    return sex_birth_num
```

Out：



结论：

① 从图上的趋势来看，从 1880 年至 1960 年，美国婴儿的出生数量随时间的变化呈增长趋势，之后，出生数量呈下降趋势至 1980 年。最后，变化趋于平缓。

② 从 1880 年至 1930 年，男孩的出生数量小于女孩的出生数量；但 1950 年之后，男婴的出生数量反而大于女婴的数量。

- 比较男孩和女孩名字随时间的变化

代码如下：

Input：便于比较，增加了一列数据（每个孩子名字相对于出生总数的比例 prop = births / total_births）。取每年出生数量最多的前 1000 行数据，然后计算总比例，以及查找有多少名字是最受欢迎的的那 50%。

```python
def add_prop(data):
    data['prop'] = data.births / data.births.sum()
    return data


def addprop(rawdata):
    rawdata = rawdata.groupby(['year', 'sex']).apply(add_prop)
    return rawdata
```

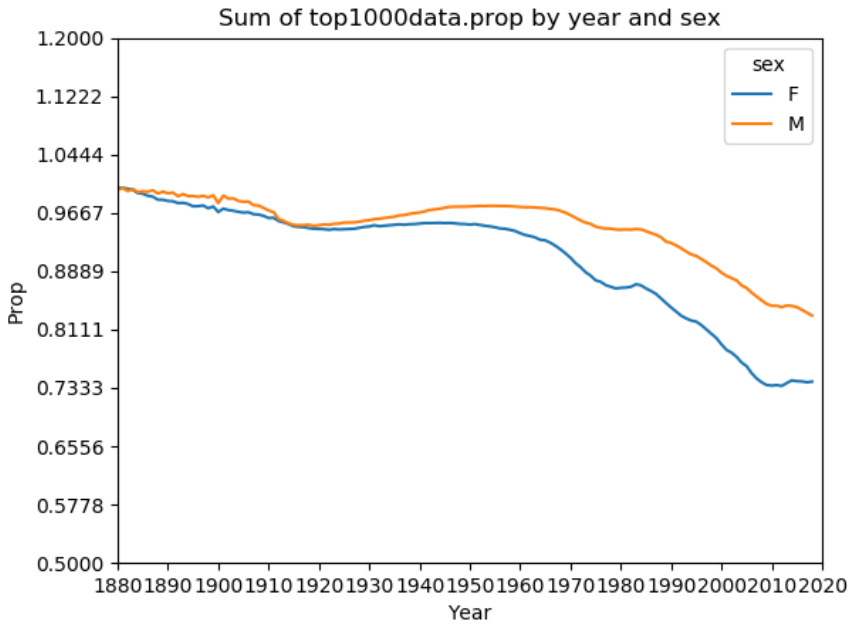
```python
def count_number(data, q=0.5):
    data = data.sort_values(by='prop', ascending=False)
    position = data.prop.cumsum().values.searchsorted(q) + 1
    return position
```

```python
def get_top1000(rawdata):
    top1000_list = []
    for year, group in rawdata.groupby(['year', 'sex']):
        top1000_list.append(
            group.sort_values(by='births', ascending=False)[:1000])

    top1000_df = pd.concat(top1000_list, ignore_index=True)
    return top1000_df
```

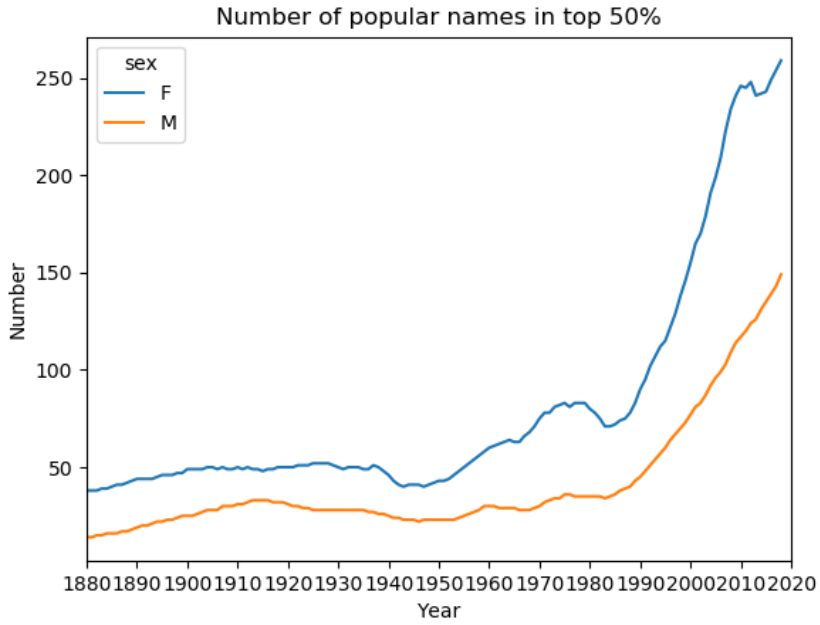
```python
def name_diversity(top1000data):
    table = pd.pivot_table(top1000data, index='year', columns='sex',
                           aggfunc='sum', values='prop')  # Top1000 中 sum(prop)
    table.plot(title='Sum of top1000data.prop by year and sex',
               yticks=np.linspace(0.3, 1.2, 10), xticks=range(1880, 2025, 10))
    plt.xlabel('Year')
    plt.ylabel('Prop')
    plt.savefig('1 Sum of top1000data.prop by year and sex.png')
    number_name_q = top1000data.groupby(['year', 'sex']).apply(
        count_number)  # Top1000 中 sum(prop) = 50% 的 num(name)
    numbername = number_name_q.unstack('sex')
    numbername.plot(title='Number of popular names in top 50%',
                    xticks=range(1880, 2025, 10))
    plt.xlabel('Year')
    plt.ylabel('Number')
    plt.savefig('2 Number of popular names in top 50%.png')

    return table, number_name_q
```

Out：

图一：



图二：



结论：

① 从图一的结果来看，随着时间的变化，Top1000 名字的总比例下降，表明家长给孩子取名字更趋向于多样化。

② 从图二的结果来看，随着时间的变化，前 50% 最受欢迎的 name 数量呈增长趋势，且女孩子的数量一直大于男孩子，增长的趋势也更加明显，说明女孩子的名字一直都比男孩子的名字更加多样化，且多样化的程度更大。

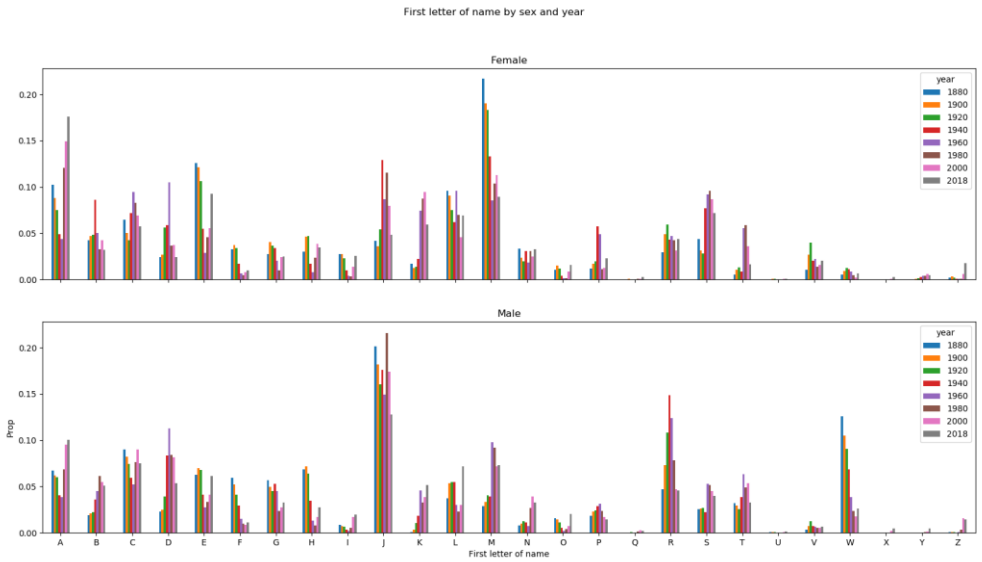
- 统计随时间的变化，名字开头字母的变化

代码如下：

Input：清洗后的数据。

```
def get_first_letters(rawdata):
    rawdata['first_letter'] = rawdata.name.map(lambda x: x[0])

    table = pd.pivot_table(rawdata, index='first_letter',
                           columns=['sex', 'year'], values='births',
                           aggfunc='sum')
    letter_prop_table = table / table.sum()
    letter_prop_year = letter_prop_table.reindex(
        columns=[1880, 1900, 1920, 1940, 1960, 1980, 2000, 2018], level='year')

    fig, axes = plt.subplots(2, 1, sharex=True, sharey=True, figsize=(20, 10))
    letter_prop_year['F'].plot(kind='bar', rot=0, ax=axes[0], title='Female')

    letter_prop_year['M'].plot(kind='bar', rot=0, ax=axes[1], title='Male')
    plt.xlabel('First letter of name')
    plt.ylabel('Prop')
    plt.suptitle('First letter of name by sex and year')
    plt.savefig('4 First letter of name by sex and year.png')
    plt.show()
```

Out：



结论：

① 男孩和女孩基本更喜欢用取以 A-P、R-T 开头的名字，不喜欢以 Q、X-Z 开头的名字。

② 以 M 开头的女孩名字一直都较受欢迎。从 20 世纪 60 年代以来，以 A 开头的女孩名

字经历了显著的增长。
③ 以 J 开头的男孩名字一直都较受欢迎。以 W 开头的男孩名字经历了显著的降低。

- 查看每年都在用的名字

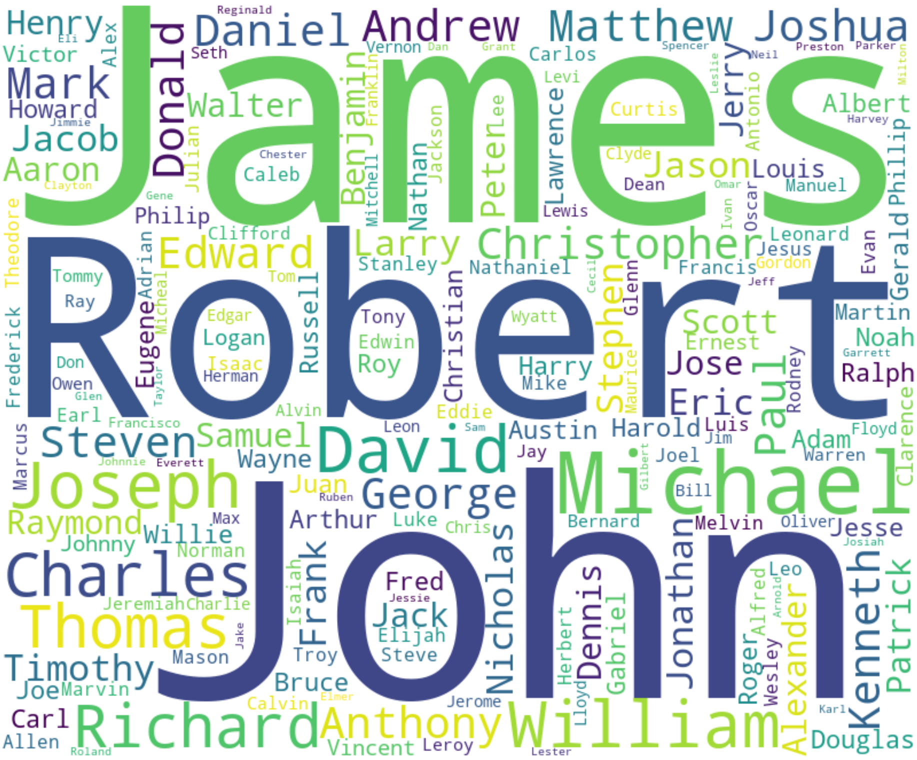
代码如下：
Input：清洗后的数据。

```
def nochangename(rawdata):
    name_pivot = pd.pivot_table(rawdata, index='year', columns=['sex', 'name'],
                                values='births', aggfunc='sum')
    girls = name_pivot['F'].dropna(axis=1, how='any')
    boys = name_pivot['M'].dropna(axis=1, how='any')
    girls_sort = girls.sum(axis=0).sort_values(ascending=False)
    girls_sort_txt = ' '.join(list(girls_sort.index))

    boys_sort = boys.sum(axis=0).sort_values(ascending=False)
    boys_sort_txt = ' '.join(list(boys_sort.index))

    return girls_sort_txt, boys_sort_txt
```

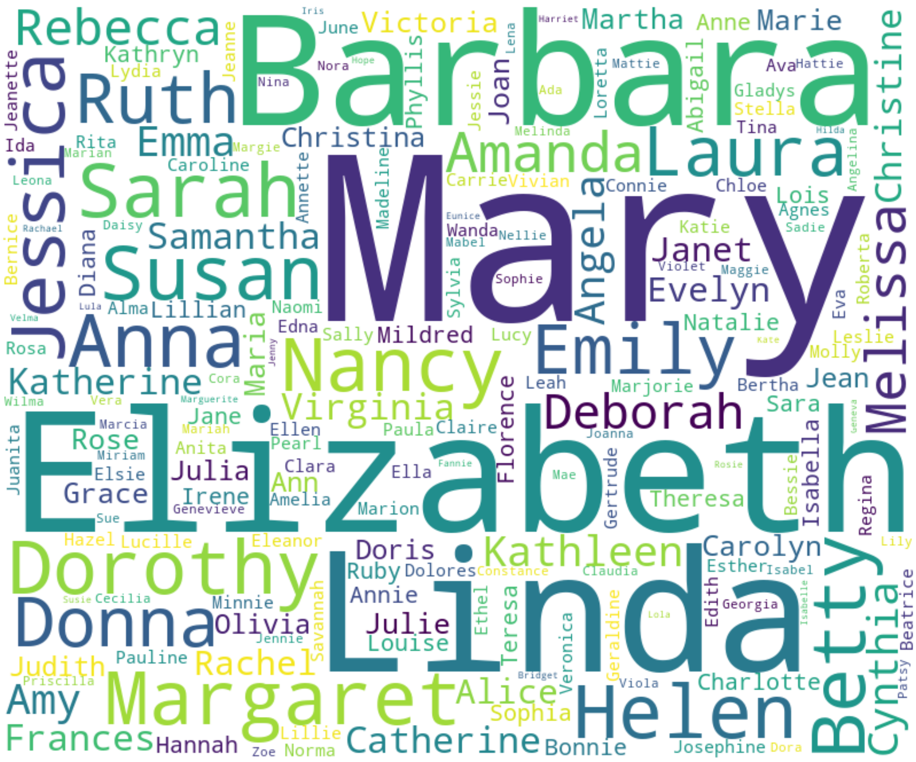
Out：

```
def wordcloudfunc(data, pname):
    wordcloud = WordCloud(background_color='white', width=800, height=660,
                          margin=2, collocations=False, relative_scaling=0,
                          normalize_plurals=False).generate(
        data)
    plt.imshow(wordcloud)
    plt.axis('off')
    wordcloud.to_file(pname)
```

男孩：



女孩：



结论：

① 自 19 世纪八十年代以来，每年都在用且前五个最受欢迎的男孩姓名：James, John, Robert, Michael, William。

② 自 19 世纪八十年代以来，每年都在用且前五个最受欢迎的女孩姓名：Mary, Elizabeth,

Linda, Barbara, Margaret。

➢ 总结

1、自 19 世纪 80 年代以来，美国新出生婴儿的数量呈增长趋势，至 20 世纪后期变化趋势趋于平稳。

2、从 20 世纪中期开始，美国父母给孩子取名字的趋势更加趋于多样化，且女孩的多样化程度更高。

3、不管给男孩还是女孩取名字，父母都不喜欢用以 X、Y 和 Z 开头的名字。

4、自 19 世纪 80 年代以来，父母偏向于给男孩取以 J 开头的名字。