In [1]:
```
import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_style('darkgrid')
```

In [3]:
```
df = pd.read_csv("master.csv")
df.head()
```

Out[3]:

	country	year	sex	age	suicides_no	population	suicides/100k pop	country-year	HDI for year	gdp_for_year ($)	gdp_p
0	Albania	1987	male	15-24 years	21	312900	6.71	Albania1987	NaN	2,156,624,900	
1	Albania	1987	male	35-54 years	16	308000	5.19	Albania1987	NaN	2,156,624,900	
2	Albania	1987	female	15-24 years	14	289700	4.83	Albania1987	NaN	2,156,624,900	
3	Albania	1987	male	75+ years	1	21800	4.59	Albania1987	NaN	2,156,624,900	
4	Albania	1987	male	25-34 years	9	274300	3.28	Albania1987	NaN	2,156,624,900	

In [4]:
```
df.describe()
```

Out[4]:

	year	suicides_no	population	suicides/100k pop	HDI for year	gdp_per_capita ($)
count	27820.000000	27820.000000	2.782000e+04	27820.000000	8364.000000	27820.000000
mean	2001.258375	242.574407	1.844794e+06	12.816097	0.776601	16866.464414
std	8.469055	902.047917	3.911779e+06	18.961511	0.093367	18887.576472
min	1985.000000	0.000000	2.780000e+02	0.000000	0.483000	251.000000
25%	1995.000000	3.000000	9.749850e+04	0.920000	0.713000	3447.000000
50%	2002.000000	25.000000	4.301500e+05	5.990000	0.779000	9372.000000
75%	2008.000000	131.000000	1.486143e+06	16.620000	0.855000	24874.000000
max	2016.000000	22338.000000	4.380521e+07	224.970000	0.944000	126352.000000

In [5]:
```
df.columns
```

Out[5]:
```
Index(['country', 'year', 'sex', 'age', 'suicides_no', 'population',
       'suicides/100k pop', 'country-year', 'HDI for year',
       ' gdp_for_year ($) ', 'gdp_per_capita ($)', 'generation'],
      dtype='object')
```

In [6]:
```
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        8364 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: 2.5+ MB
```

In [7]:
```
def missing_check(df):
    total = df.isnull().sum().sort_values(ascending=False)   # total number of null valu
    percent = (df.isnull().sum()/df.isnull().count()).sort_values(ascending=False)  # pe
    missing_data = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])  # put
    return missing_data # return the dataframe
missing_check(df)
```

Out[7]:

	Total	Percent
HDI for year	19456	0.699353
country	0	0.000000
year	0	0.000000
sex	0	0.000000
age	0	0.000000
suicides_no	0	0.000000
population	0	0.000000
suicides/100k pop	0	0.000000
country-year	0	0.000000
gdp_for_year ($)	0	0.000000
gdp_per_capita ($)	0	0.000000
generation	0	0.000000

In [8]:
```
df[['suicides_no','population','suicides/100k pop','gdp_per_capita ($)']].describe() #de
```

Out[8]:

	suicides_no	population	suicides/100k pop	gdp_per_capita ($)
count	27820.000000	2.782000e+04	27820.000000	27820.000000
mean	242.574407	1.844794e+06	12.816097	16866.464414
std	902.047917	3.911779e+06	18.961511	18887.576472
min	0.000000	2.780000e+02	0.000000	251.000000
25%	3.000000	9.749850e+04	0.920000	3447.000000
50%	25.000000	4.301500e+05	5.990000	9372.000000
75%	131.000000	1.486143e+06	16.620000	24874.000000
max	22338.000000	4.380521e+07	224.970000	126352.000000

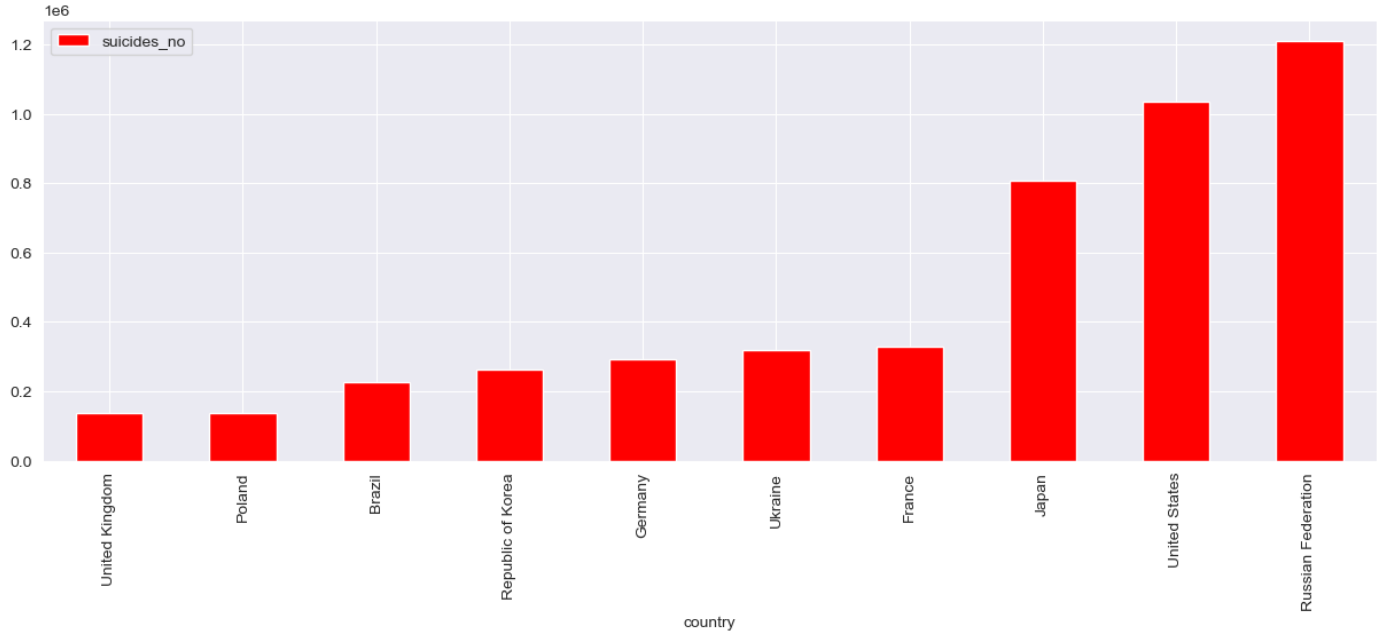
In [9]:
```
#frequency table for age
my_tab = pd.crosstab(index=df["age"],  # Make a crosstab
                     columns="count")                  # Name the count column
my_tab
```

Out[9]:

col_0	count
age	
15-24 years	4642
25-34 years	4642
35-54 years	4642
5-14 years	4610
55-74 years	4642
75+ years	4642

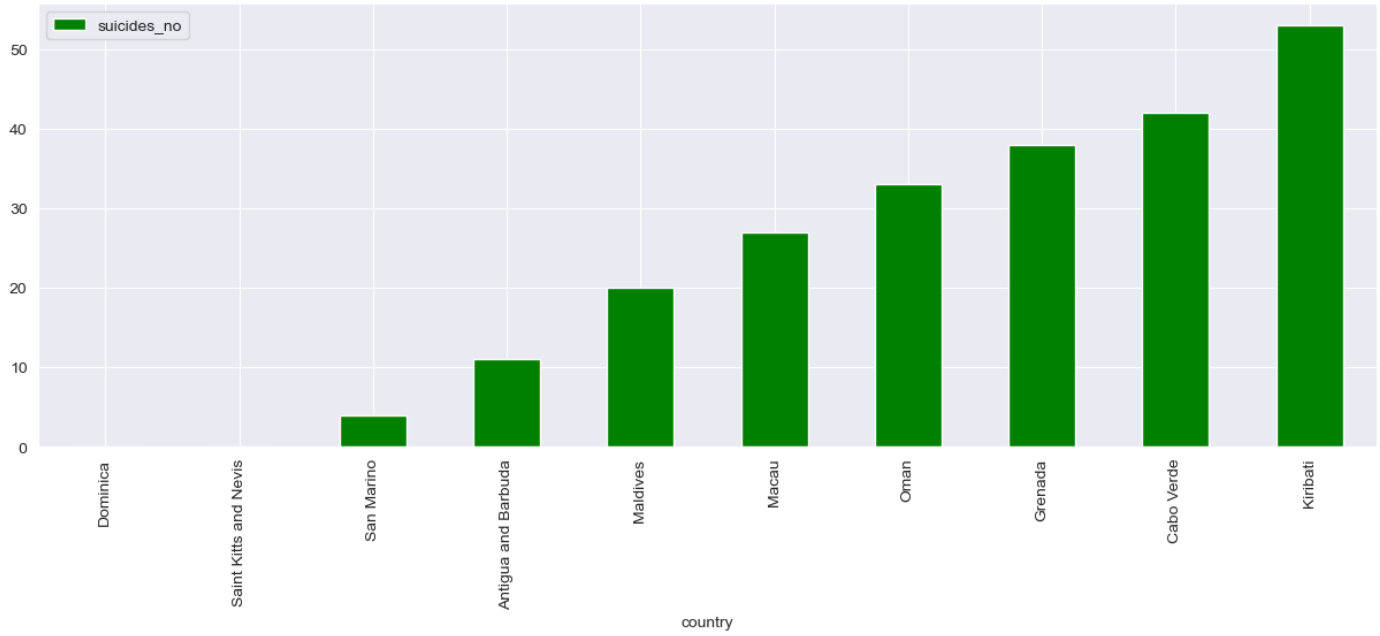
number of suicides in top countries

In [17]:
```
#group number of suicides by top 10 countries and plot it on a bar graph
df.groupby(
    by=['country'])['suicides_no'].sum().reset_index().sort_values(['suicides_no']).tail



plt.show()
```

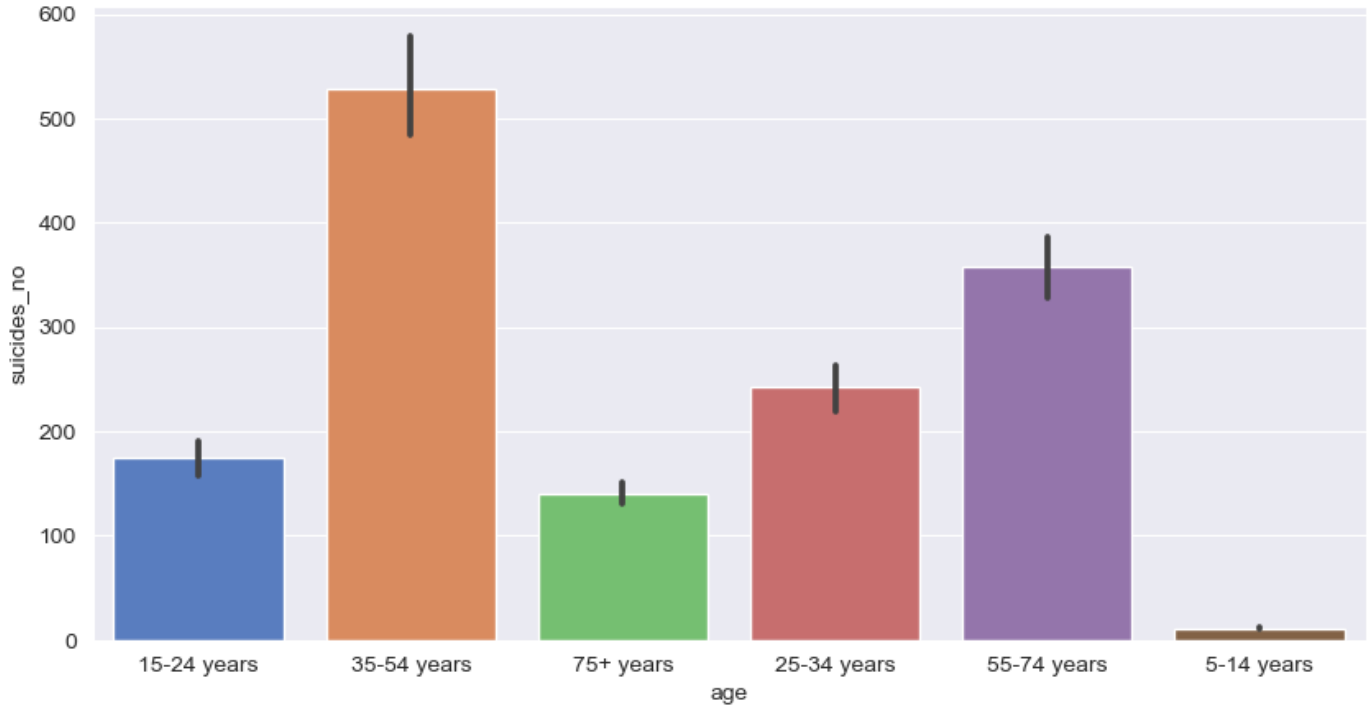


the top 3 countries with the highest suicides are Russia, USA and Japan

In [16]:
```
#group number of suicides by buttom 10 countries and plot it on a bar graph
df.groupby(
    by=['country'])['suicides_no'].sum().reset_index().sort_values(['suicides_no'], asce



plt.show()
```

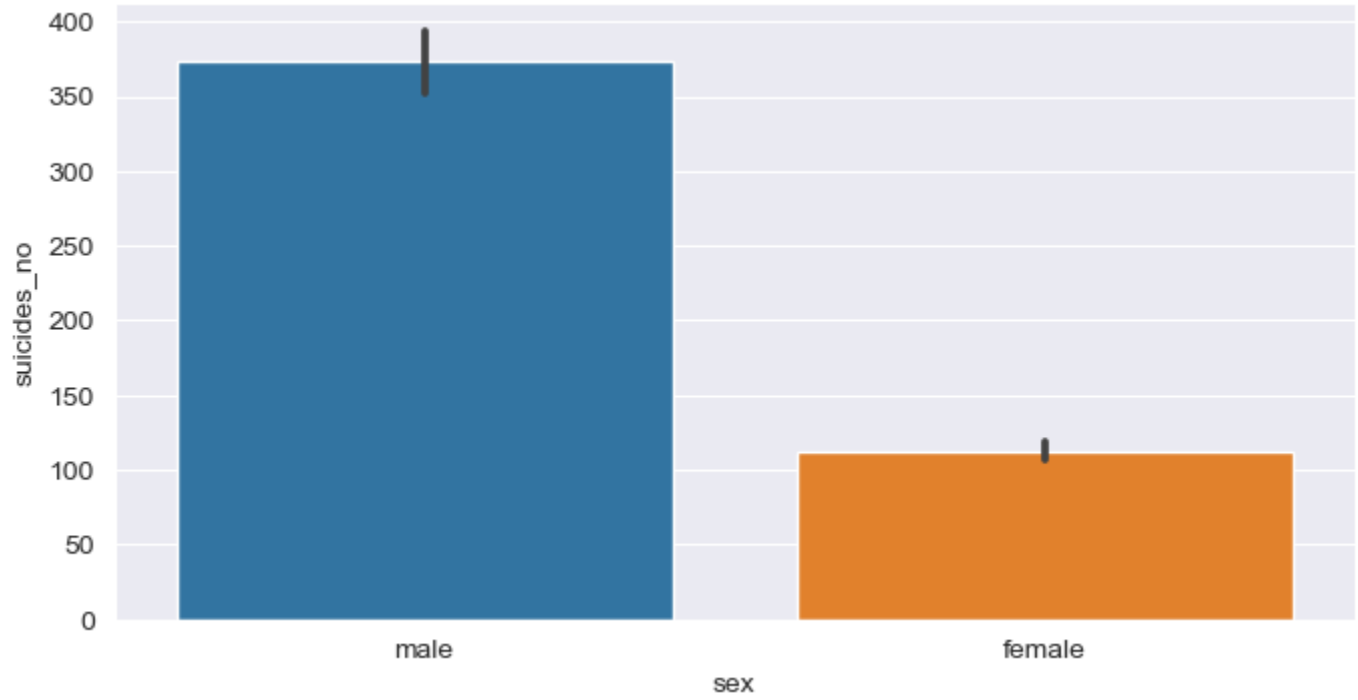


In [18]:
```
#### number of suicides vs age
plt.figure(figsize=(10,5)) # setting the figure size

ax = sns.barplot(x='age', y='suicides_no', data=df, palette='muted')  # barplot
```



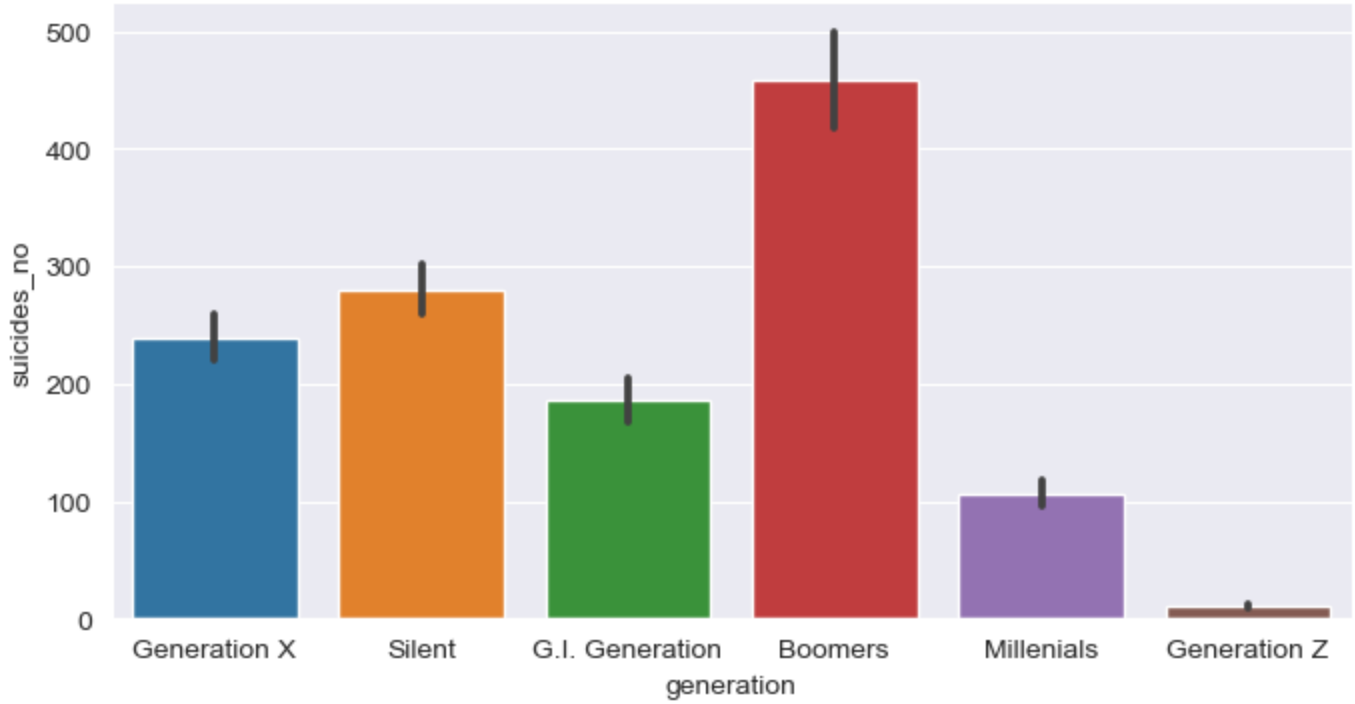
we find more people in the 35-54 years committing suicide followed by 55-74 years

In [19]:
```
#### number of suicides vs sex
plt.figure(figsize=(8,4)) # setting the figure size

ax = sns.barplot(x='sex', y='suicides_no', data=df)  # barplot
```



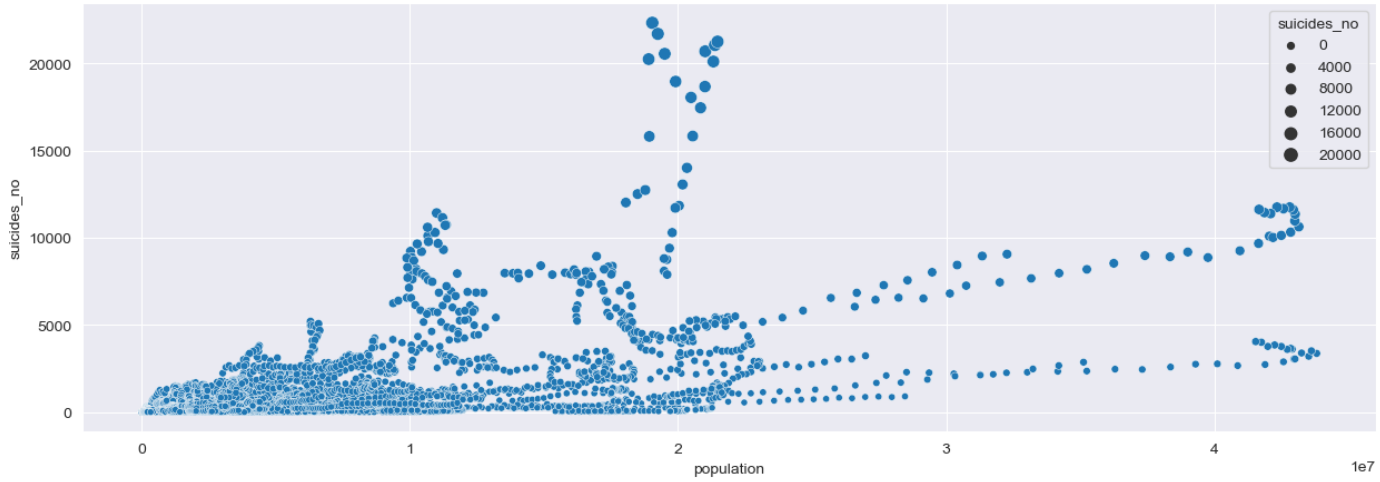
we find more males committing suicide compared to females

In [20]:
```
#### number of suicides vs generation
plt.figure(figsize=(8,4)) # setting the figure size

ax = sns.barplot(x='generation', y='suicides_no', data=df)  # barplot
```

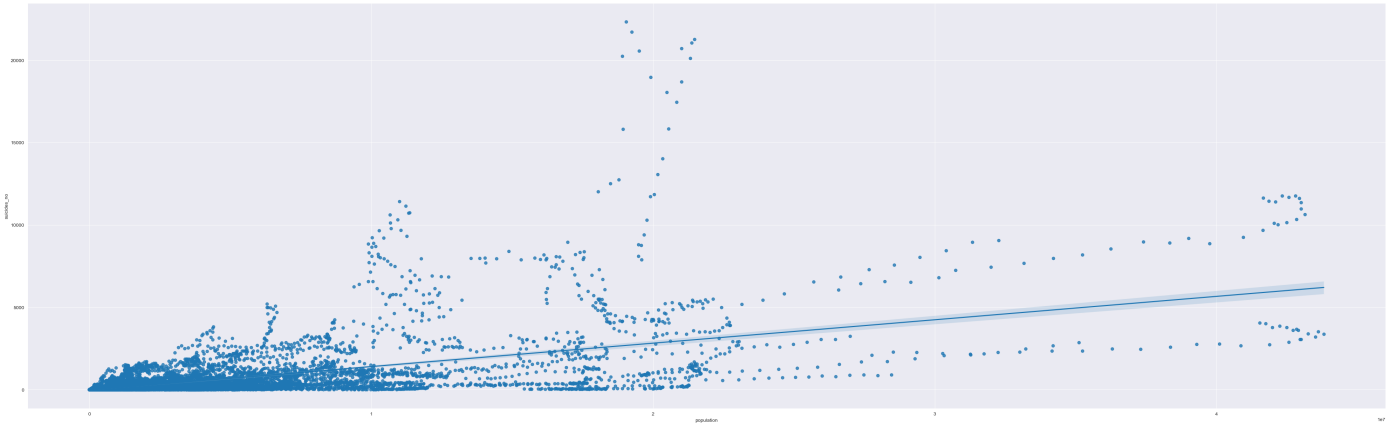


the generation with the most suicide are the Boomers

In [21]:
```
## number of suicide vs population
figure = plt.figure(figsize=(15,5))

ax = sns.scatterplot(x=df['population'],y='suicides_no', data=df, size = "suicides_no")
```

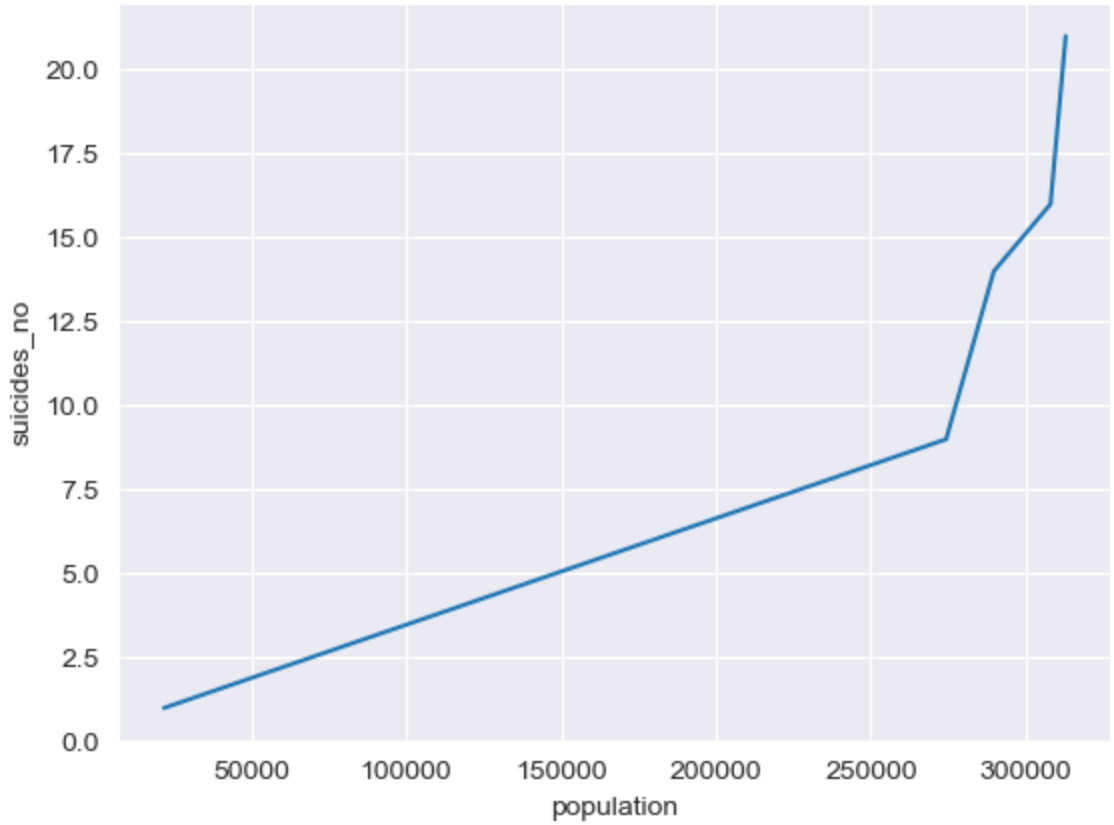


In [23]:
```
 # regression plot - scatter plot with a regression line
figure = plt.figure(figsize=(50,15))

ax = sns.regplot(x='population',y='suicides_no', data=df ) # regression plot - scatter p
```



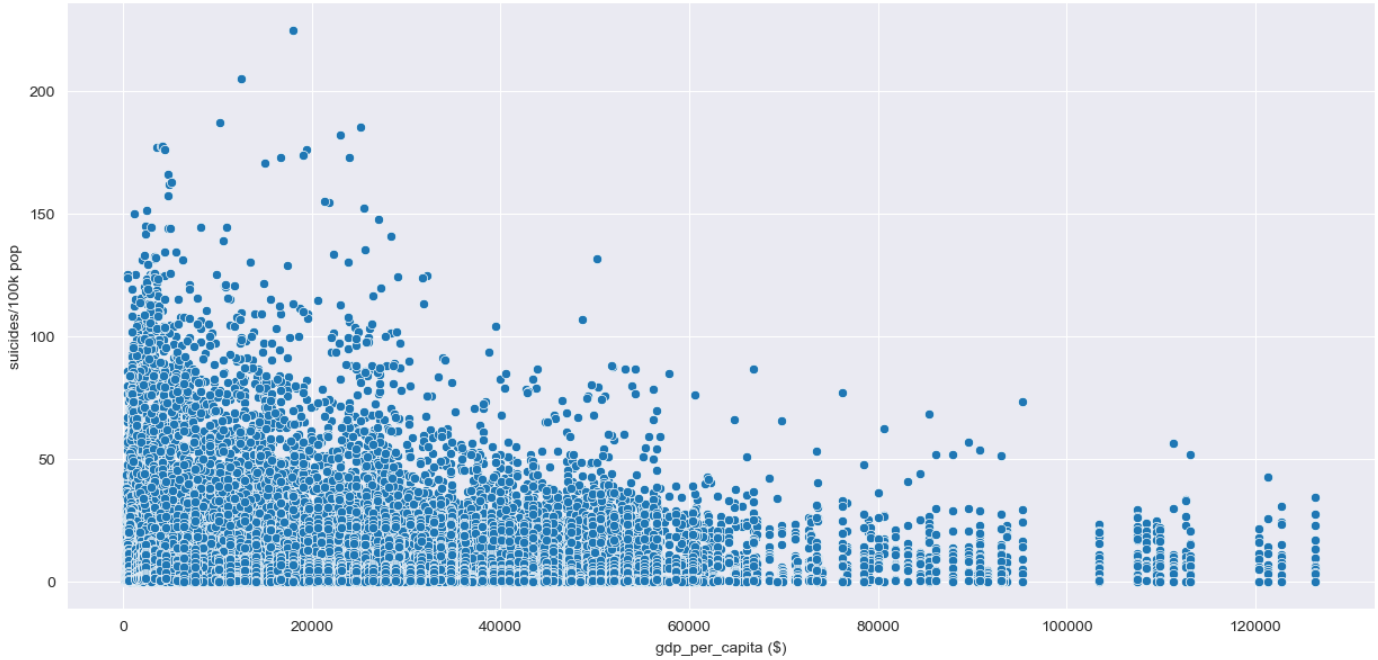
In [28]:
```
#Here we plotting a line plot.
sns.lineplot(x='population',y='suicides_no', data=df.head(),)
```

Out[28]:
```
<AxesSubplot:xlabel='population', ylabel='suicides_no'>
```



Scatter plot for Number of Suicides/100k Population Vs GDP Per Capita

In [29]:
```
figure = plt.figure(figsize=(15,7))

sns.scatterplot(x='gdp_per_capita ($)', y='suicides/100k pop', data=df) # scatter plot
plt.show()
```

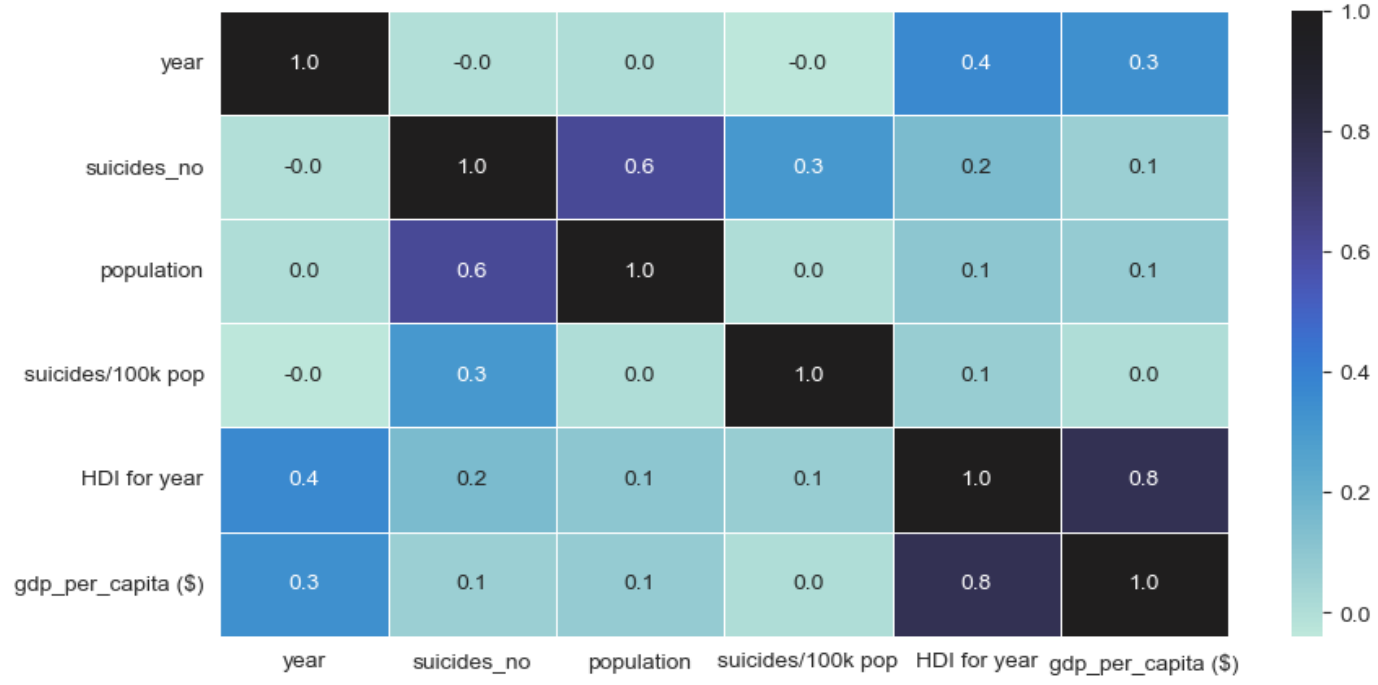


Looks like higher suicide rates are a bit more prevalent in countries with lower GDP Per Capita.

However, it doesn't look like there is any significant correlation between the two.

Checking the correlation among pairs of continuous variables

In [30]:
```
plt.figure(figsize=(10,5))
sns.heatmap(df.corr(), annot=True, linewidths=.5, fmt= '.1f', center = 1 )  # heatmap
plt.show()
```



The darker the color the higher the correlation.

None of the attributes seem to have a correlation of real significance.

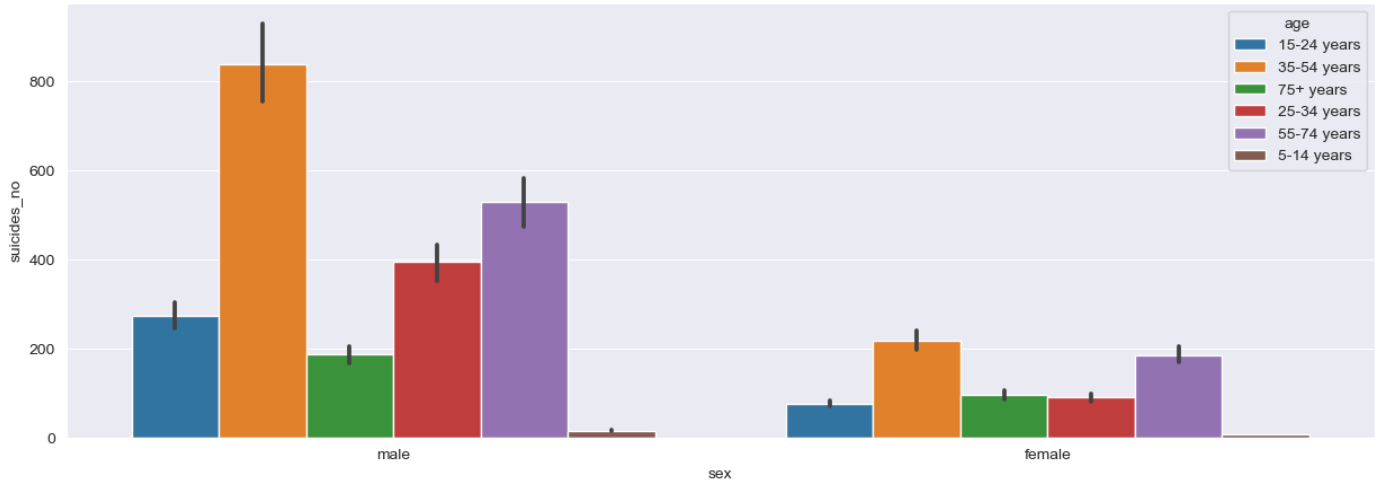
Some obvious correlations are that in a larger population, it is very likely that the number of suicides will be more in number.

Human Development Index - gdp per capita is the only pair with the high correlation.

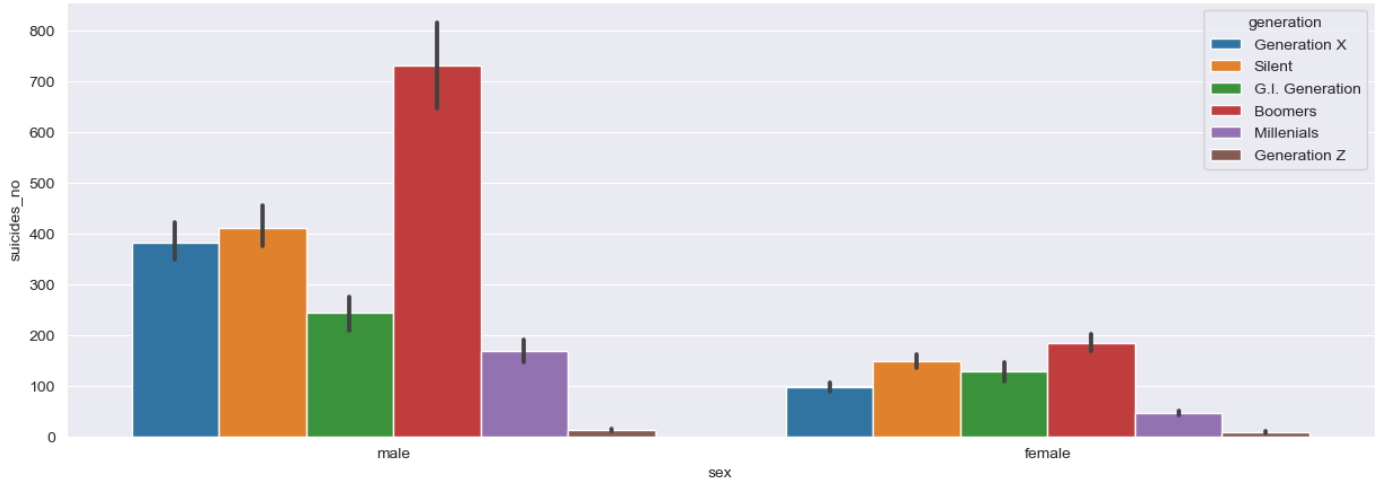
s not soo high correlation

Bar plot to check Number of suicides by sex and age (three variables used to generate a single plot)

In [31]:
```
plt.figure(figsize=(15,5))
sns.barplot(data=df,x='sex',y='suicides_no',hue='age')
plt.show()
```



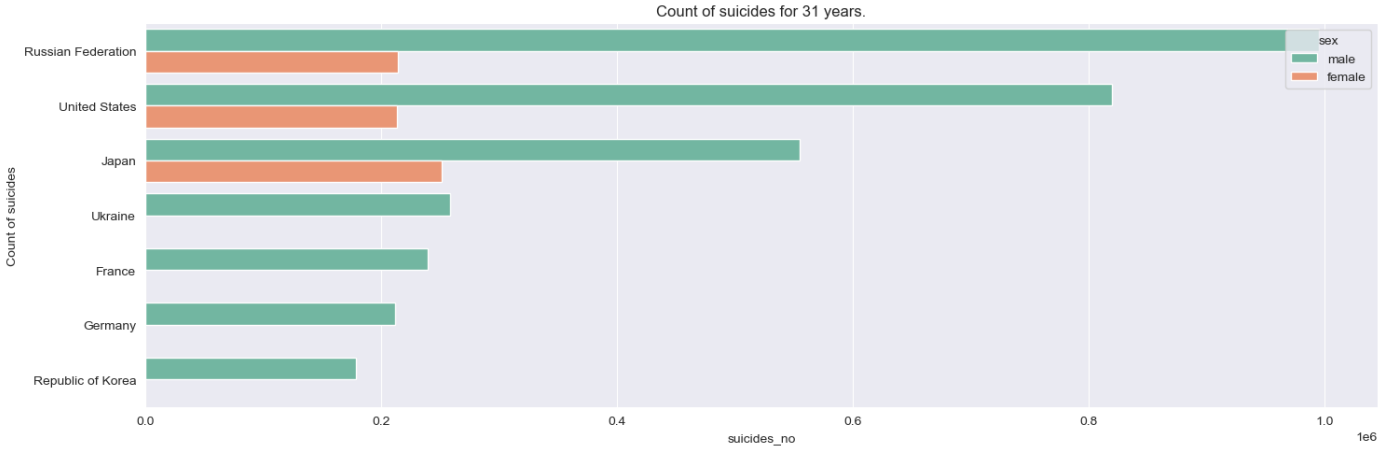
In [33]:
```
plt.figure(figsize=(15,5))
sns.barplot(data=df,x='sex',y='suicides_no',hue='generation')
plt.show()
```



Suicides are high among Boomers, both male and female.

No.of suicides: Country Vs Sex

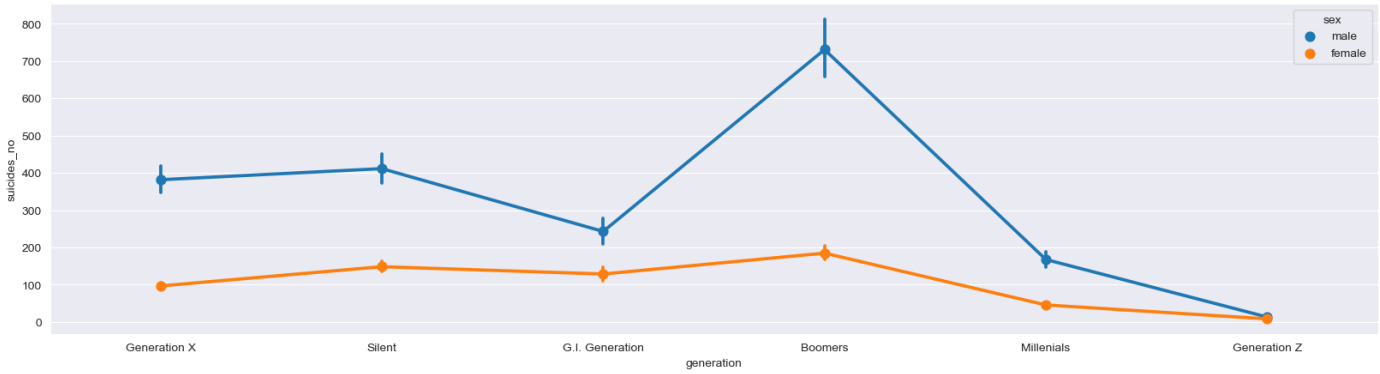
In [35]:
```
suic_sum_m = df['suicides_no'].groupby([df['country'],df['sex']]).sum()  # number of sui
suic_sum_m = suic_sum_m.reset_index().sort_values(by='suicides_no',ascending=False) # so
most_cont_m = suic_sum_m.head(10)  # getting the top ten countries in terms of suicides

fig = plt.figure(figsize=(15,5))
plt.title('Count of suicides for 31 years.')

sns.barplot(y='country',x='suicides_no',hue='sex',data=most_cont_m,palette='Set2');

plt.ylabel('Count of suicides')
plt.tight_layout()
```



*In comparison to other countries with high suicide rates, Japan has a larger proportion of female suicides.

Average number of suicides across each generation for a given gender along with the confidence intervals - Point Plot

In [37]:
```
plt.figure(figsize=(20,5))

sns.pointplot(x="generation", y="suicides_no", hue = 'sex',  data=df)
plt.show()
```

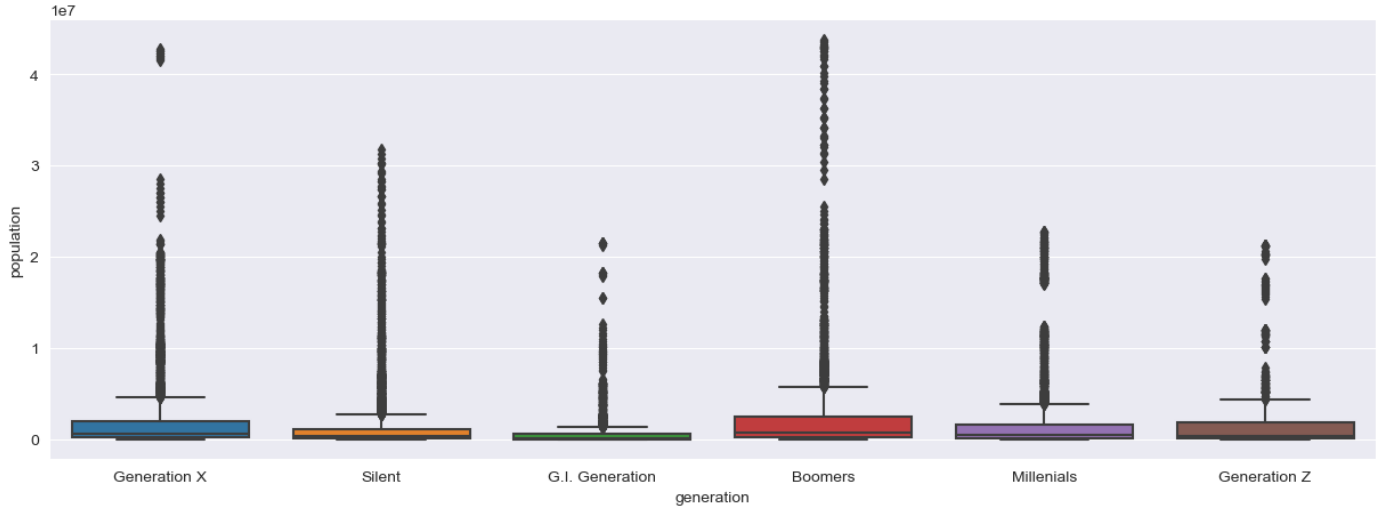


generation Z have an almost an equal number of suicides among males and females

suicide among men flatuates among the men but its fairly equal among the females

distribution of population across each generation

In [39]:
```
plt.figure(figsize=(15,5))
sns.boxplot(x=df.generation, y=df['population'])
plt.show()
```



from the boxplot distribution of population across generation is highly skewed with alot of outliers

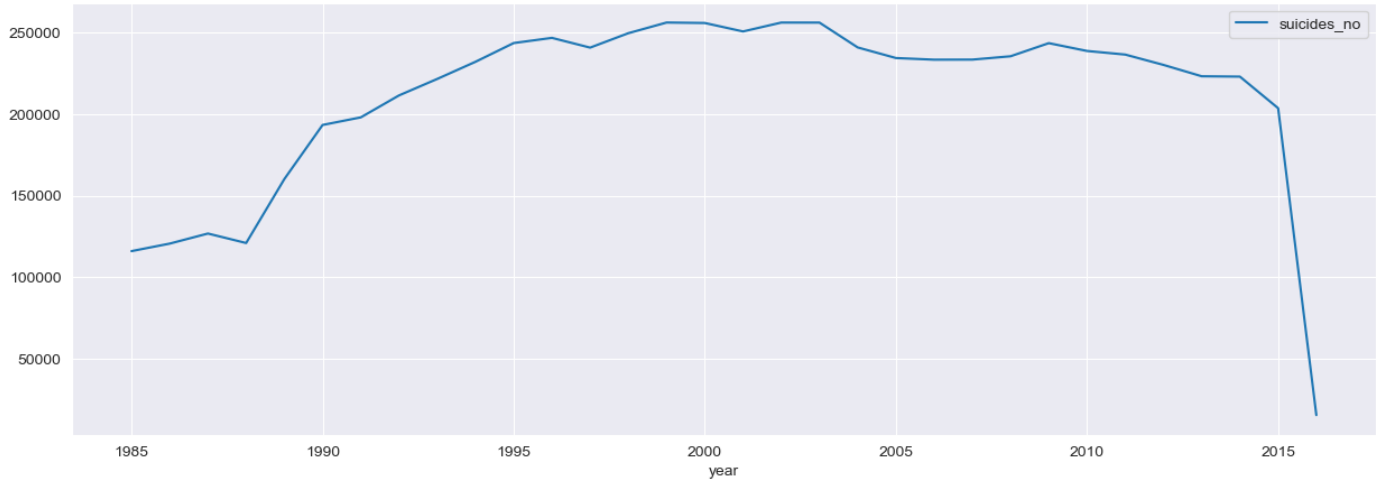
Trend of suicide accross the years

In [42]:
```
df.head()
```

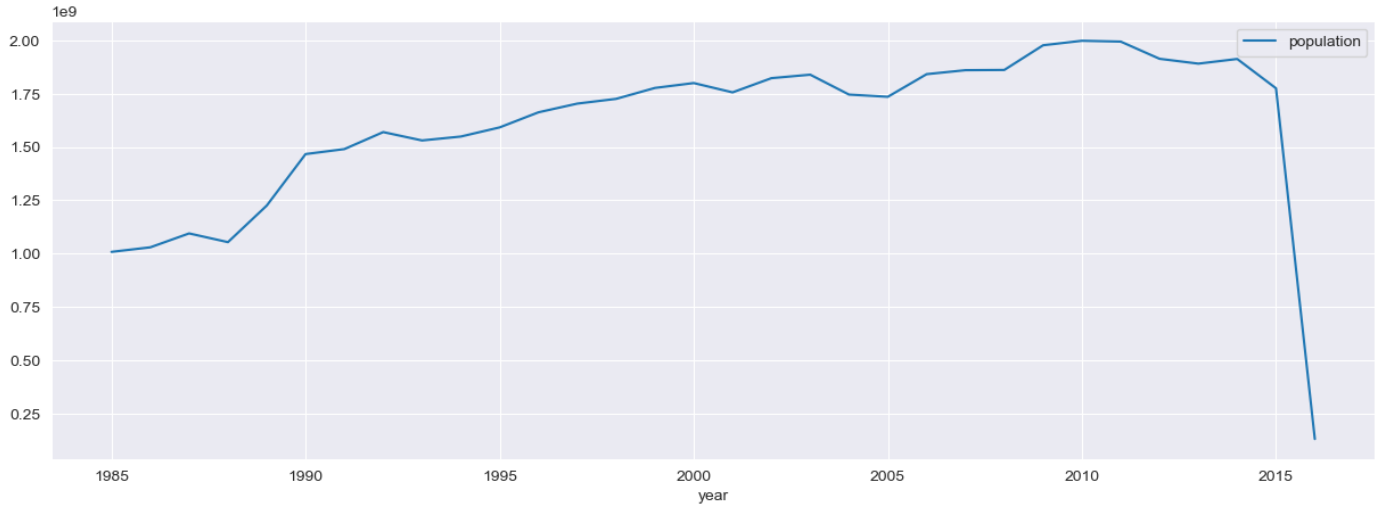
Out[42]:

	country	year	sex	age	suicides_no	population	suicides/100k pop	country-year	HDI for year	gdp_for_year ($)	gdp_p
0	Albania	1987	male	15-24 years	21	312900	6.71	Albania1987	NaN	2,156,624,900	
1	Albania	1987	male	35-54 years	16	308000	5.19	Albania1987	NaN	2,156,624,900	
2	Albania	1987	female	15-24 years	14	289700	4.83	Albania1987	NaN	2,156,624,900	
3	Albania	1987	male	75+ years	1	21800	4.59	Albania1987	NaN	2,156,624,900	
4	Albania	1987	male	25-34 years	9	274300	3.28	Albania1987	NaN	2,156,624,900	

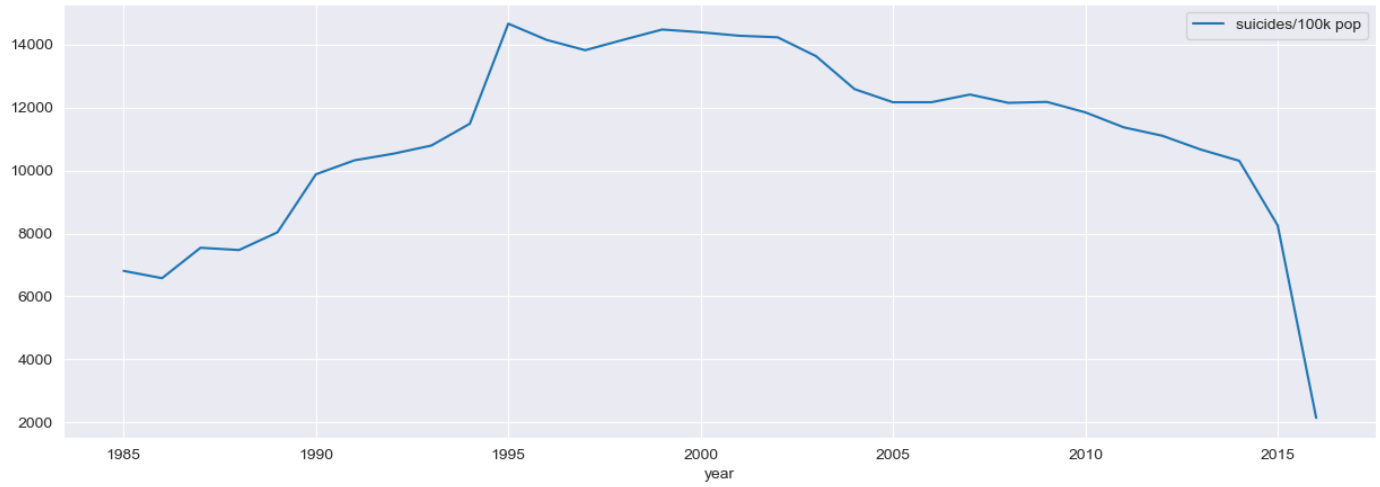
In [43]:
```
df[["year", "suicides_no"]].groupby(["year"]).sum().plot(figsize=(15,5))
plt.show()
```



In [44]:
```
df[["year", "population"]].groupby(["year"]).sum().plot(figsize=(15,5))
plt.show()
```



In [45]:
```
df[["year", "suicides/100k pop"]].groupby(["year"]).sum().plot(figsize=(15,5))
plt.show()
```



we find suicides/100k pop peaking in 1995

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: