In [124]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
```

In [158]:
```
df = pd.read_csv('suicide-rates-overview.csv')
sns.set(style="whitegrid")
df.head()
```

Out[158]:

	country	year	sex	age	suicides_no	population	suicides/100k pop	country-year	HDI for year	gdp_f
0	Albania	1987	male	15-24 years	21	312900	6.71	Albania1987	nan	2,156,
1	Albania	1987	male	35-54 years	16	308000	5.19	Albania1987	nan	2,156,
2	Albania	1987	female	15-24 years	14	289700	4.83	Albania1987	nan	2,156,
3	Albania	1987	male	75+ years	1	21800	4.59	Albania1987	nan	2,156,
4	Albania	1987	male	25-34 years	9	274300	3.28	Albania1987	nan	2,156,

Исследование данных о суицидах по половозрастным группам (с 1985 по 2015гг.)

на примере трех стран - Мексика, Россия и США. Цель работы - анализ и сравнение

Проверка и обработка пропущенных значений

In [126]:
```
df = df.fillna(0)
df.isna().sum()
```

Out[126]:
```
country               0
year                  0
sex                   0
age                   0
suicides_no           0
population            0
suicides/100k pop     0
country-year          0
HDI for year          0
 gdp_for_year ($)     0
gdp_per_capita ($)    0
generation            0
dtype: int64
```

In [127]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27820 entries, 0 to 27819
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   country             27820 non-null  object
 1   year                27820 non-null  int64
 2   sex                 27820 non-null  object
 3   age                 27820 non-null  object
 4   suicides_no         27820 non-null  int64
 5   population          27820 non-null  int64
```

```
 6   suicides/100k pop   27820 non-null  float64
 7   country-year        27820 non-null  object
 8   HDI for year        27820 non-null  float64
 9    gdp_for_year ($)   27820 non-null  object
 10  gdp_per_capita ($)  27820 non-null  int64
 11  generation          27820 non-null  object
dtypes: float64(2), int64(4), object(6)
memory usage: 1.9+ MB
```

Файл содержит 27 820 строк × 12 колонок, 0 пропущенных значений

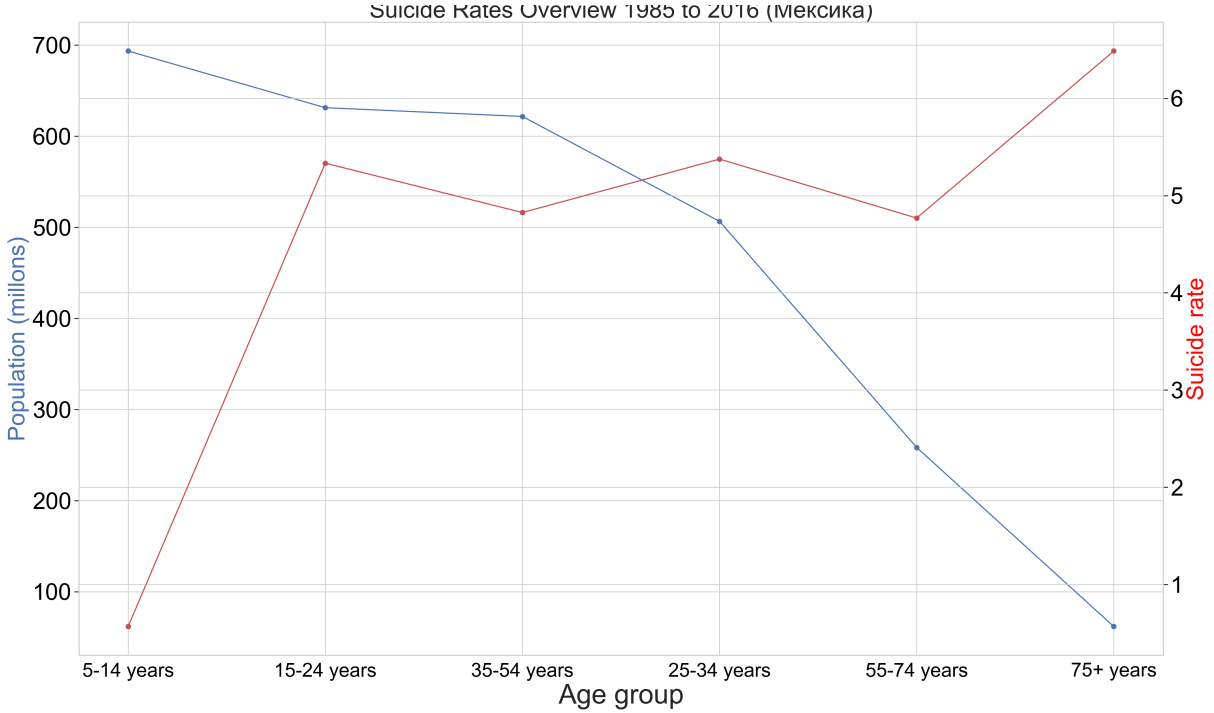
Мексика

In [128]:
```
table_mexico = df[df.country=='Mexico'].groupby(by='age').sum().sort_values(by='population', ascending=False)
mexico_data = table_mexico.drop(['year', 'gdp_per_capita ($)', 'HDI for year', 'suicides/100k pop'], axis=1)
mexico_data['suicide_rate'] = (mexico_data['suicides_no'] / (mexico_data['population'] / 100000))


mexico_data
```

Out[128]:

age	suicides_no	population	suicide_rate
5-14 years	3930	693477763	0.57
15-24 years	33664	631268656	5.33
35-54 years	29997	621648773	4.83
25-34 years	27226	506518247	5.38
55-74 years	12318	258319544	4.77
75+ years	4004	61726176	6.49

In [129]:
```
labels_1 = mexico_data.index[:10]
values_1 = mexico_data['population'] / 1000000
suicides_1 = mexico_data['suicide_rate']
Legend = ['Суициды', 'Население']


fig, ax1 = plt.subplots(figsize=(25, 15), dpi=200)
plt.title('Suicide Rates Overview 1985 to 2016 (Мексика)', fontsize=30)


ax1.set_xlabel('Age group',fontsize=35 )
ax1.set_ylabel('Population (millons)', fontsize=30, color='b')

ax1.plot(labels_1, values_1, color='b', marker='o')

ax1.tick_params(axis='y',labelsize=30, labelcolor='black')
ax1.tick_params(axis='x',labelsize=25, labelcolor='black')
ax2 = ax1.twinx()

ax2.set_ylabel('Suicide rate', fontsize=30, color='red')

ax2.plot(labels_1, suicides_1, color='r', marker='o')
ax2.tick_params(axis='y', labelsize=30, labelcolor='black')


plt.show()
```



Преобладет население в возрасте 5-14 лет, население Мексики довольно молодое. Минимальное количество суицидов приходится на 5-14 лет, а максимальное на 75+

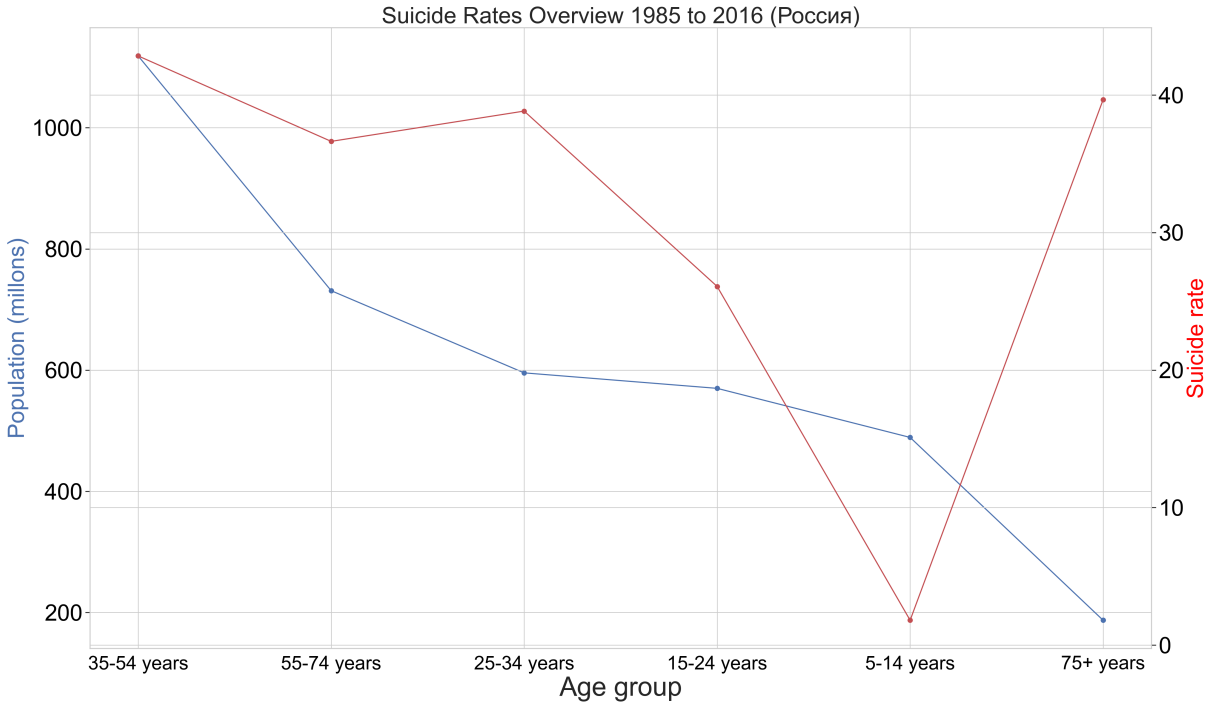
Данные по России

In [130]:
```
russian_tab = df[df.country=='Russian Federation'].groupby(by='age').su
m().sort_values(by='population', ascending=False)
russian_data = russian_tab.drop(['year', 'gdp_per_capita ($)', 'HDI for
 year', 'suicides/100k pop'], axis=1)
russian_data['suicide_rate'] = (russian_data['suicides_no'] / (russian_
data['population'] / 100000))

russian_data
```

Out[130]:

age	suicides_no	population	suicide_rate
35-54 years	479140	1118486996	42.84
55-74 years	267753	731053070	36.63
25-34 years	231187	595413982	38.83
15-24 years	148611	569864937	26.08
5-14 years	8840	488859625	1.81
75+ years	74211	187124010	39.66

In [131]:
```
labels_2 = russian_data.index[:10]
values_2 = russian_data['population'] / 1000000
suicides_2 = russian_data['suicide_rate']

fig, ax1 = plt.subplots(figsize=(25, 15), dpi=200)
plt.title('Suicide Rates Overview 1985 to 2016 (Россия)', fontsize=30)


ax1.set_xlabel('Age group',fontsize=35 )
ax1.set_ylabel('Population (millons)', fontsize=30, color='b')

ax1.plot(labels_2, values_2, color='b', marker='o')

ax1.tick_params(axis='y',labelsize=30, labelcolor='black')
ax1.tick_params(axis='x',labelsize=25, labelcolor='black')
ax2 = ax1.twinx()

ax2.set_ylabel('Suicide rate', fontsize=30, color='red')

ax2.plot(labels_2, suicides_2, color='r', marker='o')
ax2.tick_params(axis='y', labelsize=30, labelcolor='black')


plt.show()
```



В России преобладет население в возрасте 35-54 лет на которое приходится максимальная величина само, второе по величине число самоубийств приходится на молодой возраст 25-34 лет, минимальное количество суицидов приходится на 5-14 лет, а максимальное на 75+

Данные по США

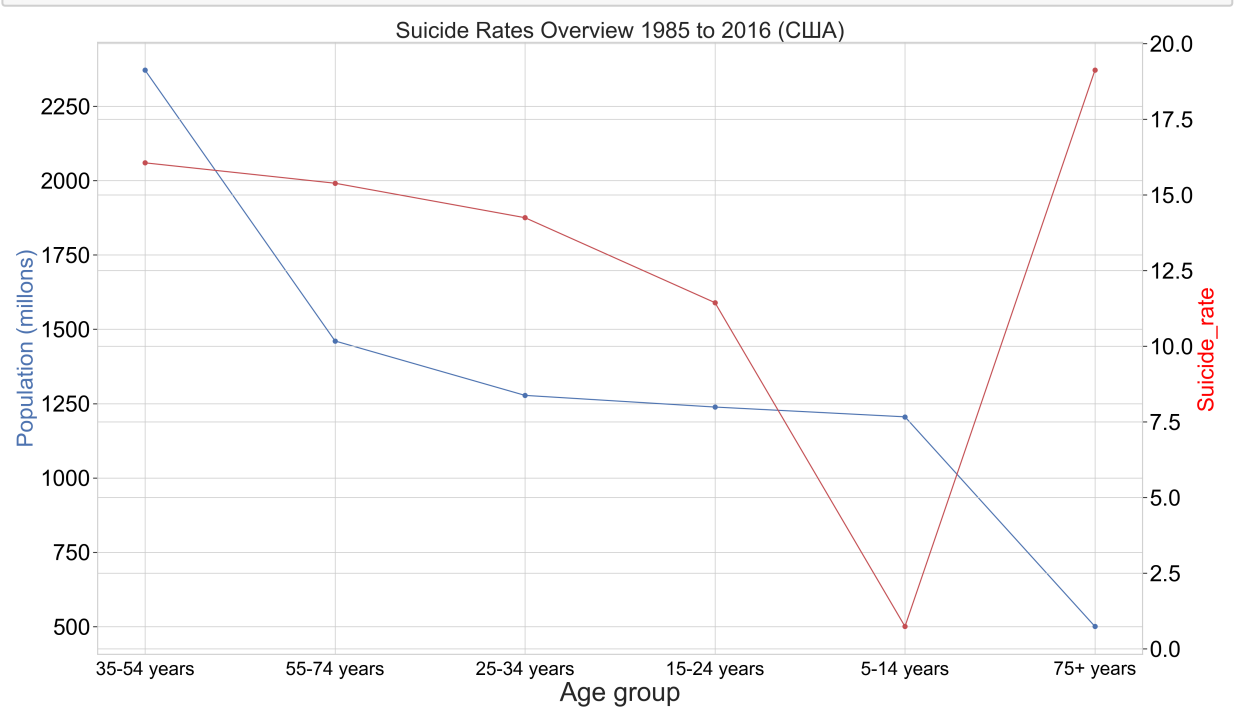
In [132]:
```
usa_tab = df[df.country=='United States'].groupby(by='age').sum().sort_
values(by='population', ascending=False)
usa_data = usa_tab.drop(['year', 'gdp_per_capita ($)', 'HDI for year',
'suicides/100k pop'], axis=1)
usa_data['suicide_rate'] = (usa_data['suicides_no'] / (usa_data['popula
tion'] / 100000))

usa_data
```

Out[132]:

age	suicides_no	population	suicide_rate
35-54 years	380917	2371577220	16.06
55-74 years	224770	1460664960	15.39
25-34 years	182047	1277625343	14.25
15-24 years	141679	1238381995	11.44
5-14 years	8923	1205493232	0.74
75+ years	95677	500284451	19.12

In [133]:

```
labels_3 = usa_data.index[:10]
values_3 = usa_data['population'] / 1000000
suicides_3 = usa_data['suicide_rate']


fig, ax1 = plt.subplots(figsize=(25, 15), dpi=200)
plt.title('Suicide Rates Overview 1985 to 2016 (CWA)', fontsize=30)


ax1.set_xlabel('Age group',fontsize=35 )
ax1.set_ylabel('Population (millons)', fontsize=30, color='b')

ax1.plot(labels_3, values_3, color='b', marker='o')

ax1.tick_params(axis='y',labelsize=30, labelcolor='black')
ax1.tick_params(axis='x',labelsize=25, labelcolor='black')
ax2 = ax1.twinx()

ax2.set_ylabel('Suicide_rate', fontsize=30, color='red')

ax2.plot(labels_3, suicides_3, color='r', marker='o')
ax2.tick_params(axis='y', labelsize=30, labelcolor='black')


plt.show()
```

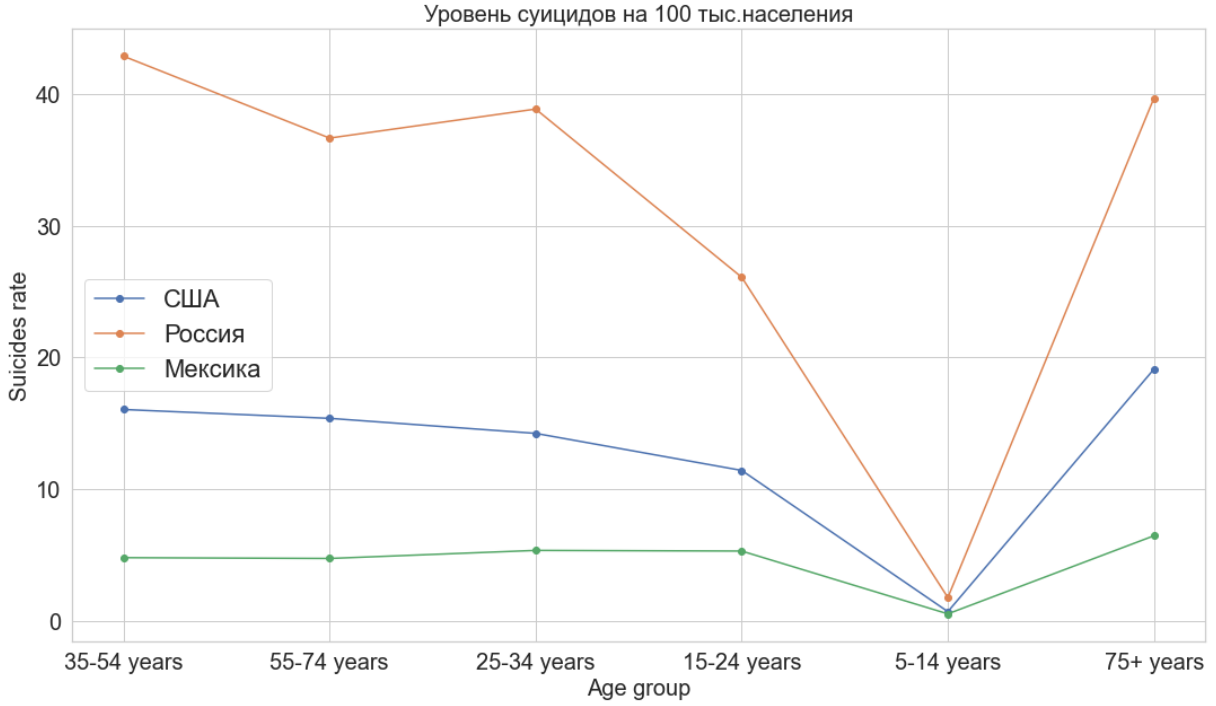


Преобладает возрастная группа 35-54, минимальное количество суицидов на группу 5-14, а максимальная -75+

#

In [134]:
```
x = usa_data.index[:10]
y_usa = 16.06, 15.39, 14.25,11.44, 0.74, 19.12
y_rus = 42.84, 36.63, 38.83, 26.08, 1.81, 39.66
y_mex = 4.83, 4.77, 5.38, 5.33, 0.57, 6.49

Legend=('США', 'Россия', 'Мексика')


plt.figure(figsize=(18, 10))
plt.title('Уровень суицидов на 100 тыс.населения', fontsize=20)
```

```
plt.plot(x, y_usa, marker='o')
plt.plot(x, y_rus, marker='o')
plt.plot(x, y_mex, marker='o')

plt.yticks(fontsize=20)
plt.xticks(fontsize=20)
plt.ylabel('Suicides rate', fontsize=20)
plt.xlabel('Age group', fontsize=20)


plt.legend(Legend, fontsize=22)
plt.show()
```

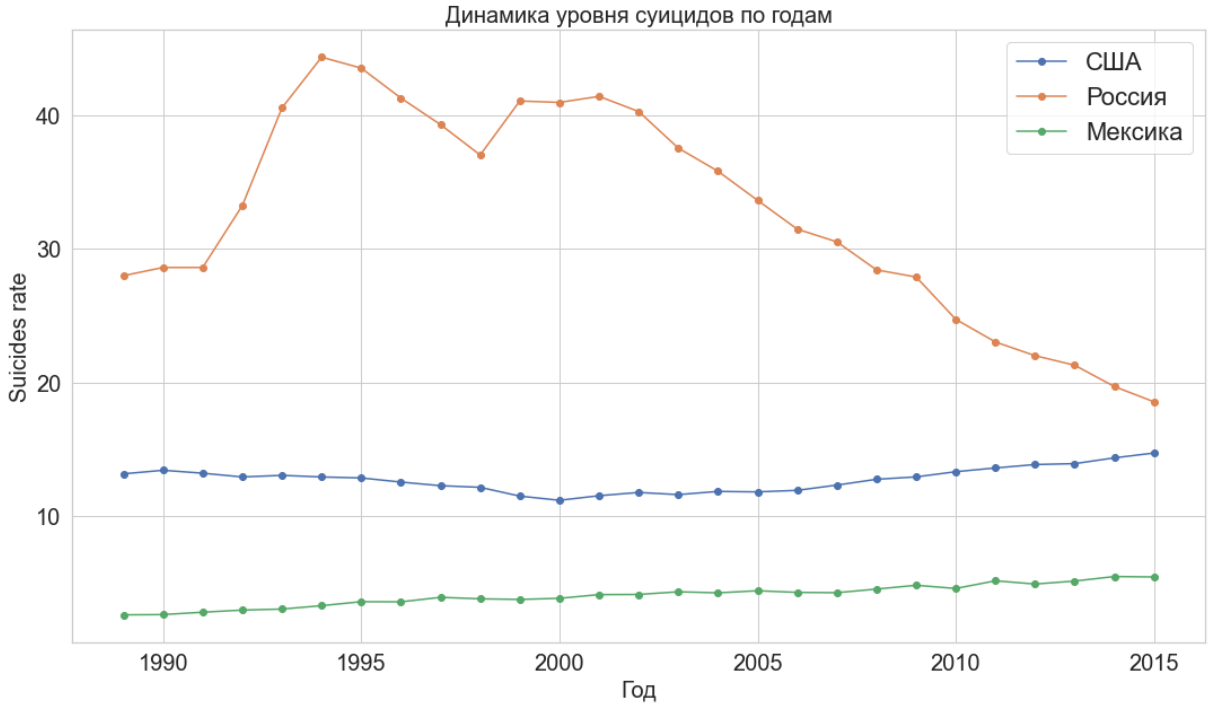


Российская Федерация лидирует по числу самоубийств на 100 тыс.населения, разрыв от США более чем в два раза. Латиноамериканская страна ничем не выделяется

#

In [264]:
```
suicides_dynamic_1 = df[df.country == 'Mexico'].groupby(by='year').sum
().drop(['suicides/100k pop', 'HDI for year', 'gdp_per_capita ($)'], ax
is=1)
suicides_dynamic_1['dynamic'] = (suicides_dynamic_1['suicides_no'] / su
icides_dynamic_1['population']) * 100000

suicides_dynamic_2 = df[df.country == 'Russian Federation'].groupby(by=
'year').sum().drop(['suicides/100k pop', 'gdp_per_capita ($)', 'HDI for
 year'], axis=1)
suicides_dynamic_2['dynamic'] = (suicides_dynamic_2['suicides_no'] / su
icides_dynamic_2['population']) * 100000

suicides_dynamic_3 = df[df.country == 'United States'].groupby(by='yea
r').sum().drop(['suicides/100k pop', 'gdp_per_capita ($)', 'HDI for yea
r'], axis=1)
suicides_dynamic_3['dynamic'] = (suicides_dynamic_3['suicides_no'] / su
icides_dynamic_3['population']) * 100000

Legend=('США', 'Россия', 'Мексика')
xs = suicides_dynamic_1.index[4:31]
xs1 = suicides_dynamic_2.index[:31]
xs2 = suicides_dynamic_3.index[4:31]

ys = suicides_dynamic_1.dynamic[4:31]
ys1 = suicides_dynamic_2.dynamic[:31]
ys2 = suicides_dynamic_3.dynamic[4:31]

plt.figure(figsize=(18, 10))
plt.title('Динамика уровня суицидов по годам', fontsize=20)


plt.plot(xs2, ys2, marker='o')
plt.plot(xs1, ys1, marker='o')
plt.plot(xs, ys, marker='o')

plt.yticks(fontsize=20)
plt.xticks(fontsize=20)
```

```
plt.ylabel('Suicides rate', fontsize=20)
plt.xlabel('Год', fontsize=20)


plt.legend(Legend, fontsize=22)
plt.show()
```



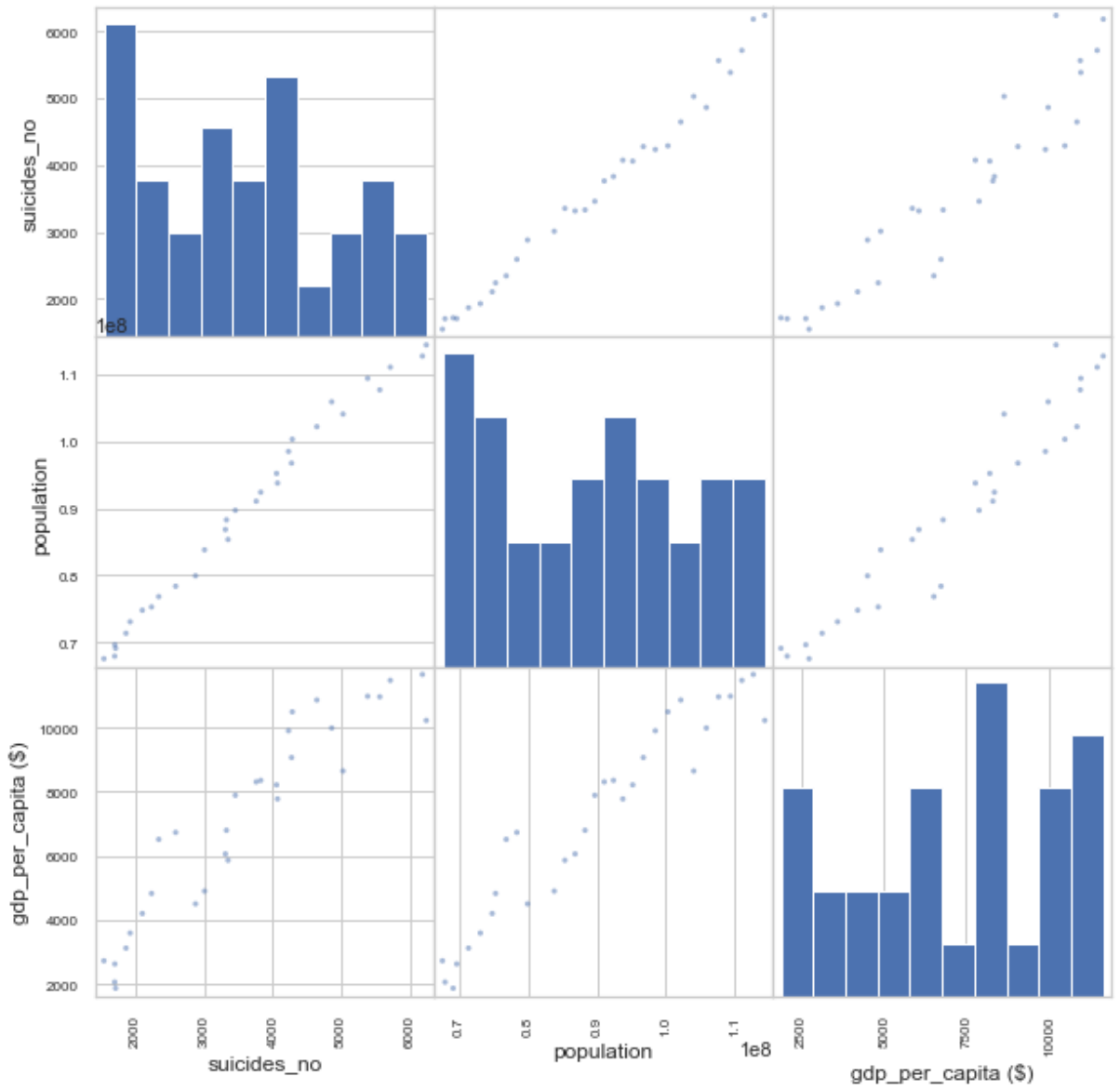
f(max) по России приходится на 1994 год

Корреляционный анализ величин по выбранным странам (количество населения, число суицидов, ВНП на душу населения). Диаграммы рассеяния

Мексика

In [190]:
```
from pandas.plotting import scatter_matrix


data_new_mexico =  df[df.country=='Mexico'].groupby(by='year').sum().sort_values(by='population', ascending=False)
datacorr_new_mexico = data_new_mexico.drop(['HDI for year', 'suicides/100k pop'], axis=1)
datacorr_new_mexico['gdp_per_capita ($)'] = datacorr_new_mexico['gdp_per_capita ($)'] / 12


scatter_matrix(datacorr_new_mexico, figsize=(10, 10))
scatter_matrix
```

Out[190]:
```
<function pandas.plotting._misc.scatter_matrix(frame, alpha=0.5, figsize=None, ax=None, grid=False, diagonal='hist', marker='.', density_kwds=None, hist_kwds=None, range_padding=0.05, **kwargs)>
```



Высокая связь между величинами, особенно между населением и уровнем суицидов

Россия

In [194]:
```
data_new_russia =  df[df.country=='Russian Federation'].groupby(by='year').sum().sort_values(by='population', ascending=False)
datacorr_new_russia = data_new_russia.drop(['HDI for year', 'suicides/100k pop'], axis=1)
datacorr_new_russia['gdp_per_capita ($)'] = datacorr_new_russia['gdp_per_capita ($)'] / 12
datacorr_new_russia


scatter_matrix(datacorr_new_russia, figsize=(10, 10))
scatter_matrix
```

Out[194]:
```
<function pandas.plotting._misc.scatter_matrix(frame, alpha=0.5, figsize=None, ax=None, grid=False, diagonal='hist', marker='.', density_kwds=None, hist_kwds=None, range_padding=0.05, **kwargs)>
```

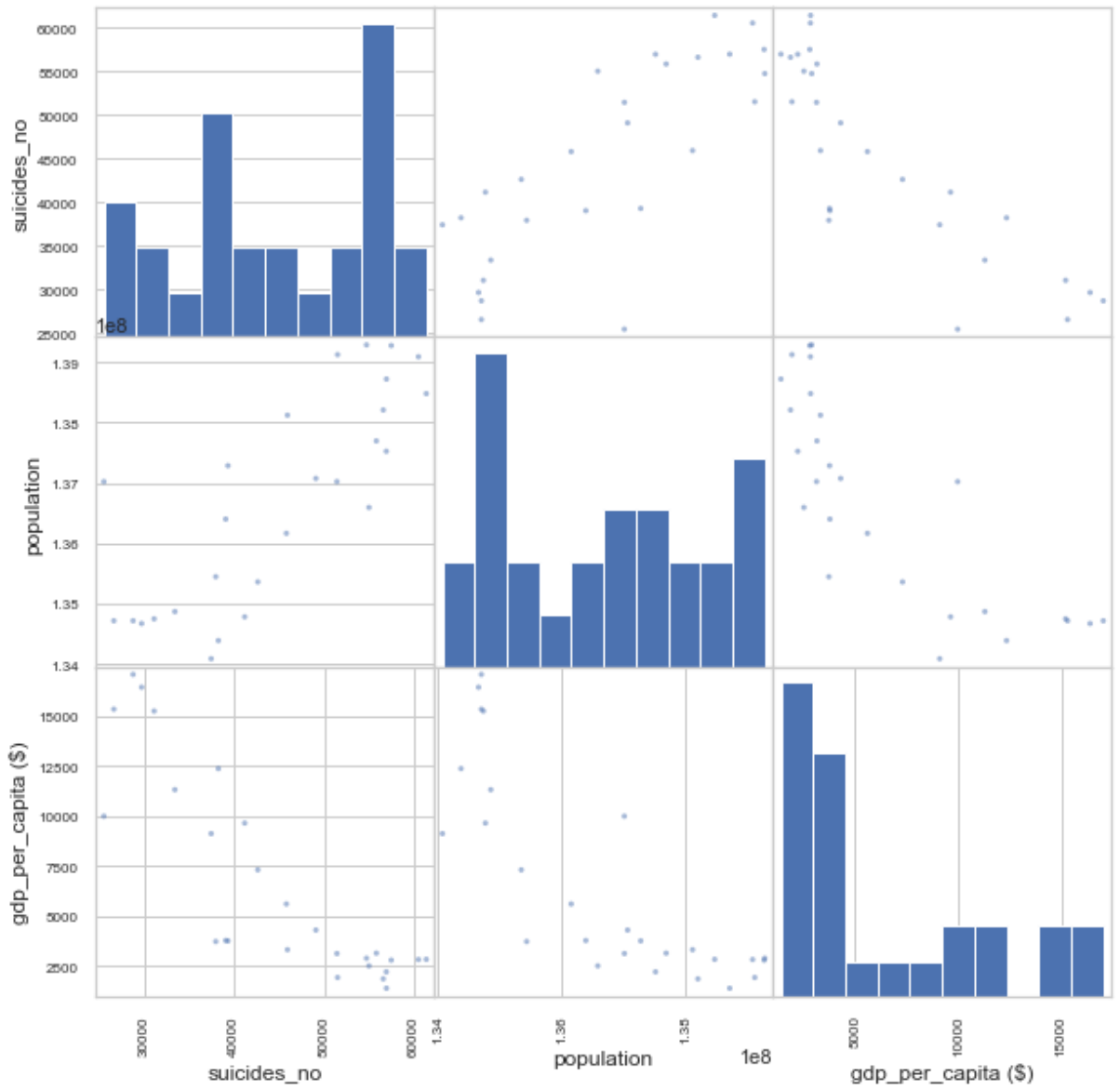


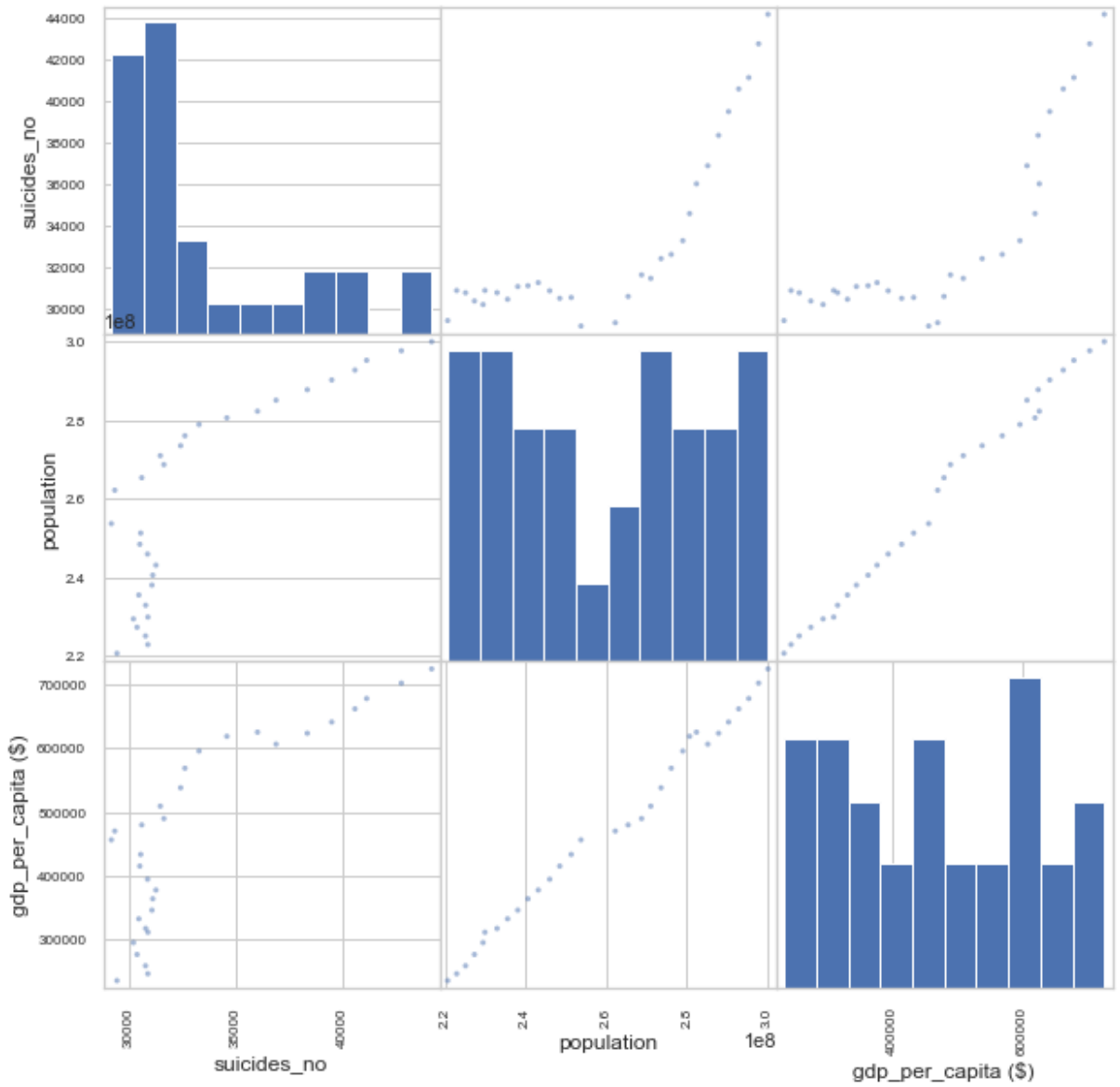
Диаграмма рассеяния по России очень интересен, отрицательная корреляция между

ростом населения и ВНП на душу населения, нет корреляции между суицидами и ВНП

США

In [196]:
```
data_new_usa =  df[df.country=='United States'].groupby(by='year').sum
().sort_values(by='population', ascending=False)
datacorr_new_usa = data_new_usa.drop(['HDI for year', 'suicides/100k po
p'], axis=1)


scatter_matrix(datacorr_new_usa, figsize=(10, 10))
scatter_matrix
```

Out[196]:
```
<function pandas.plotting._misc.scatter_matrix(frame, alpha=0.5, figsiz
e=None, ax=None, grid=False, diagonal='hist', marker='.', density_kwds=
None, hist_kwds=None, range_padding=0.05, **kwargs)>
```



США показывает максимальную связь между ВНП и населением

#

Анализ доли суицидов по гендерным группам

Сводные данные (Мексика, Россия и США)

In [139]:
```
def find_sex_suicides(df, country):
    sex_list = df[(df['country'] == country)]
    summ_suicides = sex_list.groupby('sex')['suicides_no'].sum()
    summ_pop = sex_list.groupby('sex')['population'].sum()
    return summ_suicides, summ_pop
```

In [140]:
```
find_sex_suicides(df, 'Mexico')
```

Out[140]:
```
(sex
 female    19334
 male      91805
 Name: suicides_no, dtype: int64,
 sex
 female    1397570687
 male      1375388472
 Name: population, dtype: int64)
```

In [141]:
```
find_sex_suicides(df, 'Russian Federation')
```

Out[141]:
```
(sex
 female    214330
 male      995412
 Name: suicides_no, dtype: int64,
 sex
 female    1980710973
 male      1710091647
 Name: population, dtype: int64)
```

In [142]:
```
find_sex_suicides(df, 'United States')
```

Out[142]:
```
(sex
 female    213797
 male      820216
 Name: suicides_no, dtype: int64,
 sex
 female    4113698286
 male      3940328915
 Name: population, dtype: int64)
```

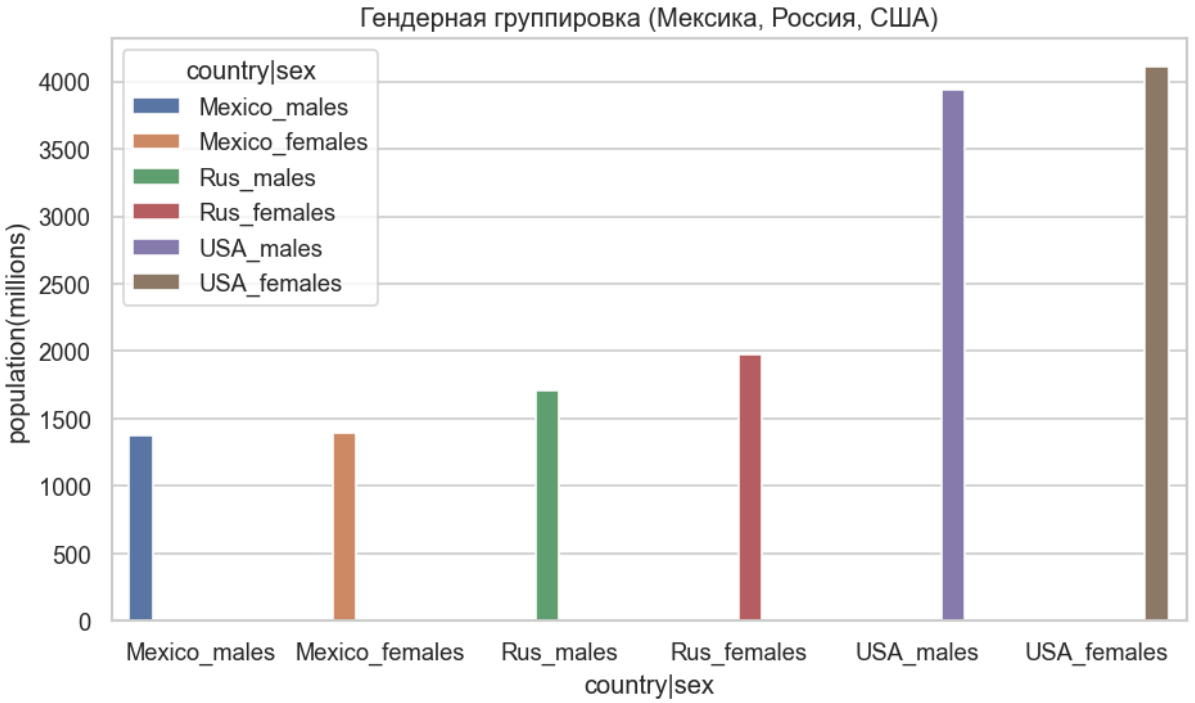
In [143]:
```
data = [['Mexico_males', 1375388472,  91805], ['Mexico_females', 139757
0687, 19334], ['Rus_males', 1710091647, 995412], ['Rus_females', 198071
0973, 214330], ['USA_males', 3940328915, 820216], ['USA_females', 41136
98286, 213797]]
columns = ['country|sex', 'population(millions)', 'suicides_K']
table = pd.DataFrame(data=data, columns=columns)
table['population(millions)'] = table['population(millions)'] // 100000
0
table['suicides_K'] = table['suicides_K'] / 1000
table['percentage_of_gender'] = round(table['suicides_K'] / table['popu
lation(millions)'] * 10, 3)

table
```

Out[143]:

	country|sex	population(millions)	suicides_K	percentage_of_gender
0	Mexico_males	1375	91.81	0.67
1	Mexico_females	1397	19.33	0.14
2	Rus_males	1710	995.41	5.82
3	Rus_females	1980	214.33	1.08
4	USA_males	3940	820.22	2.08
5	USA_females	4113	213.80	0.52

In [144]:
```
plt.figure(figsize=(9,5), dpi=120)
```

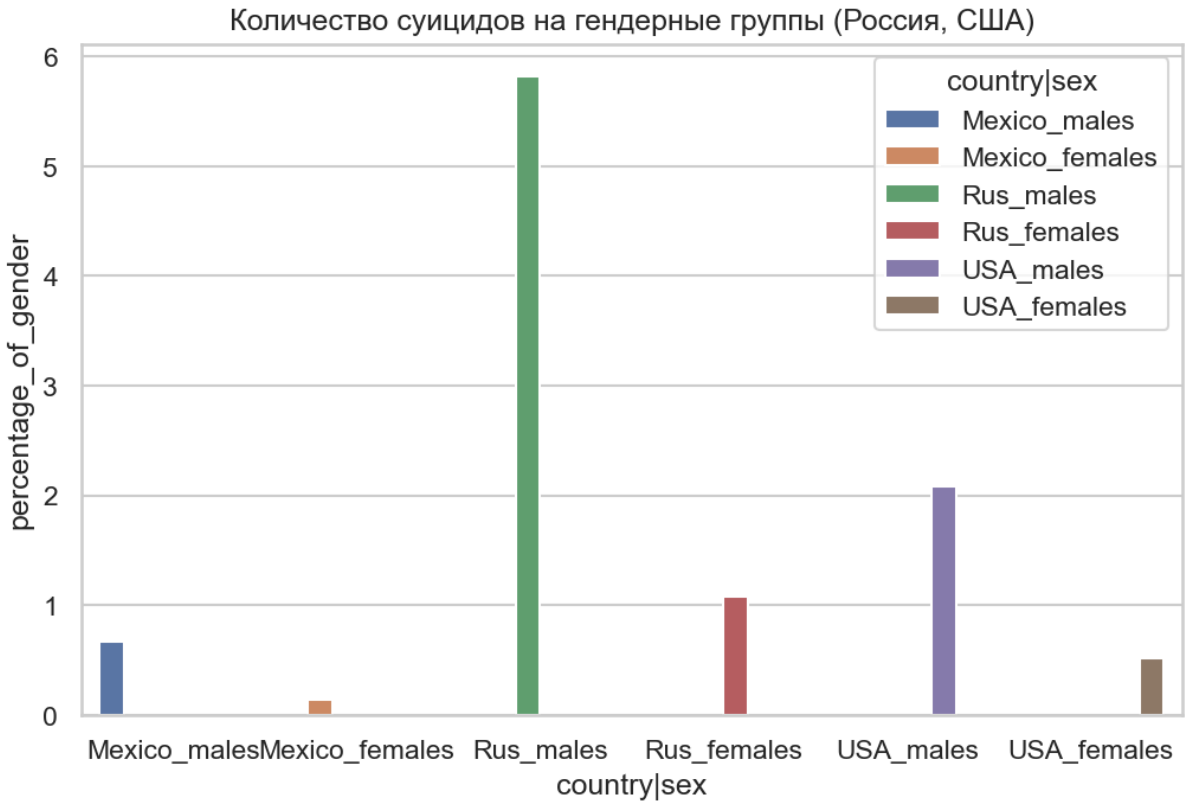
```
ax = sns.barplot(x="country|sex", y="population(millions)", hue = 'coun
try|sex',
                 data=table).set_title('Гендерная группировка (Мексика,
 Россия, США)')
```



В России и США преобладает население мужского пола (~15%), в Мексике количество мужского и женского населения примерно равны

#

In [145]:
```
plt.figure(figsize=(8,5), dpi=150)
ax = sns.barplot
ax = sns.barplot(x="country|sex", y="percentage_of_gender", hue = 'coun
try|sex',
                 data=table).set_title('Количество суицидов на гендерны
е группы (Россия, США)')
```



Российские мужчины заканчивали жизнь самоубийством в 5 раз больше, чем женщины, в США - в 4 раза. Примечательно то, что число суицидов среди женщин в России и США находятся на одинаковом уровне, несмотря на существенную разницу в количестве. Латиноамериканские мужчины также совершают больше самоубийств, чем женщины

#

Суициды по поколениям (Мир)

In [146]:
```
pd.options.display.float_format = '{:,.2f}'.format
gen_world = df.groupby('generation')['population', 'suicides_no'].sum()
.sort_values(by='suicides_no', ascending=False)
gen_world['population'] = gen_world['population'] / 1000000
gen_world['suicides_no'] = gen_world['suicides_no'] / 1000000
gen_world.set_axis(['population (billions)','suicides (millons)'],axis
= 'columns',inplace = True)
gen_world
```

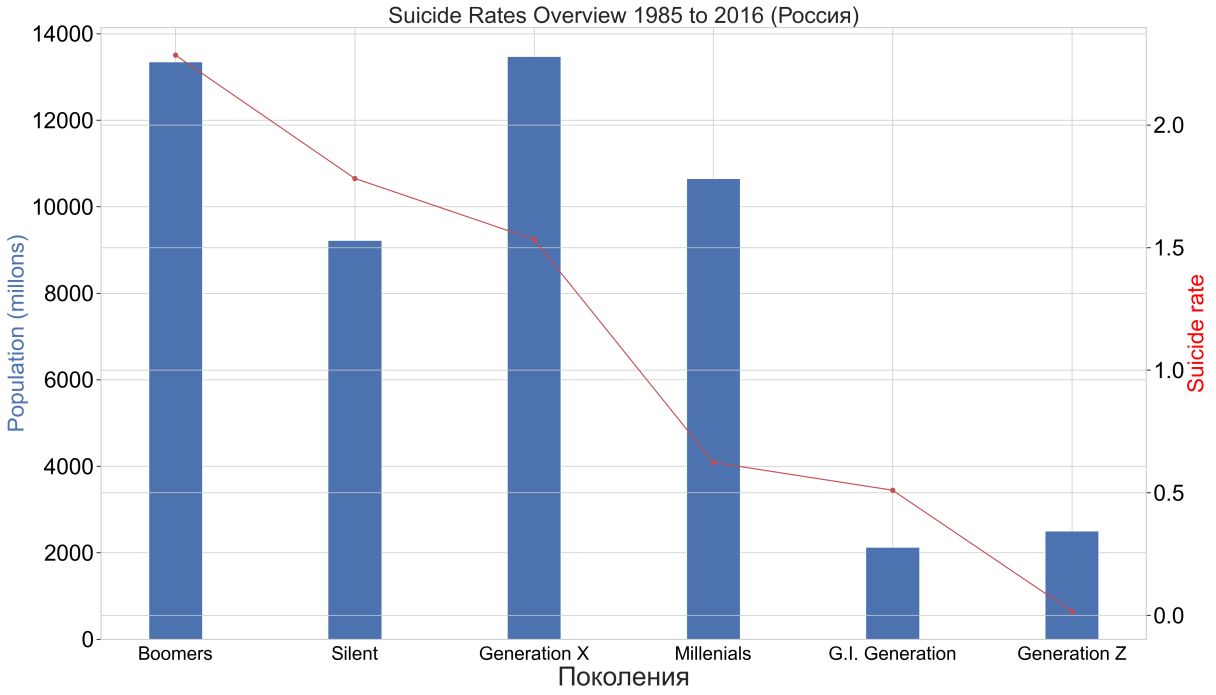
```
<ipython-input-146-8da2198c47b8>:2: FutureWarning: Indexing with multip
le keys (implicitly converted to a tuple of keys) will be deprecated, u
se a list instead.
  gen_world = df.groupby('generation')['population', 'suicides_no'].sum
().sort_values(by='suicides_no', ascending=False)
```

Out[146]:

generation	population (billions)	suicides (millons)
Boomers	13,350.51	2.28
Silent	9,220.33	1.78
Generation X	13,472.11	1.53
Millenials	10,649.46	0.62
G.I. Generation	2,126.20	0.51
Generation Z	2,503.54	0.02

In [155]:
```
gen_labels = gen_world.index[:6]
gen_values = gen_world['population (billions)']
gen_suicides = gen_world['suicides (millons)']


fig, ax1 = plt.subplots(figsize=(25, 15), dpi=200)
plt.title('', fontsize=30)


ax1.set_xlabel('Поколения',fontsize=35 )
ax1.set_ylabel('Population (millons)', fontsize=30, color='b')
```

```
ax1.bar(gen_labels, gen_values, width=0.3, color='b')

ax1.tick_params(axis='y',labelsize=30, labelcolor='black')
ax1.tick_params(axis='x',labelsize=25, labelcolor='black')
ax2 = ax1.twinx()

ax2.set_ylabel('Suicide rate', fontsize=30, color='red')

ax2.plot(gen_labels, gen_suicides, color='r', marker='o')
ax2.tick_params(axis='y', labelsize=30, labelcolor='black')


plt.show()
```



Самая высокая доля суицидов приходится на поколения Boomers, Silent и G.I. Generation, поколение Миллениалов и Поколение Z меньшей степени склонны к суицидам

#

Выводы

Количество суицидов в России остается на очень высоком уровне, разрыв от США - существенный, пик суицидов приходится на 1994 год. Нет зависимостей между ростом населения и ростом национальных благ, а также между суицидами и ВНП. В основном, высокое количество суицидов приходится на возраст 75+, а в России на группу 35-54, возрастная группа 75+ на втором вместе. Во всех трёх странах преобладает население мужского пола, на котрое приходится основная доля суицидов. С течением времени, Российская Федерация показывает отрицательную динамику по уровню суицидов