

Contents

- 1 分析过程与目标
  - 1.1 数据来源
  - 1.2 熟悉数据
- 2 各国与地区GDP数据分析
  - 2.1 导入数据
  - 2.2 清洗数据
  - 2.3 设定分析目标
- 3 主要国家DGP分析
  - 3.1 主要国家GDP趋势
  - 3.2 1990年开始GDP对比
- 4 中国GDP分析
  - 4.1 从1990年开始GDP变化
  - 4.2 中国GDP分析增长超过10%的年份
    - 4.2.1 计算每年增长率思路
    - 4.2.2 获取增长大于10%的年份
  - 4.3 5年连续累加增长率最高年份
    - 4.3.1 rolling:移动窗口方法
    - 4.3.2 代码实现

1 分析过程与目标

- 数据来源
- 熟悉数据
- 分析过程
- 分析结果呈现
- 使用知识点与代码实现

1.1 数据来源

- 企业内部采集数据：web端,小程序,Android或者IOS应用，智能设备(智能电表，温度传感器等)
- 开方数据平台：国家数据统计局，世界银行数据等
- 第三方数据集：kaagle等竞赛平台
- 爬虫抓取第三方数据
- 数据可能由多源组成

1.2 熟悉数据

- 通过工具展示数据
- 查看数据字段
- 多个数据源观察，数据源关系

2 各国与地区GDP数据分析

2.1 导入数据

```
In [9]: 1 import pandas as pd
2 import numpy as np
3 %matplotlib inline
4 #读取excel文件
5 fpath = r'F:\database\pandas_dir\GDP.csv'
6 f = open(fpath)
7 pdata = pd.read_csv(f)
```

```
In [10]: 1 pdata[:3]
```

Out[10]:

	Country Name	Country Code	Indicator Name	Indicator Code	1960	1961	1962	1963	1964
0	Aruba	ABW	GDP (current US\$)	NY.GDP.MKTP.CD	NaN	NaN	NaN	NaN	NaN
1	Afghanistan	AFG	GDP (current US\$)	NY.GDP.MKTP.CD	537777811.1	548888895.6	546666677.8	751111191.1	800000044.4
2	Angola	AGO	GDP (current US\$)	NY.GDP.MKTP.CD	NaN	NaN	NaN	NaN	NaN

3 rows × 62 columns

2.2 清洗数据

观察数据，删除无用数据；

```
In [3]: 1 pdata.columns
```

```
Out[3]: Index(['Country Name', 'Country Code', 'Indicator Name', 'Indicator Code',
              '1960', '1961', '1962', '1963', '1964', '1965', '1966', '1967', '1968',
              '1969', '1970', '1971', '1972', '1973', '1974', '1975', '1976', '1977',
              '1978', '1979', '1980', '1981', '1982', '1983', '1984', '1985', '1986',
              '1987', '1988', '1989', '1990', '1991', '1992', '1993', '1994', '1995',
              '1996', '1997', '1998', '1999', '2000', '2001', '2002', '2003', '2004',
              '2005', '2006', '2007', '2008', '2009', '2010', '2011', '2012', '2013',
              '2014', '2015', '2016', '2017'],
              dtype='object')
```

```
In [3]: 1 #删除数据
2 pdata = pdata.drop(['Country Code', 'Indicator Name', 'Indicator Code'], axis=1)
3 #重置索引
4 pdata = pdata.set_index('Country Name')
5 pdata = pdata.stack()
6 pdata = pd.DataFrame(pdata)
7 pdata.columns = ['GDP']
```

```
In [ ]: 1 pdata
```

2.3 设定分析目标

- 主要国家GDP数据变化
- 从1990年开始主要国家GDP数据变化
- 中国GDP1990年开始GDP增长与累积增长
- 中国GDP1990年开始，增长超过10%年份
- 中国GDP连续5年增长最高的年份

3 主要国家DGP分析

选择国家：['China', 'Japan', 'United States', 'Germany', 'France', 'United Kingdom']

3.1 主要国家GDP趋势

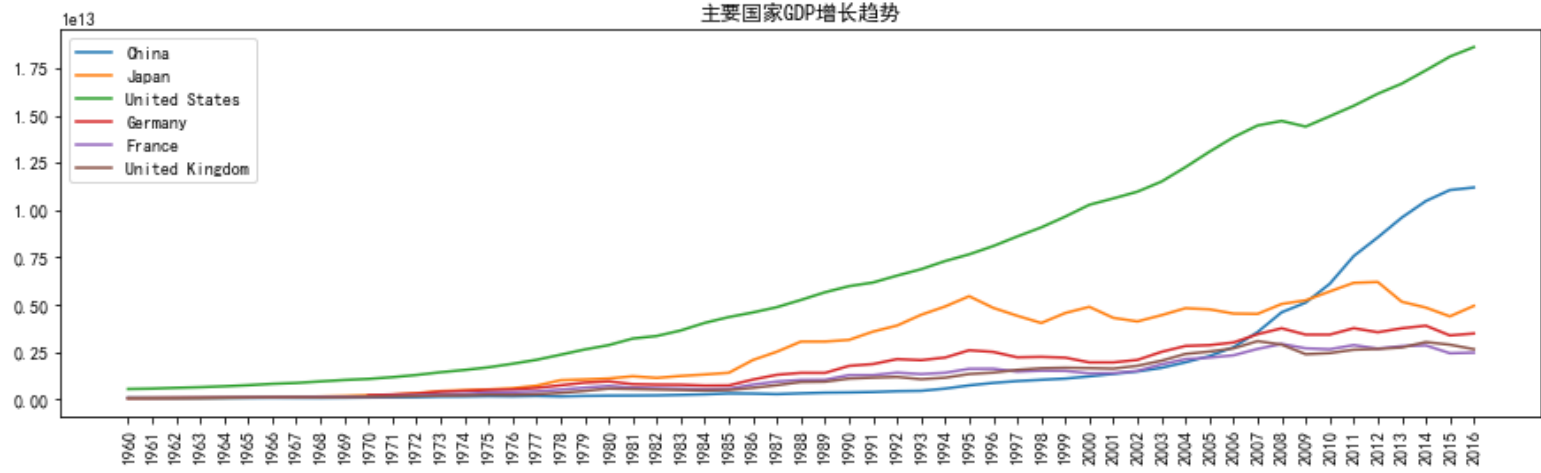
问题：选择合适图表代表数据趋势？折线图

Contents

- 1 分析过程与目标
  - 1.1 数据来源
  - 1.2 熟悉数据
- 2 各国与地区GDP数据分析
  - 2.1 导入数据
  - 2.2 清洗数据
  - 2.3 设定分析目标
- 3 主要国家DGP分析
  - 3.1 主要国家GDP趋势
  - 3.2 1990年开始GDP对比
- 4 中国GDP分析
  - 4.1 从1990年开始GDP变化
  - 4.2 中国GDP分析增长超过10%的年份
    - 4.2.1 计算每年增长率思路
    - 4.2.2 获取增长大于10%的年份
  - 4.3 5年连续累加增长率最高年份
    - 4.3.1 rolling:移动窗口方法
    - 4.3.2 代码实现

In [6]:

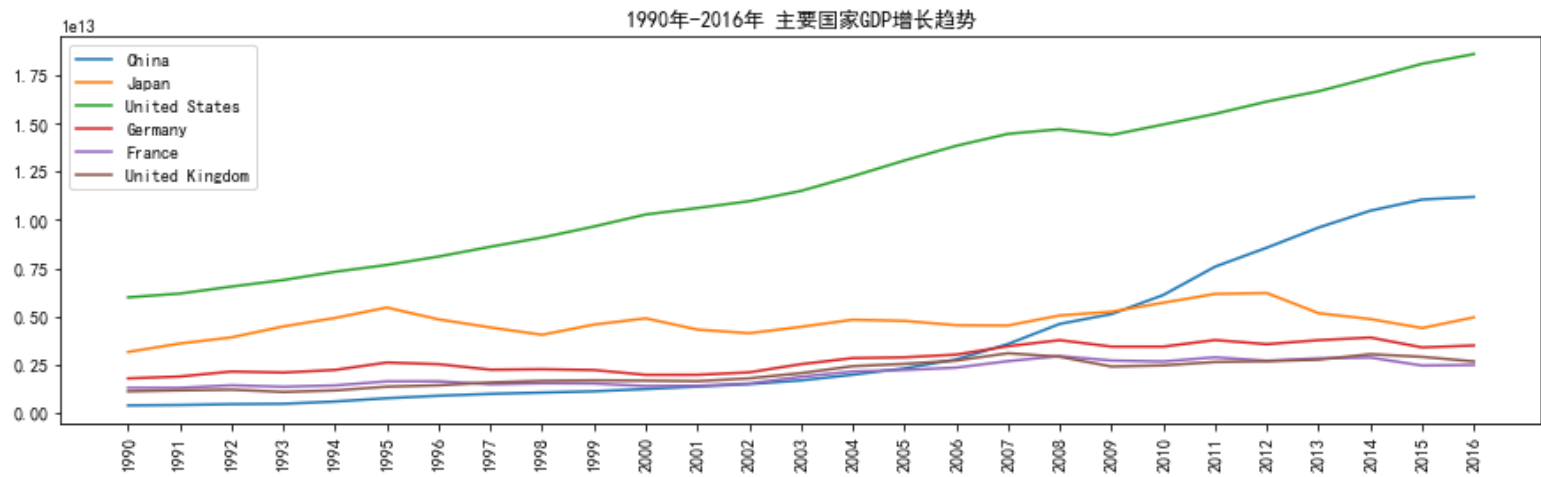
```
1 import matplotlib.pyplot as plt
2 import matplotlib
3 import matplotlib.pyplot as plt
4 #设置支持中文
5 matplotlib.rcParams['font.sans-serif'] = ['SimHei']
6 matplotlib.rcParams['axes.unicode_minus']=False
7 plt.figure(l,figsize=(15, 4))
8 countrys = ['China', 'Japan','United States', 'Germany', 'France', 'United Kingdom']
9 for c in countrys:
10     plt.plot(pdata.loc[c])
11 plt.title("主要国家GDP增长趋势")
12 plt.legend(countrys)
13 _ = plt.xticks(rotation=90)
```



3.2 1990年开始GDP对比

In [7]:

```
1 import matplotlib.pyplot as plt
2 plt.figure(l,figsize=(15, 4))
3 countrys = ['China', 'Japan','United States', 'Germany', 'France', 'United Kingdom']
4 for c in countrys:
5     #取国家, 切片, 取年代
6     plt.plot(pdata.loc[c]['1990':])
7 plt.title("1990年-2016年 主要国家GDP增长趋势")
8 plt.legend(countrys)
9 _ = plt.xticks(rotation=90)
```

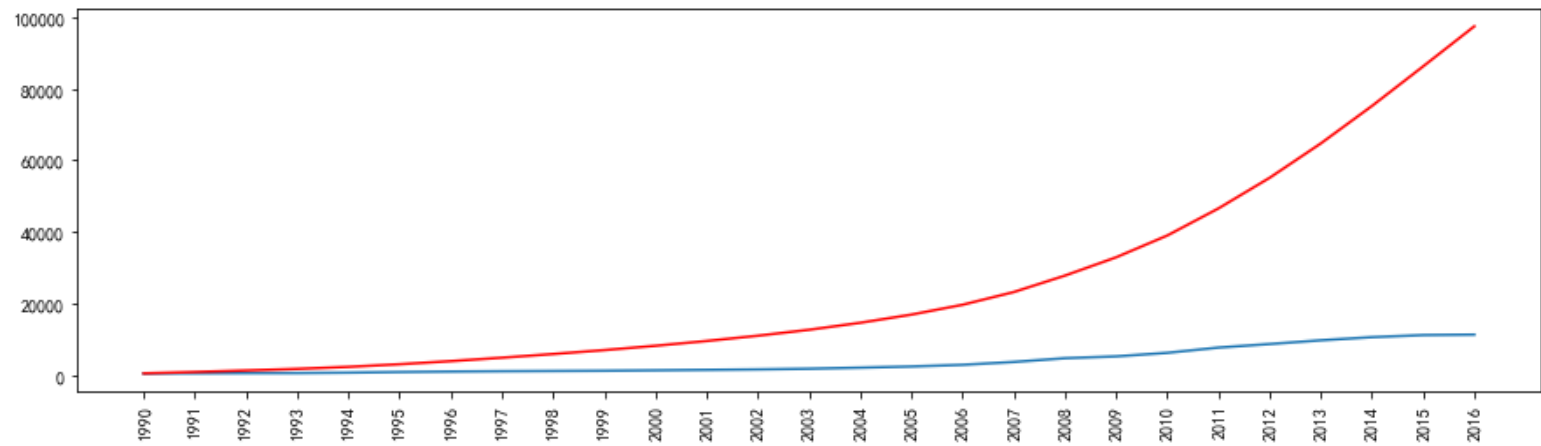


4 中国GDP分析

4.1 从1990年开始GDP变化

In [8]:

```
1 import matplotlib.pyplot as plt
2 plt.rcParams['font.family'] = ['sans-serif']
3 plt.rcParams['font.sans-serif'] = ['SimHei']
4 _ = plt.figure(l,figsize=(15, 4))
5 #国家
6 countrys = ['China']
7 base1 = 10000*10000*10
8 base2 = 10000*10000*10
9 for c in countrys:
10     #中国, 年份1990-
11     data = pdata.loc[c]['1990':]
12     plt.plot(data/base1, label='每年值')
13     #累计值
14     plt.plot(data.cumsum()/base1,label='累积GDP', color='r')
15 _ = plt.xticks(rotation=90)
```



4.2 中国GDP分析增长超过10%的年份

问题： 计算每一年增长率，使用什么知识点？

In [20]:

```
1 gdp = pdata.loc['China']['1990':]
```

4.2.1 计算每年增长率思路

思路1: 循环迭代： 第二年-第一年/第一年

思路2: 利用numpy计算

- 第一年数据tmp1: [开始: 结束-1]
- 第二年数据tmp2: [第二年:结束]
- 结果: tmp2-tmp1/tmp1 \* 100

Contents

- 1 分析过程与目标
  - 1.1 数据来源
  - 1.2 熟悉数据
- 2 各国与地区GDP数据分析
  - 2.1 导入数据
  - 2.2 清洗数据
  - 2.3 设定分析目标
- 3 主要国家DGP分析
  - 3.1 主要国家GDP趋势
  - 3.2 1990年开始GDP对比
- 4 中国GDP分析
  - 4.1 从1990年开始GDP变化
  - 4.2 中国GDP分析增长超过10%的年份
    - 4.2.1 计算每年增长率思路
    - 4.2.2 获取增长大于10%的年份
  - 4.3 5年连续累加增长率最高年份
    - 4.3.1 rolling:移动窗口方法
    - 4.3.2 代码实现

```
In [21]: 1 #获取第一年与第二年数据
2 tmp1 = gdp.loc[:, '2015'] / 1000000
3 tmp2 = gdp.loc[:, '1991':] / 1000000
4 #转换成整数
5 tmp1 = tmp1.astype('i')
6 tmp2 = tmp2.astype('i')
7 #计算增长率
8 index = (tmp2.values - tmp1.values) / tmp1.values * 100
9 #第一年插入0
10 grow = np.insert(index, 0, 0)
11 #插入新的列
12 gdp['grow'] = grow
```

```
In [23]: 1 gdp
```

Out[23]:

	GDP	grow
1990	3.608580e+11	0.000000
1991	3.833730e+11	6.239296
1992	4.269160e+11	11.357868
1993	4.447310e+11	4.172952
1994	5.643250e+11	26.891312
1995	7.345480e+11	30.164001
1996	8.637470e+11	17.588912
1997	9.616040e+11	11.329359
1998	1.029040e+12	7.012866
1999	1.094000e+12	6.312680
2000	1.211350e+12	10.726691
2001	1.339400e+12	10.570851
2002	1.470550e+12	9.791698
2003	1.660290e+12	12.902655
2004	1.955350e+12	17.771594
2005	2.285970e+12	16.908482
2006	2.752130e+12	20.392219
2007	3.552180e+12	29.070211
2008	4.598210e+12	29.447551
2009	5.109950e+12	11.129113
2010	6.100620e+12	19.387078
2011	7.572550e+12	24.127548
2012	8.560550e+12	13.047124
2013	9.607220e+12	12.226668
2014	1.048240e+13	9.109607
2015	1.106470e+13	5.555026
2016	1.119910e+13	1.214674

4.2.2 获取增长大于10%的年份

```
In [25]: 1 vals = gdp[gdp.grow > 10]
2 vals
```

Out[25]:

	GDP	grow
1992	4.269160e+11	11.357868
1994	5.643250e+11	26.891312
1995	7.345480e+11	30.164001
1996	8.637470e+11	17.588912
1997	9.616040e+11	11.329359
2000	1.211350e+12	10.726691
2001	1.339400e+12	10.570851
2003	1.660290e+12	12.902655
2004	1.955350e+12	17.771594
2005	2.285970e+12	16.908482
2006	2.752130e+12	20.392219
2007	3.552180e+12	29.070211
2008	4.598210e+12	29.447551
2009	5.109950e+12	11.129113
2010	6.100620e+12	19.387078
2011	7.572550e+12	24.127548
2012	8.560550e+12	13.047124
2013	9.607220e+12	12.226668

4.3 5年连续累加增长率最高年份

分析问题：5年连续增长：第一年+第二年+第三年...+第五年

如何计算并获取最大？

4.3.1 rolling:移动窗口方法

应用场景：金融，股票，统计等

一个需求：连续2个数据中最大值 例如：

1,2,3,1,1,0

结果：

1,2,3,3,1,1

主要参数：

参数	说明
window	窗口,数值或者时间
min_periods	最小周期，默认为window

Contents 🔁 ⚙️

- ▼ 1 分析过程与目标
  - 1.1 数据来源
  - 1.2 熟悉数据
- ▼ 2 各国与地区GDP数据分析
  - 2.1 导入数据
  - 2.2 清洗数据
  - 2.3 设定分析目标
- ▼ 3 主要国家DGP分析
  - 3.1 主要国家GDP趋势
  - 3.2 1990年开始GDP对比
- ▼ 4 中国GDP分析
  - 4.1 从1990年开始GDP变化
  - ▼ 4.2 中国GDP分析增长超过10%的年份
    - 4.2.1 计算每年增长率思路
    - 4.2.2 获取增长大于10%的年份
  - ▼ 4.3 5年连续累加增长率最高年份
    - 4.3.1 rolling:移动窗口方法
    - 4.3.2 代码实现

参数	说明
center	取值方式，True:前后取值，False:向前取值
on	指定列，一般为datetime-like的column
axis	指定轴

```
In [ ]:
```

4.3.2 代码实现

```
In [36]:
```

```
In [37]:
```

Out[37]:

	GDP	grow
1990	NaN	NaN
1991	NaN	NaN
1992	NaN	NaN
1993	NaN	NaN
1994	2.180203e+12	48.661428
1995	2.553893e+12	78.825430
1996	3.034267e+12	90.175045
1997	3.568955e+12	90.146536
1998	4.153264e+12	92.986450
1999	4.682939e+12	72.407818
2000	5.159741e+12	52.970508

```
In [38]:
```

Out[38]:

	GDP	grow
2008	1.514384e+13	113.590056

```
In [ ]:
```