Московский государственный технический университет им. Н.Э. Баумана.

Факультет «Информатика и управление»

Кафедра ИУ5. Курс «Разработка интернет-приложений»

Отчет по лабораторной работе №1:

«Разведочный анализ данных. Исследование и визуализация данных»

Выполнил:
студент группы ИУ5-63
Курганова Александра

Подпись и дата:

Проверил:

Подпись и дата:

Москва, 2019 г.

Цель ЛР1

Изучить различные методы визуализации данных.

Текстовое описание набора данных

В качестве набора данных использован набор данных по информации о населении, регионе, размере территории, младенческой смертности и т.д. - https://www.kaggle.com/fernandol/countries-of-the-world/version/1

Все эти наборы данных состоят из данных правительства США за 1970-2017. При создании визуализаций, связанных со странами, иногда интересно сгруппировать их по таким атрибутам, как регион, или взвесить их важность по численности населения, ВВП или другим переменным.

Датасет состоит из 20 колонок:

- Country – страна
- Region – регион
- Population – количество жителей
- Area – площадь (кв.миля)
- Pop. Density – плотность населения (на 1 кв.милю)
- Coastline – соотношение берег/площадь
- Net migration – сетевая миграция
- Infant mortality – младенческая смертность (на 1000 рождений)
- GPD – ВВП (doll. на душу населения)
- Literacy – грамотность (%)
- Phones – телефоны (на 1000)
- Arable – пашня (%)
- Crops – культура (%)
- Other – другое (%)
- Climate – климатический пояс
- Birthrate – уровень рождаемости
- Deathrate – уровень смертности
- Agriculture – сельсткое хозяйство
- Industry – промышленность
- Service – оказание услуг

Импорт библиотек

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

Загрузка данных

Загрузка файла датасета происходит с помощью библиотеки Pandas. Данные представлены в формате CSV.

```
data = pd.read_csv('countries of the world.csv', sep=",", decimal=',')
```

Основные характеристики датасета

```
# Первые 5 строк
data.head()
```

	Country	Region	Population	Area (sq. mi.)	Pop. Density (per sq. mi.)	Coastline (coast/area ratio)	Net migration	Infant mortality (per 1000 births)	GDP ($ per capita)	Literacy (%)	Phones (per 1000)	Arable (%)	Crops (%)	Other (%)	Climate	Birthrate
0	Afghanistan	ASIA (EX. NEAR EAST)	31056997	647500	48.0	0.00	23.06	163.07	700.0	36.0	3.2	12.13	0.22	87.65	1.0	46.60
1	Albania	EASTERN EUROPE	3581655	28748	124.6	1.26	-4.93	21.52	4500.0	86.5	71.2	21.09	4.42	74.49	3.0	15.11
2	Algeria	NORTHERN AFRICA	32930091	2381740	13.8	0.04	-0.39	31.00	6000.0	70.0	78.1	3.22	0.25	96.53	1.0	17.14
3	American Samoa	OCEANIA	57794	199	290.4	58.29	-20.71	9.27	8000.0	97.0	259.5	10.00	15.00	75.00	2.0	22.46
4	Andorra	WESTERN EUROPE	71201	468	152.1	0.00	6.60	4.05	19000.0	100.0	497.2	2.22	0.00	97.78	3.0	8.71

Birthrate	Deathrate	Agriculture	Industry	Service
46.60	20.34	0.380	0.240	0.380
15.11	5.22	0.232	0.188	0.579
17.14	4.61	0.101	0.600	0.298
22.46	3.27	NaN	NaN	NaN
8.71	6.25	NaN	NaN	NaN

```
# Размер датасета
data.shape
 (227, 20)
rows = data.shape[0]
print('Всего строк: {}'.format(rows))
Всего строк: 227

# Список колонок
data.columns
Index(['Country', 'Region', 'Population', 'Area (sq. mi.)',
       'Pop. Density (per sq. mi.)', 'Coastline (coast/area ratio)',
       'Net migration', 'Infant mortality (per 1000 births)',
       'GDP ($ per capita)', 'Literacy (%)', 'Phones (per 1000)', 'Arable (%)',
       'Crops (%)', 'Other (%)', 'Climate', 'Birthrate', 'Deathrate',
       'Agriculture', 'Industry', 'Service'],
      dtype='object')

# Список колонок с типами данных
data.dtypes

Country                               object
Region                                object
Population                             int64
Area (sq. mi.)                         int64
Pop. Density (per sq. mi.)           float64
Coastline (coast/area ratio)         float64
Net migration                        float64
Infant mortality (per 1000 births)   float64
GDP ($ per capita)                   float64
Literacy (%)                         float64
```

```
Phones (per 1000)                   float64
Arable (%)                          float64
Crops (%)                           float64
Other (%)                           float64
Climate                             float64
Birthrate                           float64
Deathrate                           float64
Agriculture                         float64
Industry                            float64
Service                             float64
dtype: object
```

```
# Наличие пустых значений
def empty():
    for col in data.columns:
        null_col = data[data[col].isnull()].shape[0]
        print('{} - {}'.format(col, null_col))

print('Before:')
empty()

# Удаление пустых строк
data = data.dropna(axis=0, how='any', thresh=None, subset=None, inplace=False)
print('\nAfter:')
empty()
```

```
Before:
Country - 0
Region - 0
Population - 0
Area (sq. mi.) - 0
Pop. Density (per sq. mi.) - 0
Coastline (coast/area ratio) - 0
Net migration - 3
Infant mortality (per 1000 births) - 3
GDP ($ per capita) - 1
Literacy (%) - 18
Phones (per 1000) - 4
Arable (%) - 2
Crops (%) - 2
Other (%) - 2
Climate - 22
Birthrate - 3
Deathrate - 4
Agriculture - 15
Industry - 16
Service - 15

After:
Country - 0
Region - 0
Population - 0
Area (sq. mi.) - 0
```

```
Pop. Density (per sq. mi.) - 0
Coastline (coast/area ratio) - 0
Net migration - 0
Infant mortality (per 1000 births) - 0
GDP ($ per capita) - 0
Literacy (%) - 0
Phones (per 1000) - 0
Arable (%) - 0
Crops (%) - 0
Other (%) - 0
Climate - 0
Birthrate - 0
Deathrate - 0
Agriculture - 0
Industry - 0
Service - 0
```

```
# Основные статистические характеристки набора данных
data.describe()
```

	Population	Area (sq. mi.)	Pop. Density (per sq. mi.)	Coastline (coast/area ratio)	Net migration	Infant mortality (per 1000 births)	GDP ($ per capita)	Literacy (%)	Phones (per 1000)	Arable (%)	Crops (%)	Other (
count	1.790000e+02	1.790000e+02	179.000000	179.000000	179.000000	179.000000	179.000000	179.000000	179.000000	179.000000	179.000000	179.0000
mean	3.421415e+07	5.641830e+05	294.805028	16.495196	-0.206536	38.898156	9125.698324	81.944134	208.151955	14.000447	4.440838	81.5582
std	1.317639e+08	1.395657e+06	1379.352557	73.977601	4.751216	35.353003	9644.123141	19.837537	210.121875	13.152405	8.090331	16.1037
min	1.347700e+04	2.800000e+01	1.800000	0.000000	-20.990000	2.290000	500.000000	17.600000	0.200000	0.000000	0.000000	33.3300
25%	1.188580e+06	1.991500e+04	26.800000	0.090000	-1.315000	9.990000	1800.000000	69.950000	27.100000	3.675000	0.230000	72.8250
50%	6.940432e+06	1.184800e+05	66.900000	0.630000	0.000000	24.310000	5100.000000	90.900000	137.100000	10.530000	1.030000	86.0700
75%	2.086014e+07	4.964410e+05	164.700000	5.355000	0.395000	64.605000	12950.000000	97.800000	335.000000	20.000000	4.600000	94.8100
max	1.313974e+09	9.631420e+06	16183.000000	870.660000	23.060000	163.070000	37800.000000	100.000000	898.000000	62.110000	48.960000	100.0000

Other (%)	Climate	Birthrate	Deathrate	Agriculture	Industry	Service
179.000000	179.000000	179.000000	179.000000	179.000000	179.000000	179.000000
81.558212	2.108939	23.067486	9.465140	0.156905	0.288028	0.554508
16.103748	0.697611	11.287207	5.210083	0.151343	0.140310	0.165670
33.330000	1.000000	7.290000	2.410000	0.000000	0.032000	0.062000
72.825000	2.000000	13.890000	5.795000	0.039000	0.197000	0.424500
86.070000	2.000000	20.460000	7.840000	0.101000	0.274000	0.559000
94.810000	2.500000	32.315000	11.660000	0.233000	0.349000	0.668500
100.000000	4.000000	50.730000	29.740000	0.769000	0.906000	0.954000

```
# Уникальные значения для сетевой миграции
data['Net migration'].unique()
```

```
array([ 2.306e+01, -4.930e+00, -3.900e-01,  1.076e+01, -6.150e+00,
        6.100e-01, -6.470e+00,  0.000e+00,  3.980e+00,  2.000e+00,
       -4.900e+00, -2.200e+00,  1.050e+00, -7.100e-01, -3.100e-01,
        2.540e+00,  1.230e+00,  2.490e+00, -1.320e+00, -3.000e-02,
        1.001e+01,  3.590e+00, -4.580e+00, -1.800e+00, -6.000e-02,
       -1.207e+01,  1.875e+01, -1.100e-01, -4.000e-01, -1.700e-01,
        5.100e-01, -7.000e-02, -1.580e+00,  9.700e-01,  2.480e+00,
       -1.387e+01, -3.220e+00, -8.580e+00, -2.200e-01, -3.740e+00,
       -3.160e+00, -3.140e+00,  9.500e-01,  6.600e-01,  6.270e+00,
        2.940e+00,  1.570e+00, -4.700e+00,  2.180e+00, -6.400e-01,
        2.350e+00, -1.392e+01, -1.500e-01, -1.670e+00, -3.060e+00,
       -1.570e+00, -2.070e+00, -3.400e+00, -1.990e+00,  5.240e+00,
```

```
 8.600e-01,  2.380e+00, -8.400e-01,  4.990e+00,  6.800e-01,
-4.920e+00,  6.590e+00, -3.350e+00, -1.000e-01,  1.418e+01,
-2.450e+00, -2.230e+00, -7.400e-01,  4.850e+00,  4.860e+00,
-3.300e-01, -6.040e+00, -5.000e-02, -9.000e-01, -4.870e+00,
-2.099e+01,  2.910e+00, -4.100e-01,  4.050e+00, -1.220e+00,
-6.700e-01,  2.600e-01,  1.740e+00,  2.800e-01, -2.770e+00,
 2.850e+00, -9.100e-01, -8.000e-02, -1.050e+00, -1.500e+00,
-4.900e-01,  3.570e+00, -1.460e+00,  1.629e+01, -1.300e-01,
-7.110e+00, -2.670e+00, -7.640e+00, -1.170e+01, -2.720e+00,
-2.710e+00,  2.000e-01, -5.690e+00,  1.153e+01,  5.370e+00,
-2.900e-01,  9.900e-01, -1.310e+00, -2.000e-02, -8.810e+00,
 1.670e+00, -2.860e+00, -1.083e+01, -5.700e-01, -8.600e-01,
 1.030e+00,  2.190e+00,  3.410e+00, -3.200e-01, -1.720e+00,
-4.000e-02, -4.500e-01])
```

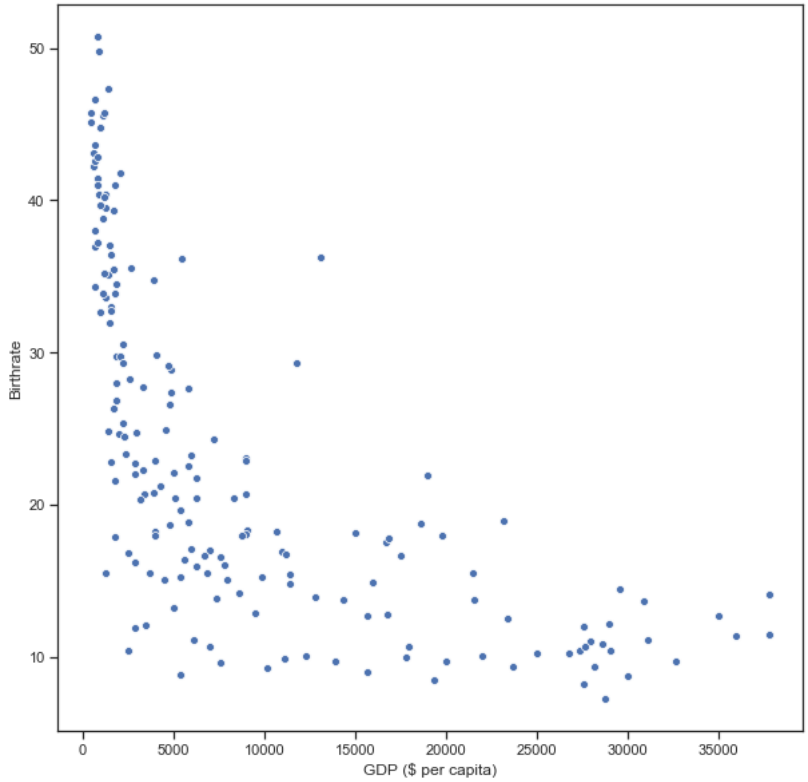
Визуальное исследование датасета

Диаграмма рассеяния

Позволяет построить распределение двух колонок данных (уровень рождаемости от ВВП) и визуально обнаружить наличие зависимости (почти линейная).

```
fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='GDP ($ per capita)', y='Birthrate', data=data)
```

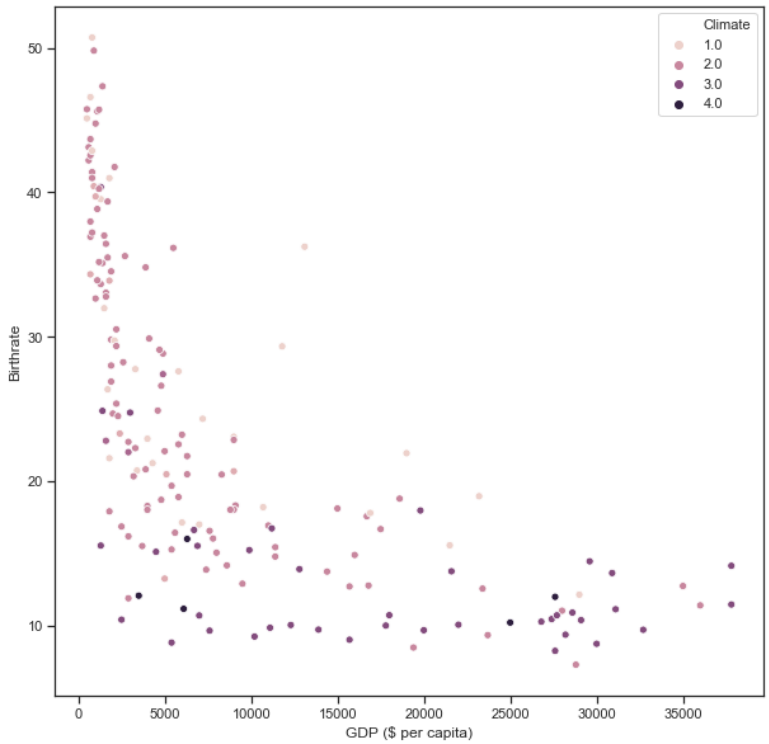
```
<matplotlib.axes._subplots.AxesSubplot at 0x152340630>
```



Посмотрим насколько на эту зависимость влияет климат

```
fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='GDP ($ per capita)', y='Birthrate', data=data, hue='Climate')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x153c836a0>
```

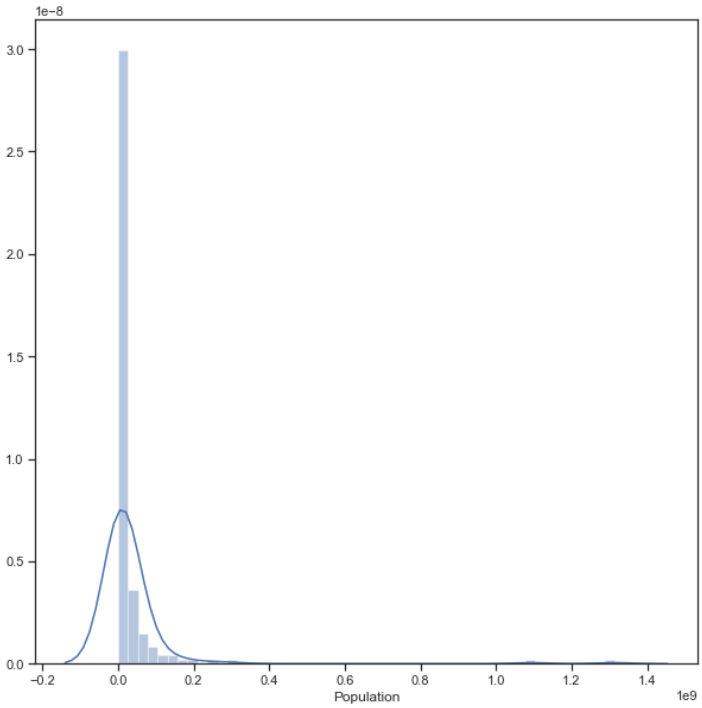


Гистограмма

Позволяет оценить плотность вероятности распределения данных (младенческая смертность (на 1000 рождений)).

```
fig, ax = plt.subplots(figsize=(10,10))
sns.distplot(data['Population'])
```

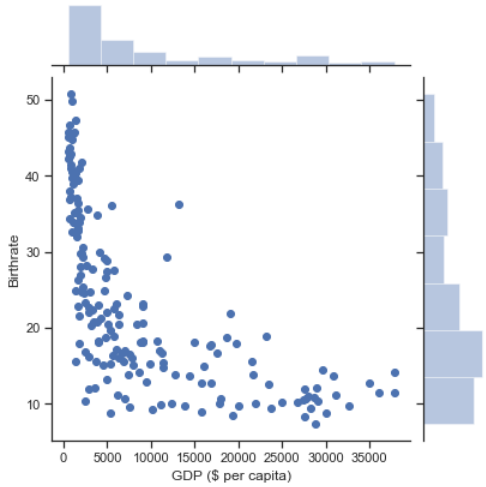
```
<matplotlib.axes._subplots.AxesSubplot at 0x14d0194e0>
```



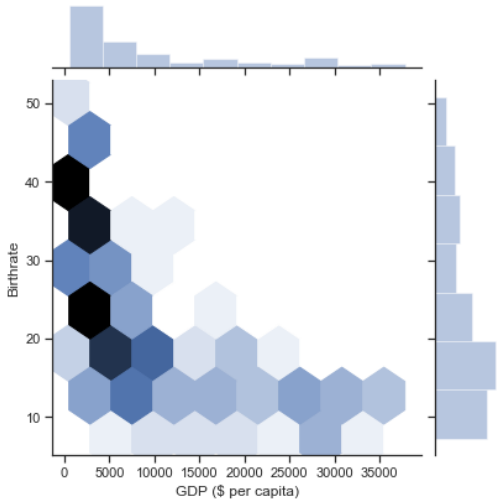
Jointplot

Комбинация гистограмм и диаграмм рассеивания.

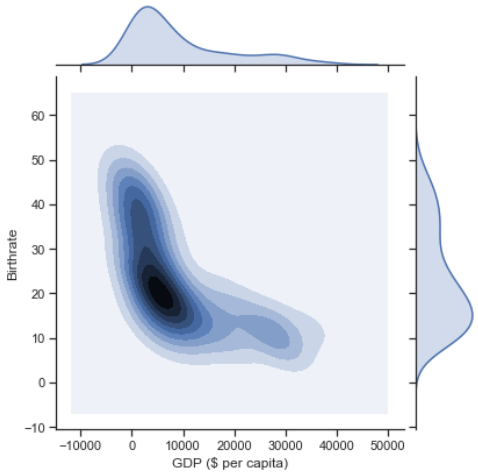
```
sns.jointplot(x='GDP ($ per capita)', y='Birthrate', data=data)
<seaborn.axisgrid.JointGrid at 0x147b2a080>
```




```
sns.jointplot(x='GDP ($ per capita)', y='Birthrate', data=data, kind="hex")
<seaborn.axisgrid.JointGrid at 0x147eb2ba8>
```




```
sns.jointplot(x='GDP ($ per capita)', y='Birthrate', data=data, kind="kde")
<seaborn.axisgrid.JointGrid at 0x1480a9ba8>
```



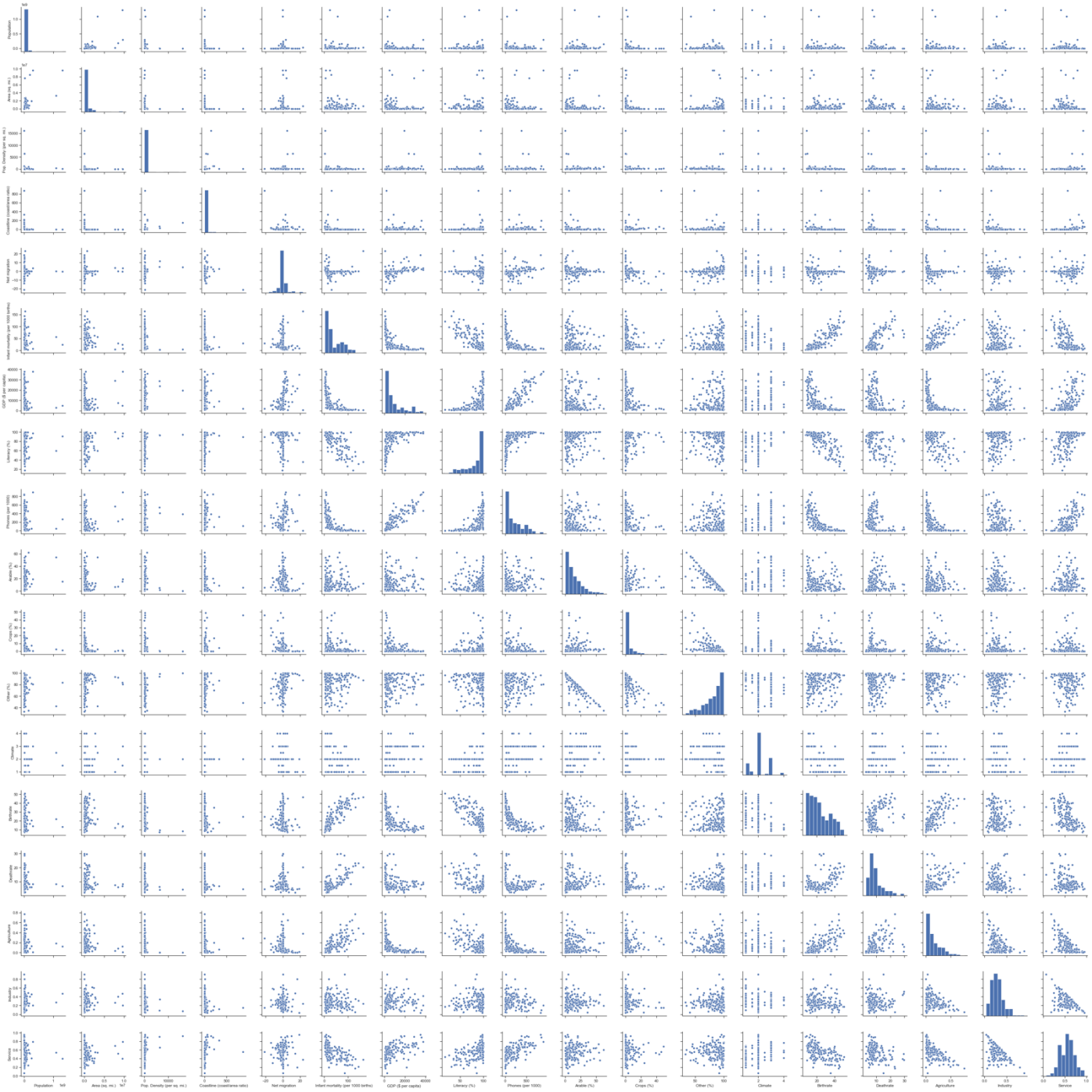
Парные диаграммы

Комбинация гистограмм и диаграмм рассеивания для всего набора данных.

Выводится матрица графиков. На пересечении строки и столбца, которые соответствуют двум показателям, строится диаграмма рассеивания. В главной диагонали матрицы строятся гистограммы распределения соответствующих показателей.

```
sns.pairplot(data)
```

```
<seaborn.axisgrid.PairGrid at 0x14841bba8>
```

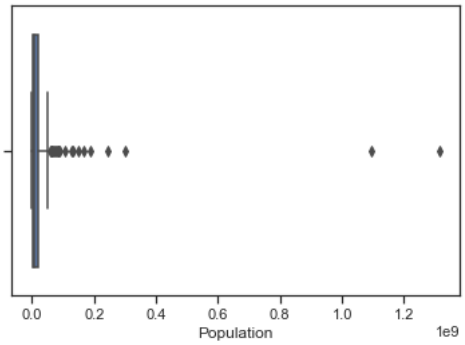


Ящик с усами

Отображает одномерное распределение вероятности.

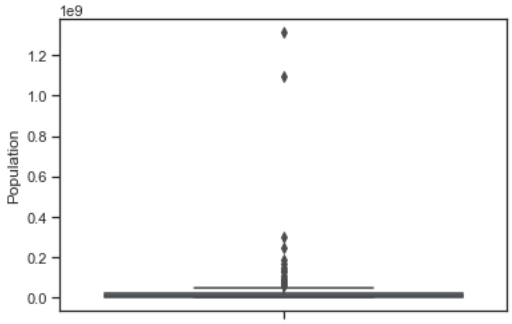
```
sns.boxplot(x=data['Population'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x132366518>
```



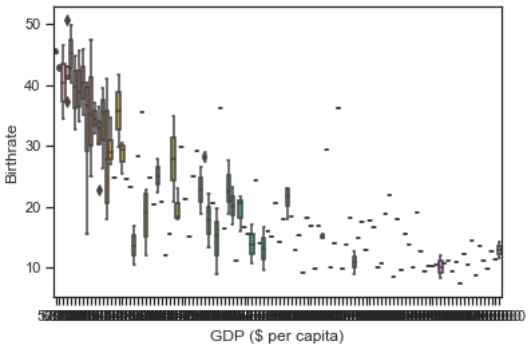

```
# По вертикали
sns.boxplot(y=data['Population'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x136eb5470>
```




```
# Распределение параметра Birthrate сгруппированные по GDP ($ per capita).
sns.boxplot(x='GDP ($ per capita)', y='Birthrate', data=data)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x158a07278>
```

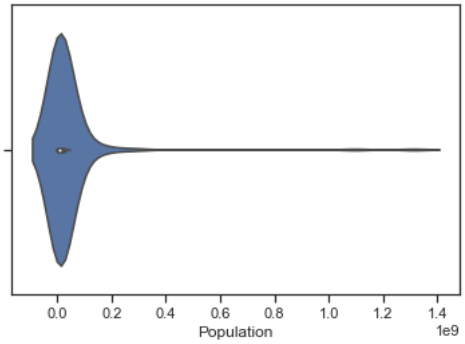


Violin plot

По краям отображаются распределения плотности

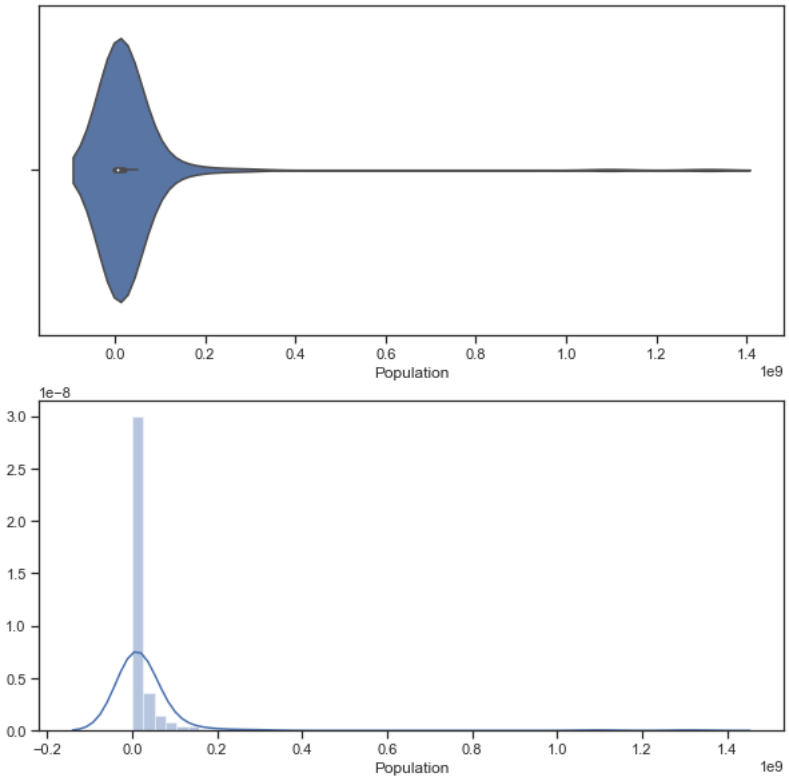
```
sns.violinplot(x=data['Population'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1354ae240>
```




```
fig, ax = plt.subplots(2, 1, figsize=(10,10))
sns.violinplot(ax=ax[0], x=data['Population'])
sns.distplot(data['Population'], ax=ax[1])
```

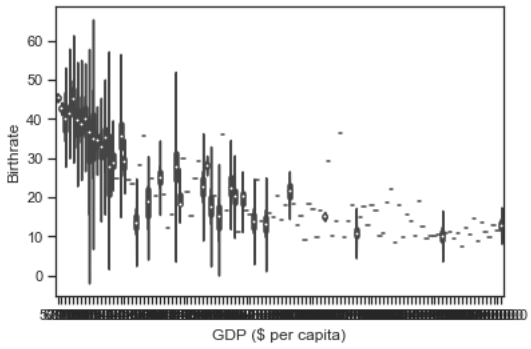
```
<matplotlib.axes._subplots.AxesSubplot at 0x131bc24a8>
```



Из приведенных графиков видно, что violinplot действительно показывает распределение плотности.

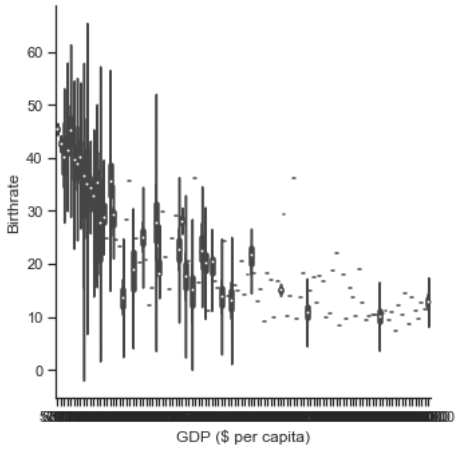
```
# Распределение параметра GDP ($ per capita) сгруппированные по Birthrate.
sns.violinplot(x='GDP ($ per capita)', y='Birthrate', data=data)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x131490e48>
```




```
sns.catplot(y='Birthrate', x='GDP ($ per capita)',data=data, kind="violin", split=True)
```

```
<seaborn.axisgrid.FacetGrid at 0x162554470>
```



Информация о корреляции признаков

Проверка корреляции признаков позволяет решить две задачи:

1) Понять какие признаки (колонки датасета) наиболее сильно коррелируют с целевым признаком ("Birthrate"). Именно эти признаки будут наиболее информативными для моделей машинного обучения. Признаки, которые слабо коррелируют с целевым признаком, можно попробовать исключить из построения модели, иногда это повышает качество модели.

2) Понять какие нецелевые признаки линейно зависимы между собой.

```
data.corr()
```

	Population	Area (sq. mi.)	Pop. Density (per sq. mi.)	Coastline (coast/area ratio)	Net migration	Infant mortality (per 1000 births)	GDP ($ per capita)	Literacy (%)	Phones (per 1000)	Arable (%)	Crops (%)	Other (%)	Climate	Bir
Population	1.000000	0.610850	-0.019010	-0.054617	0.010332	0.002438	-0.033618	-0.038162	-0.003942	0.206667	-0.062567	-0.137345	-0.018471	-0.0
Area (sq. mi.)	0.610850	1.000000	-0.069010	-0.088162	0.052730	0.002924	0.068356	0.000682	0.077864	-0.053747	-0.160433	0.124528	-0.094852	-0.0
Pop. Density (per sq. mi.)	-0.019010	-0.069010	1.000000	0.164036	0.161705	-0.143214	0.190122	0.086090	0.157072	-0.059223	-0.036580	0.066753	-0.012370	-0.1
Coastline (coast/area ratio)	-0.054617	-0.088162	0.164036	1.000000	-0.241629	-0.105956	0.035815	0.099605	0.098367	-0.077800	0.399358	-0.137085	-0.027063	-0.0
Net migration	0.010332	0.052730	0.161705	-0.241629	1.000000	0.013053	0.378790	-0.053788	0.232446	-0.065846	-0.405355	0.257420	-0.070413	-0.0
Infant mortality (per 1000 births)	0.002438	0.002924	-0.143214	-0.105956	0.013053	1.000000	-0.639090	-0.761224	-0.699199	-0.123033	-0.095712	0.148600	-0.366672	0.8
GDP ($ per capita)	-0.033618	0.068356	0.190122	0.035815	0.378790	-0.639090	1.000000	0.522880	0.883520	0.046465	-0.207844	0.066445	0.360567	-0.6
Literacy (%)	-0.038162	0.000682	0.086090	0.099605	-0.053788	-0.761224	0.522880	1.000000	0.592042	0.086519	0.060741	-0.101167	0.395194	-0.7
Phones (per 1000)	-0.003942	0.077864	0.157072	0.098367	0.232446	-0.699199	0.883520	0.592042	1.000000	0.124116	-0.124819	-0.038643	0.410691	-0.7
Arable (%)	0.206667	-0.053747	-0.059223	-0.077800	-0.065846	-0.123033	0.046465	0.086519	0.124116	1.000000	0.098265	-0.866058	0.392914	-0.1
Crops (%)	-0.062567	-0.160433	-0.036580	0.399358	-0.405355	-0.095712	-0.207844	0.060741	-0.124819	0.098265	1.000000	-0.582627	-0.003734	0.0
Other (%)	-0.137345	0.124528	0.066753	-0.137085	0.257420	0.148600	0.066445	-0.101167	-0.038643	-0.866058	-0.582627	1.000000	-0.318964	0.1
Climate	-0.018471	-0.094852	-0.012370	-0.027063	-0.070413	-0.366672	0.360567	0.395194	0.410691	0.392914	-0.003734	-0.318964	1.000000	-0.4
Birthrate	-0.064719	-0.037473	-0.174565	-0.063464	-0.035102	0.862113	-0.658795	-0.788349	-0.732985	-0.198438	0.075813	0.123943	-0.456312	1.0
Deathrate	-0.050578	-0.024266	-0.130624	-0.148592	0.042805	0.665729	-0.247562	-0.401696	-0.317530	0.047770	-0.208984	0.066027	0.021979	0.4
Agriculture	-0.007401	-0.017035	-0.144315	-0.032327	-0.096617	0.758537	-0.616919	-0.620514	-0.631578	-0.018610	0.084289	-0.027108	-0.187472	0.7
Industry	0.092468	0.103225	-0.145370	-0.188972	-0.004402	-0.085310	0.032855	0.105703	-0.084247	-0.073380	-0.124211	0.122303	-0.077286	-0.1
Service	-0.070320	-0.070204	0.255477	0.190004	0.091498	-0.618259	0.536551	0.474395	0.649638	0.081982	0.029020	-0.081546	0.237414	-0.5

Birthrate	Deathrate	Agriculture	Industry	Service
-0.064719	-0.050578	-0.007401	0.092468	-0.070320
-0.037473	-0.024266	-0.017035	0.103225	-0.070204
-0.174565	-0.130624	-0.144315	-0.145370	0.255477
-0.063464	-0.148592	-0.032327	-0.188972	0.190004
-0.035102	0.042805	-0.096617	-0.004402	0.091498
0.862113	0.665729	0.758537	-0.085310	-0.618259
-0.658795	-0.247562	-0.616919	0.032855	0.536551
-0.788349	-0.401696	-0.620514	0.105703	0.474395
-0.732985	-0.317530	-0.631578	-0.084247	0.649638
-0.198438	0.047770	-0.018610	-0.073380	0.081982
0.075813	-0.208984	0.084289	-0.124211	0.029020
0.123943	0.066027	-0.027108	0.122303	-0.081546
-0.456312	0.021979	-0.187472	-0.077286	0.237414
1.000000	0.446220	0.703979	-0.120518	-0.541710
0.446220	1.000000	0.416409	-0.012611	-0.366187
0.703979	0.416409	1.000000	-0.352785	-0.613489
-0.120518	-0.012611	-0.352785	1.000000	-0.521413
-0.541710	-0.366187	-0.613489	-0.521413	1.000000

Корреляционная матрица содержит коэффициенты корреляции между всеми парами признаков.

Корреляционная матрица симметрична относительно главной диагонали. На главной диагонали расположены единицы (корреляция признака самого с собой).

На основе корреляционной матрицы можно сделать следующие выводы:

- Уровень рождаемости наиболее сильно коррелирует с младенческой смертностью (на 1000 рождений) (0.845764) и сельским хозяйством (0.678207). Эти признаки обязательно следует оставить в модели.
- Уровень рождаемости отчасти коррелирует с уровнем смертности (0.395302). Этот признак стоит также оставить в модели.
- Уровень рождаемости слабо коррелирует с культурой (%) (0.120687) и другими параметрами (0.088586). Скорее всего эти признаки стоит исключить из модели, возможно они только ухудшат качество модели.

Построение корреляционной матрицы на основе коэффициентов корреляции Пирсона, Кендалла и Спирмена.

```
data.corr(method='pearson')
```

	Population	Area (sq. mi.)	Pop. Density (per sq. mi.)	Coastline (coast/area ratio)	Net migration	Infant mortality (per 1000 births)	GDP ($ per capita)	Literacy (%)	Phones (per 1000)	Arable (%)	Crops (%)	Other (%)	Climate
Population	1.000000	0.610850	-0.019010	-0.054617	0.010332	0.002438	-0.033618	-0.038162	-0.003942	0.206667	-0.062567	-0.137345	-0.018471
Area (sq. mi.)	0.610850	1.000000	-0.069010	-0.088162	0.052730	0.002924	0.068356	0.000682	0.077864	-0.053747	-0.160433	0.124528	-0.094852
Pop. Density (per sq. mi.)	-0.019010	-0.069010	1.000000	0.164036	0.161705	-0.143214	0.190122	0.086090	0.157072	-0.059223	-0.036580	0.066753	-0.012370
Coastline (coast/area ratio)	-0.054617	-0.088162	0.164036	1.000000	-0.241629	-0.105956	0.035815	0.099605	0.098367	-0.077800	0.399358	-0.137085	-0.027063
Net migration	0.010332	0.052730	0.161705	-0.241629	1.000000	0.013053	0.378790	-0.053788	0.232446	-0.065846	-0.405355	0.257420	-0.070413
Infant mortality (per 1000 births)	0.002438	0.002924	-0.143214	-0.105956	0.013053	1.000000	-0.639090	-0.761224	-0.699199	-0.123033	-0.095712	0.148600	-0.366672
GDP ($ per capita)	-0.033618	0.068356	0.190122	0.035815	0.378790	-0.639090	1.000000	0.522880	0.883520	0.046465	-0.207844	0.066445	0.360567
Literacy (%)	-0.038162	0.000682	0.086090	0.099605	-0.053788	-0.761224	0.522880	1.000000	0.592042	0.086519	0.060741	-0.101167	0.395194
Phones (per 1000)	-0.003942	0.077864	0.157072	0.098367	0.232446	-0.699199	0.883520	0.592042	1.000000	0.124116	-0.124819	-0.038643	0.410691
Arable (%)	0.206667	-0.053747	-0.059223	-0.077800	-0.065846	-0.123033	0.046465	0.086519	0.124116	1.000000	0.098265	-0.866058	0.392914
Crops (%)	-0.062567	-0.160433	-0.036580	0.399358	-0.405355	-0.095712	-0.207844	0.060741	-0.124819	0.098265	1.000000	-0.582627	-0.003734
Other (%)	-0.137345	0.124528	0.066753	-0.137085	0.257420	0.148600	0.066445	-0.101167	-0.038643	-0.866058	-0.582627	1.000000	-0.318964
Climate	-0.018471	-0.094852	-0.012370	-0.027063	-0.070413	-0.366672	0.360567	0.395194	0.410691	0.392914	-0.003734	-0.318964	1.000000
Birthrate	-0.064719	-0.037473	-0.174565	-0.063464	-0.035102	0.862113	-0.658795	-0.788349	-0.732985	-0.198438	0.075813	0.123943	-0.456312
Deathrate	-0.050578	-0.024266	-0.130624	-0.148592	0.042805	0.665729	-0.247562	-0.401696	-0.317530	0.047770	-0.208984	0.066027	0.021979
Agriculture	-0.007401	-0.017035	-0.144315	-0.032327	-0.096617	0.758537	-0.616919	-0.620514	-0.631578	-0.018610	0.084289	-0.027108	-0.187472
Industry	0.092468	0.103225	-0.145370	-0.188972	-0.004402	-0.085310	0.032855	0.105703	-0.084247	-0.073380	-0.124211	0.122303	-0.077286
Service	-0.070320	-0.070204	0.255477	0.190004	0.091498	-0.618259	0.536551	0.474395	0.649638	0.081982	0.029020	-0.081546	0.237414

Birthrate	Deathrate	Agriculture	Industry	Service
-0.064719	-0.050578	-0.007401	0.092468	-0.070320
-0.037473	-0.024266	-0.017035	0.103225	-0.070204
-0.174565	-0.130624	-0.144315	-0.145370	0.255477
-0.063464	-0.148592	-0.032327	-0.188972	0.190004
-0.035102	0.042805	-0.096617	-0.004402	0.091498
0.862113	0.665729	0.758537	-0.085310	-0.618259
-0.658795	-0.247562	-0.616919	0.032855	0.536551
-0.788349	-0.401696	-0.620514	0.105703	0.474395
-0.732985	-0.317530	-0.631578	-0.084247	0.649638
-0.198438	0.047770	-0.018610	-0.073380	0.081982
0.075813	-0.208984	0.084289	-0.124211	0.029020
0.123943	0.066027	-0.027108	0.122303	-0.081546
-0.456312	0.021979	-0.187472	-0.077286	0.237414
1.000000	0.446220	0.703979	-0.120518	-0.541710
0.446220	1.000000	0.416409	-0.012611	-0.366187
0.703979	0.416409	1.000000	-0.352785	-0.613489
-0.120518	-0.012611	-0.352785	1.000000	-0.521413
-0.541710	-0.366187	-0.613489	-0.521413	1.000000

```
data.corr(method='kendall')
```

	Population	Area (sq. mi.)	Pop. Density (per sq. mi.)	Coastline (coast/area ratio)	Net migration	Infant mortality (per 1000 births)	GDP ($ per capita)	Literacy (%)	Phones (per 1000)	Arable (%)	Crops (%)	Other (%)	Climate
Population	1.000000	0.609566	-0.000879	-0.350371	0.020548	0.104711	-0.096495	-0.096190	-0.150615	0.204508	0.002205	-0.107355	0.006608
Area (sq. mi.)	0.609566	1.000000	-0.391399	-0.494849	0.002805	0.208418	-0.162883	-0.131616	-0.239515	-0.065239	-0.226123	0.167624	-0.110455
Pop. Density (per sq. mi.)	-0.000879	-0.391399	1.000000	0.302599	0.006003	-0.224350	0.148114	0.094761	0.221169	0.415387	0.351894	-0.468006	0.153128
Coastline (coast/area ratio)	-0.350371	-0.494849	0.302599	1.000000	0.000599	-0.349187	0.287507	0.176404	0.345505	-0.022998	0.241243	-0.102226	0.058593
Net migration	0.020548	0.002805	0.006003	0.000599	1.000000	-0.161402	0.225566	0.048999	0.136689	-0.082546	-0.258303	0.161412	0.069494
Infant mortality (per 1000 births)	0.104711	0.208418	-0.224350	-0.349187	-0.161402	1.000000	-0.705327	-0.551918	-0.712398	-0.083142	-0.010838	0.086834	-0.342567
GDP ($ per capita)	-0.096495	-0.162883	0.148114	0.287507	0.225566	-0.705327	1.000000	0.493359	0.723852	0.021926	-0.088257	0.000315	0.265482
Literacy (%)	-0.096190	-0.131616	0.094761	0.176404	0.048999	-0.551918	0.493359	1.000000	0.535389	0.144268	-0.020373	-0.120856	0.408439
Phones (per 1000)	-0.150615	-0.239515	0.221169	0.345505	0.136689	-0.712398	0.723852	0.535389	1.000000	0.080701	-0.029681	-0.073718	0.294758
Arable (%)	0.204508	-0.065239	0.415387	-0.022998	-0.082546	-0.083142	0.021926	0.144268	0.080701	1.000000	0.299491	-0.796496	0.336347
Crops (%)	0.002205	-0.226123	0.351894	0.241243	-0.258303	-0.010838	-0.088257	-0.020373	-0.029681	0.299491	1.000000	-0.504052	0.072252
Other (%)	-0.107355	0.167624	-0.468006	-0.102226	0.161412	0.086834	0.000315	-0.120856	-0.073718	-0.796496	-0.504052	1.000000	-0.296323
Climate	0.006608	-0.110455	0.153128	0.058593	0.069494	-0.342567	0.265482	0.408439	0.294758	0.336347	0.072252	-0.296323	1.000000
Birthrate	0.045009	0.151099	-0.218880	-0.248177	-0.096355	0.672254	-0.641369	-0.601260	-0.673134	-0.171994	0.013673	0.141642	-0.401170
Deathrate	0.128046	0.188458	-0.139317	-0.327845	0.033610	0.310002	-0.250671	-0.069368	-0.262232	0.137257	-0.144155	-0.045599	0.133328
Agriculture	0.114928	0.191294	-0.173153	-0.307737	-0.196053	0.593754	-0.679186	-0.395566	-0.608847	0.030267	0.101276	-0.046935	-0.146721
Industry	0.155438	0.176062	-0.057925	-0.150758	-0.018754	-0.057288	0.134457	0.077540	0.045407	-0.029437	-0.031115	0.055783	0.005674
Service	-0.196827	-0.277141	0.199447	0.376810	0.102796	-0.467413	0.427104	0.322543	0.488765	0.062361	-0.004794	-0.080767	0.196356

Birthrate	Deathrate	Agriculture	Industry	Service
0.045009	0.128046	0.114928	0.155438	-0.196827
0.151099	0.188458	0.191294	0.176062	-0.277141
-0.218880	-0.139317	-0.173153	-0.057925	0.199447
-0.248177	-0.327845	-0.307737	-0.150758	0.376810
-0.096355	0.033610	-0.196053	-0.018754	0.102796
0.672254	0.310002	0.593754	-0.057288	-0.467413
-0.641369	-0.250671	-0.679186	0.134457	0.427104
-0.601260	-0.069368	-0.395566	0.077540	0.322543
-0.673134	-0.262232	-0.608847	0.045407	0.488765
-0.171994	0.137257	0.030267	-0.029437	0.062361
0.013673	-0.144155	0.101276	-0.031115	-0.004794
0.141642	-0.045599	-0.046935	0.055783	-0.080767
-0.401170	0.133328	-0.146721	0.005674	0.196356
1.000000	0.134962	0.525605	-0.116208	-0.374823
0.134962	1.000000	0.235557	-0.045796	-0.209110
0.525605	0.235557	1.000000	-0.193831	-0.463016
-0.116208	-0.045796	-0.193831	1.000000	-0.345420
-0.374823	-0.209110	-0.463016	-0.345420	1.000000

```
data.corr(method='spearman')
```

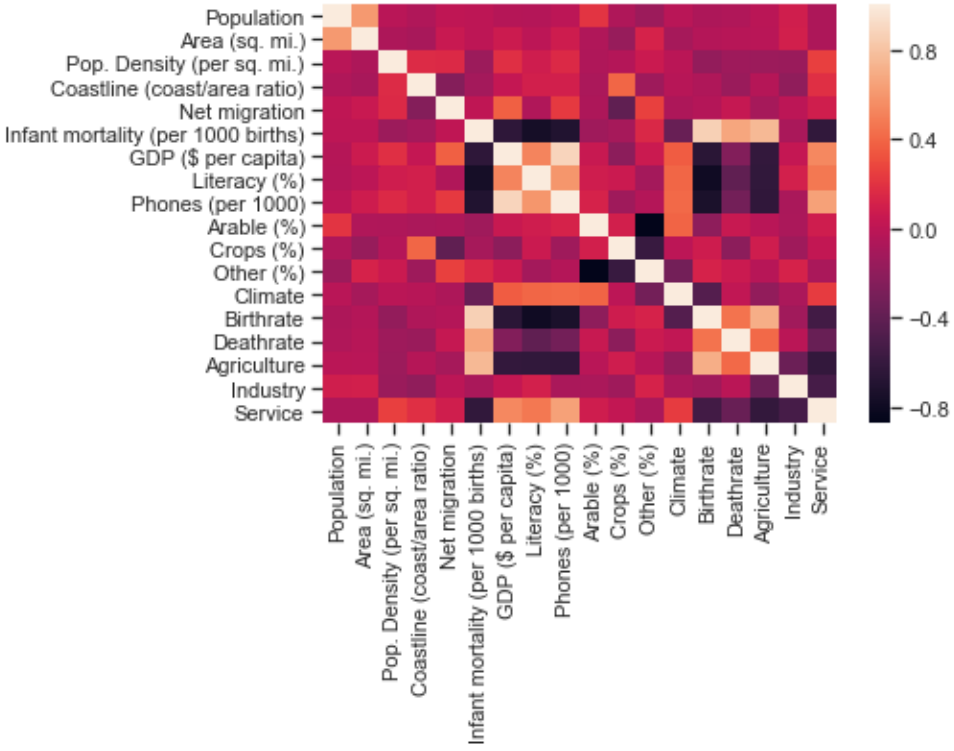
	Population	Area (sq. mi.)	Pop. Density (per sq. mi.)	Coastline (coast/area ratio)	Net migration	Infant mortality (per 1000 births)	GDP ($ per capita)	Literacy (%)	Phones (per 1000)	Arable (%)	Crops (%)	Other (%)	Climate
Population	1.000000	0.795204	-0.023268	-0.503016	0.030561	0.168955	-0.152863	-0.140565	-0.223394	0.302188	-0.009519	-0.148898	0.006739
Area (sq. mi.)	0.795204	1.000000	-0.557393	-0.649206	0.005283	0.326327	-0.244485	-0.199954	-0.350405	-0.100702	-0.310975	0.248292	-0.158786
Pop. Density (per sq. mi.)	-0.023268	-0.557393	1.000000	0.441155	0.006185	-0.321955	0.216487	0.144197	0.320258	0.570631	0.463102	-0.628396	0.203230
Coastline (coast/area ratio)	-0.503016	-0.649206	0.441155	1.000000	0.004944	-0.497061	0.411947	0.242277	0.491441	-0.030193	0.322760	-0.153607	0.080474
Net migration	0.030561	0.005283	0.006185	0.004944	1.000000	-0.279791	0.345671	0.089986	0.237618	-0.115199	-0.363960	0.231869	0.088735
Infant mortality (per 1000 births)	0.168955	0.326327	-0.321955	-0.497061	-0.279791	1.000000	-0.883782	-0.730086	-0.895244	-0.119394	-0.007476	0.131096	-0.445971
GDP ($ per capita)	-0.152863	-0.244485	0.216487	0.411947	0.345671	-0.883782	1.000000	0.674936	0.904129	0.032803	-0.132024	0.002233	0.347669
Literacy (%)	-0.140565	-0.199954	0.144197	0.242277	0.089986	-0.730086	0.674936	1.000000	0.738922	0.206588	-0.033601	-0.179432	0.512245
Phones (per 1000)	-0.223394	-0.350405	0.320258	0.491441	0.237618	-0.895244	0.904129	0.738922	1.000000	0.123371	-0.045839	-0.112715	0.394826
Arable (%)	0.302188	-0.100702	0.570631	-0.030193	-0.115199	-0.119394	0.032803	0.206588	0.123371	1.000000	0.438073	-0.911240	0.429039
Crops (%)	-0.009519	-0.310975	0.463102	0.322760	-0.363960	-0.007476	-0.132024	-0.033601	-0.045839	0.438073	1.000000	-0.689492	0.097596
Other (%)	-0.148898	0.248292	-0.628396	-0.153607	0.231869	0.131096	0.002233	-0.179432	-0.112715	-0.911240	-0.689492	1.000000	-0.383365
Climate	0.006739	-0.158786	0.203230	0.080474	0.088735	-0.445971	0.347669	0.512245	0.394826	0.429039	0.097596	-0.383365	1.000000
Birthrate	0.057770	0.230564	-0.325656	-0.357666	-0.163307	0.861648	-0.829505	-0.794650	-0.865237	-0.247097	0.021211	0.200296	-0.527452
Deathrate	0.199848	0.279218	-0.214726	-0.484122	0.052162	0.451503	-0.362430	-0.169760	-0.399334	0.193958	-0.213869	-0.061426	0.169237
Agriculture	0.177593	0.282596	-0.256509	-0.436980	-0.298429	0.789061	-0.863334	-0.562316	-0.802600	0.048188	0.150520	-0.068935	-0.194993
Industry	0.243111	0.254111	-0.081924	-0.225167	-0.028733	-0.088053	0.190807	0.119405	0.076780	-0.045884	-0.049312	0.081688	-0.002407
Service	-0.290119	-0.396463	0.292864	0.533687	0.151359	-0.667724	0.600989	0.473232	0.680298	0.088055	-0.007302	-0.114191	0.261136

Birthrate	Deathrate	Agriculture	Industry	Service
0.057770	0.199848	0.177593	0.243111	-0.290119
0.230564	0.279218	0.282596	0.254111	-0.396463
-0.325656	-0.214726	-0.256509	-0.081924	0.292864
-0.357666	-0.484122	-0.436980	-0.225167	0.533687
-0.163307	0.052162	-0.298429	-0.028733	0.151359
0.861648	0.451503	0.789061	-0.088053	-0.667724
-0.829505	-0.362430	-0.863334	0.190807	0.600989
-0.794650	-0.169760	-0.562316	0.119405	0.473232
-0.865237	-0.399334	-0.802600	0.076780	0.680298
-0.247097	0.193958	0.048188	-0.045884	0.088055
0.021211	-0.213869	0.150520	-0.049312	-0.007302
0.200296	-0.061426	-0.068935	0.081688	-0.114191
-0.527452	0.169237	-0.194993	-0.002407	0.261136
1.000000	0.258810	0.711234	-0.174757	-0.548527
0.258810	1.000000	0.347700	-0.066265	-0.302709
0.711234	0.347700	1.000000	-0.278816	-0.620622
-0.174757	-0.066265	-0.278816	1.000000	-0.435904
-0.548527	-0.302709	-0.620622	-0.435904	1.000000

Для визуализации корреляционной матрицы использована "тепловая карта" heatmap, которая показывает степень корреляции различными цветами.

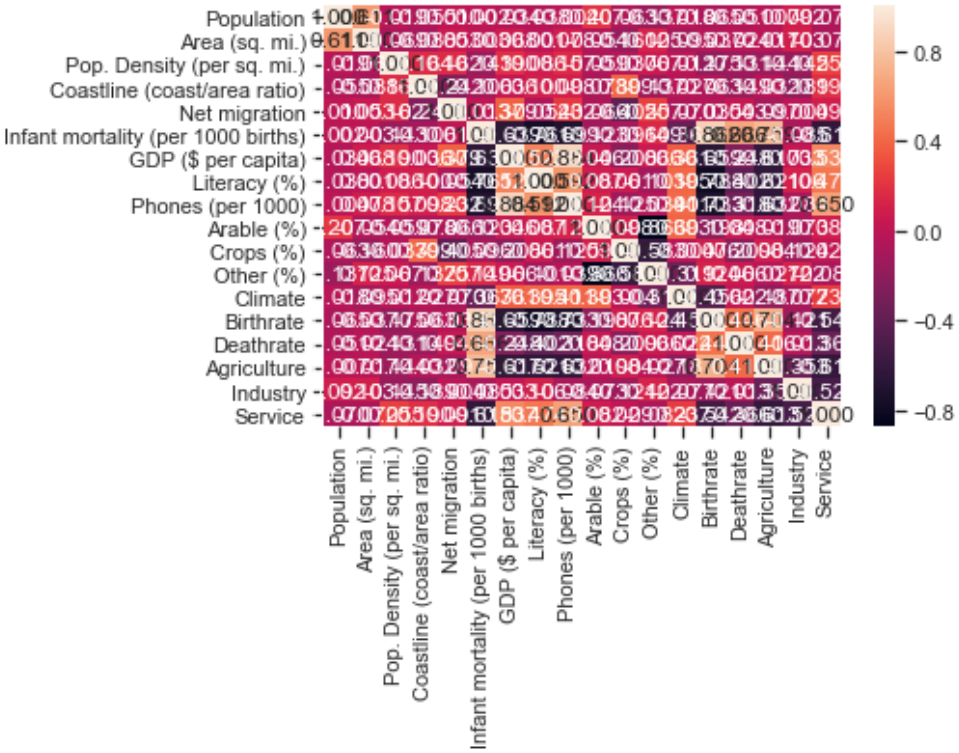
```
sns.heatmap(data.corr())
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x137d1aa58>
```



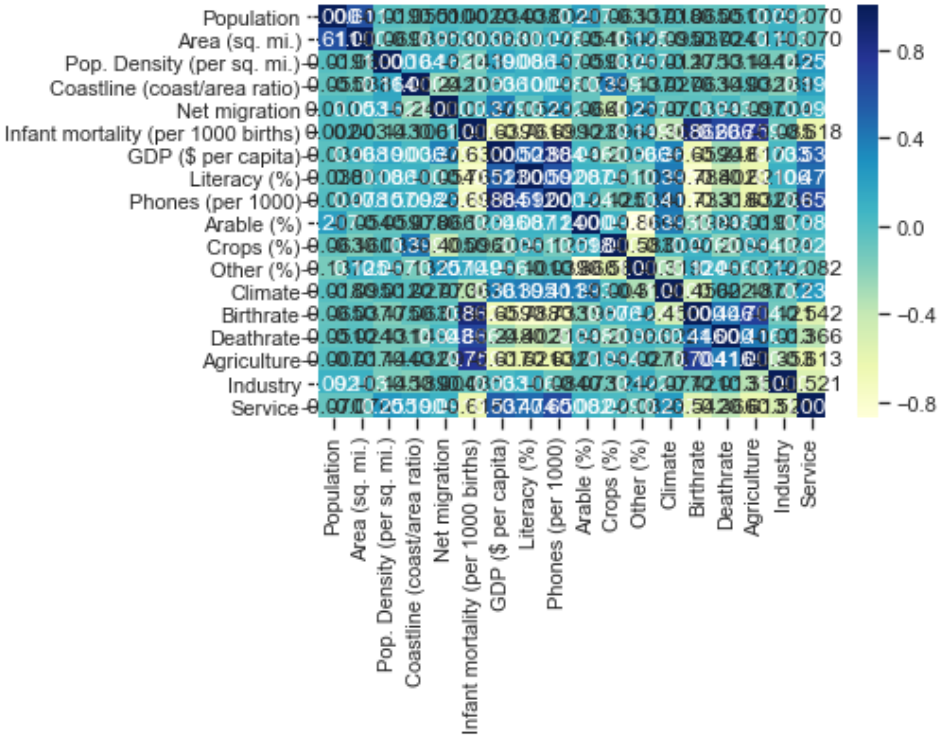

```
# Вывод значений в ячейках
sns.heatmap(data.corr(), annot=True, fmt='.3f')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x13032be48>
```



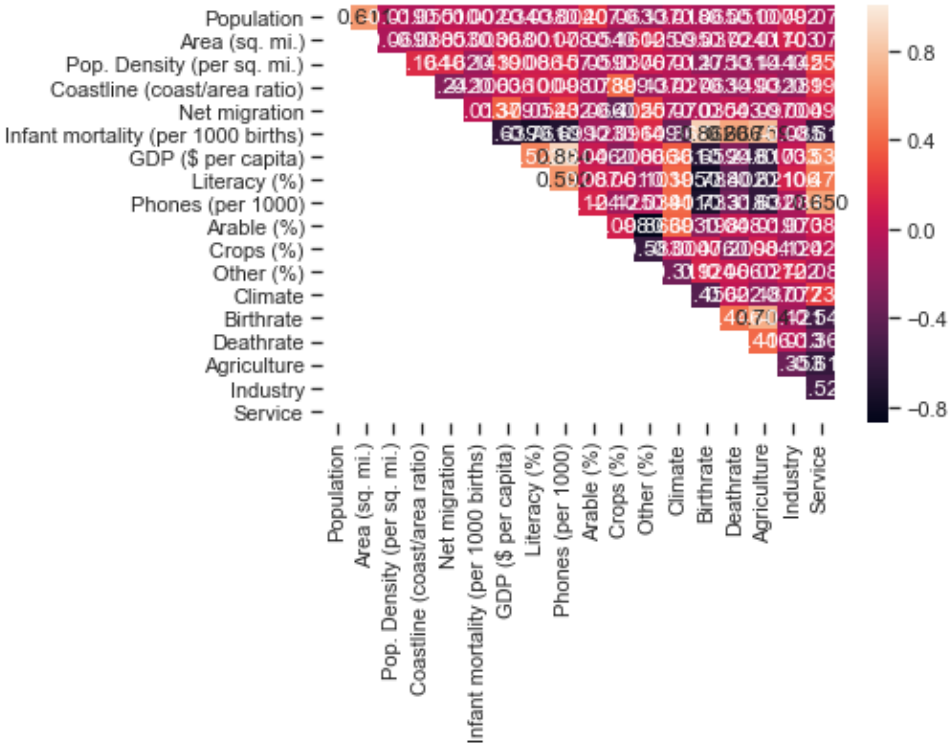

```
# Изменение цветовой гаммы
sns.heatmap(data.corr(), cmap='YlGnBu', annot=True, fmt='.3f')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x135b8d4a8>
```

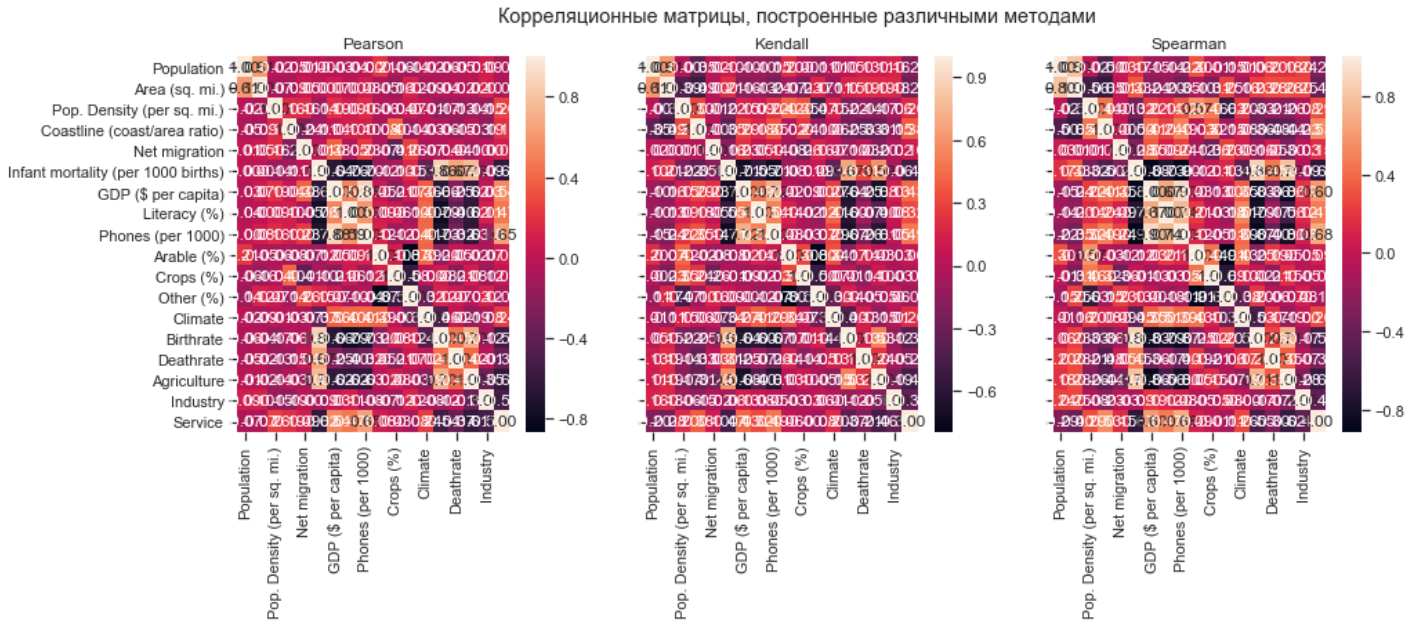



```
# Треугольный вариант матрицы
mask = np.zeros_like(data.corr(), dtype=np.bool)
# Верхняя часть матрицы
mask[np.tril_indices_from(mask)] = True
sns.heatmap(data.corr(), mask=mask, annot=True, fmt='.3f')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1356084a8>
```




```
fig, ax = plt.subplots(1, 3, sharex='col', sharey='row', figsize=(15,5))
sns.heatmap(data.corr(method='pearson'), ax=ax[0], annot=True, fmt='.2f')
sns.heatmap(data.corr(method='kendall'), ax=ax[1], annot=True, fmt='.2f')
sns.heatmap(data.corr(method='spearman'), ax=ax[2], annot=True, fmt='.2f')
fig.suptitle('Корреляционные матрицы, построенные различными методами')
ax[0].title.set_text('Pearson')
ax[1].title.set_text('Kendall')
ax[2].title.set_text('Spearman')
```



Тепловая карта не очень хорошо подходит для определения корреляции нецелевых признаков между собой.

Здесь тепловая карта помогает определить значимую корреляцию между признаками Birthrate и Infant mortality, следовательно только один из этих признаков можно включать в модель.