Data Analysis of Factors affecting Suicide Rate

Jeffrey Ma

Abstract

The dataset selected was the "Suicide Rates Overview 1985-2016" from Kaggle. Multiple research questions were asked, and they dealt with which countries have the highest average rates of suicide, whether a relationship exists between suicide rates and age group or suicide rates and sex, and finally if there was a change in the rate of suicide since 1985. The method used was to load the dataset into python, clean the data, and run some procedures on the data to develop tables, charts, and plots describing general trends found within the data. What this exploratory analysis appears to have found was that Lithuania had the highest average rate of suicide, men were more likely than women to commit suicide, adults in the 35-54 range were the most likely to commit suicide, and suicide rates as a whole seem to have been relatively constant across age groups and sex since 1985.

Motivation

People committing suicide not only involves taking one's own life tragically, but also has significant impact on the lives of friends, colleagues, and especially close family members. It is a serious social issue that has significant impact on society. The main objective of this study is, through data mining, to find out factors affecting people who have committed suicide so that solutions may be developed to minimize it. We will also be looking for factors that could possibly be linked to individuals at risk of committing suicide.

Dataset(s)

The dataset used for this final project can be found on Kaggle. The title of the dataset is Suicide Rates Overview 1985-2016. Inside the dataset, we have 12 features in 12 columns. They are: "country, year, sex, age, suicides_no, population, suicides / 100k population, country-year, HDI for year, gdp_for_year ($), gdp_per_capita ($) , and generation". There are 27820 rows of data, and the features of interest are the total number of suicides for each row and number of suicides per 100k population for each row.

The url to access the data set: https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

Data Preparation and Cleaning

This dataset was clean. When checking for null values in the dataset, "HDI for year" column had 19456 null values, while the rest had none. Thus to make the dataframe easier to work with, "HDI for year" variable was dropped from the dataset.

Research Question(s)

How has the rate of suicide changed since 1985?

Which countries have the highest average rate of suicide?

Is there a relationship between suicide rates and age group?

Is there a relationship between suicide rates and sex?

Methods

First the data was loaded into a jupyter notebook. Checks were done to make sure there were no null values. Outliers were checked for. Then a table was created to add up all the cases of suicide per 100k population for each country. This was done to evaluate average suicides per 100k population from 1985 to 2016. Then a linear plot of the rate of change in suicide rate per 100k population vs sex was created. Then multiple line plots comparing different age groups with suicide rate were generated, but all the values were rescaled to be between 0 and 1 to fit within one plot and generate comparisons with other plots. Finally, seaborn was used to create different pie charts from the dataset to see which features were most likely predictors of suicide. In particular, sex and age were two features of interest.

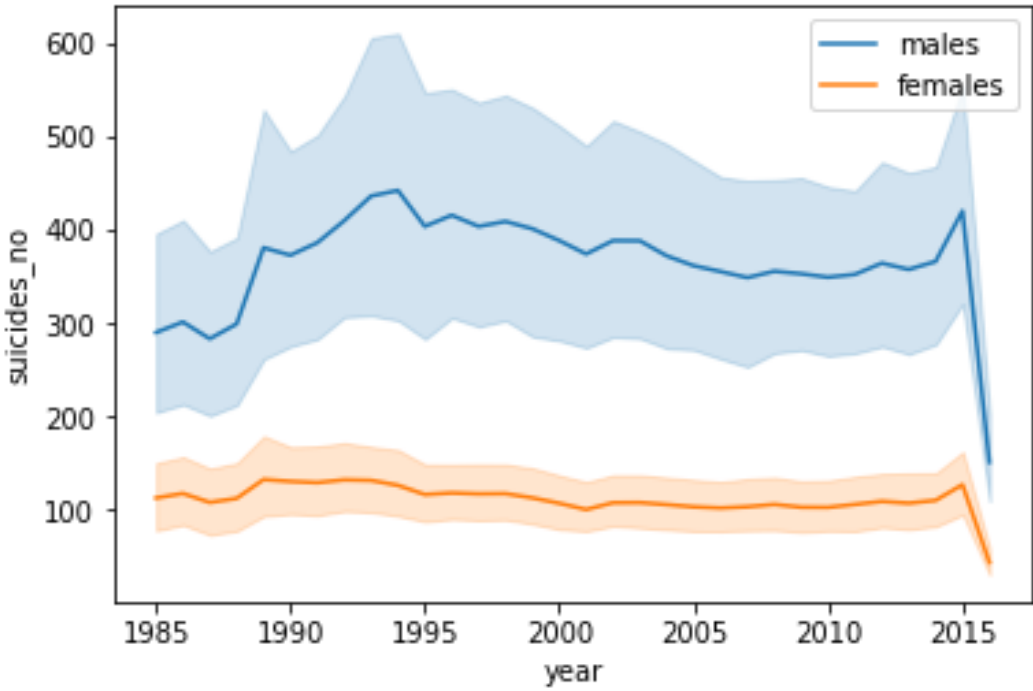
Findings-Country with most average suicides per 100k population: Lithuania

Top 10 countries with highest suicide averages per 100k population

	country	suicides/100k pop	population
0	Lithuania	40.415573	2.598672e+05
1	Sri Lanka	35.295152	1.382770e+06
2	Russian Federation	34.892377	1.139137e+07
3	Hungary	32.761516	8.020782e+05
4	Belarus	31.075913	7.832234e+05
5	Kazakhstan	30.511282	1.209980e+06
6	Latvia	29.259325	1.779867e+05
7	Slovenia	27.827857	1.597961e+05
8	Estonia	27.276905	1.075032e+05
9	Ukraine	26.582321	3.828777e+06

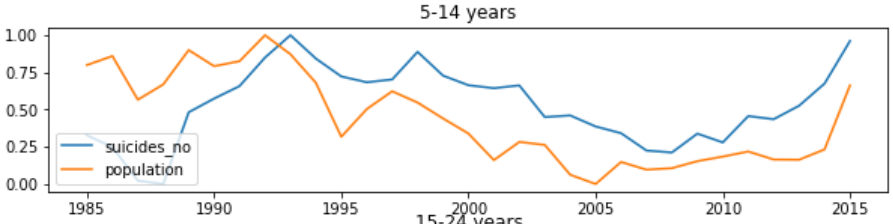
Findings

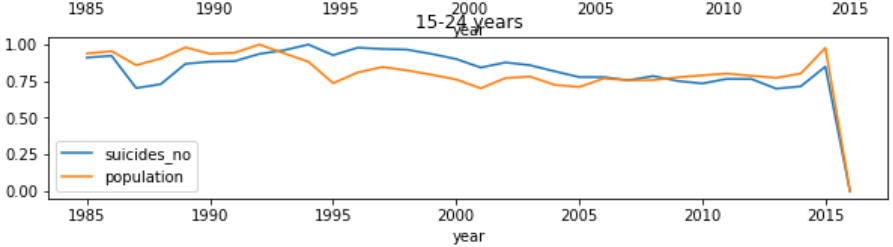
- In this graph, we can see the change in total suicides for each sex per year
- Visually it appears the rate is relatively constant for males and females, except for a precipitous drop in 2016
- 2016 could be an outlier year

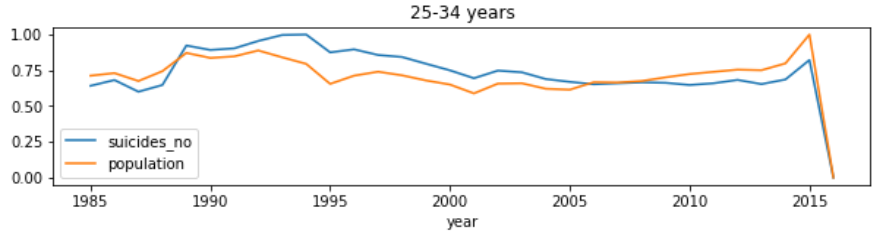


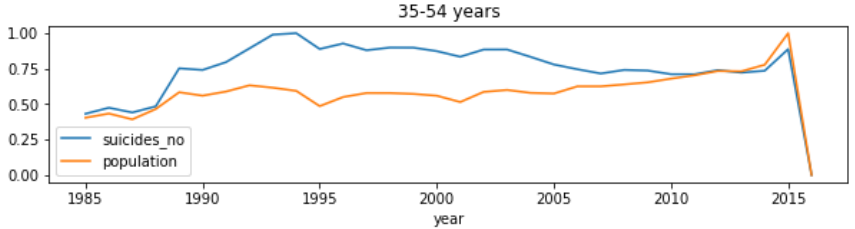
Rate of Suicide Change Based on Age

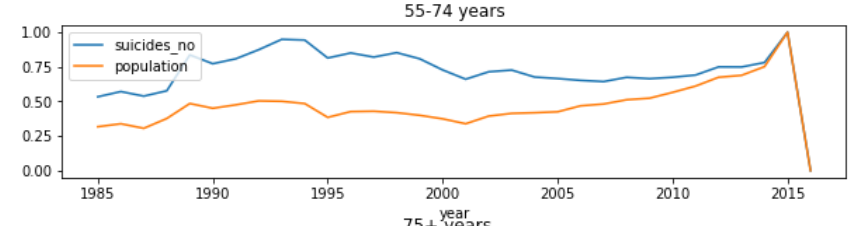
- It appears rates are relatively constant for age groups except for pre-teens.

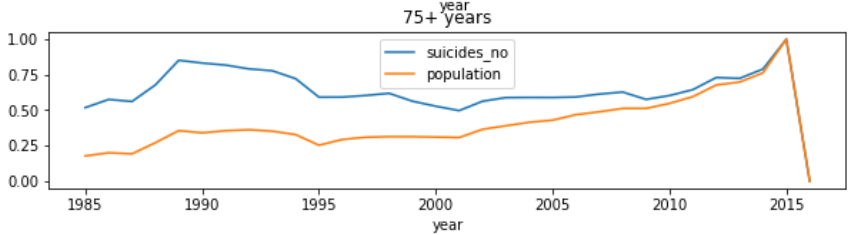






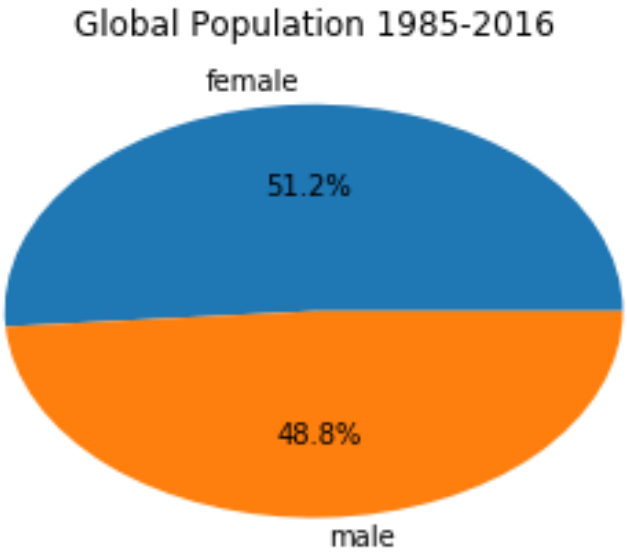


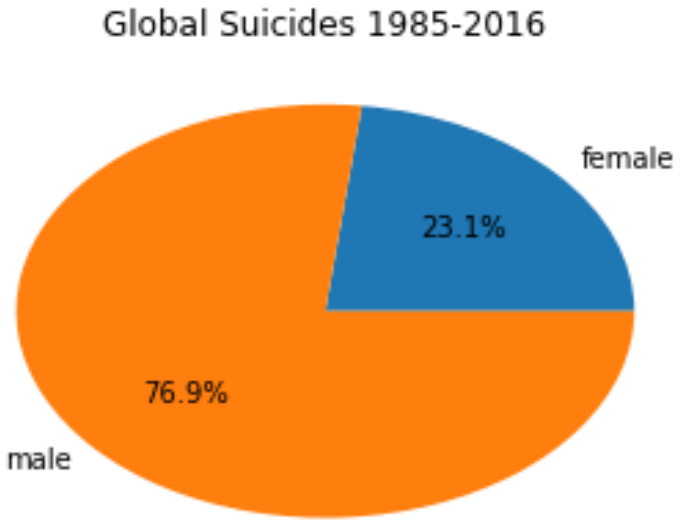




Findings

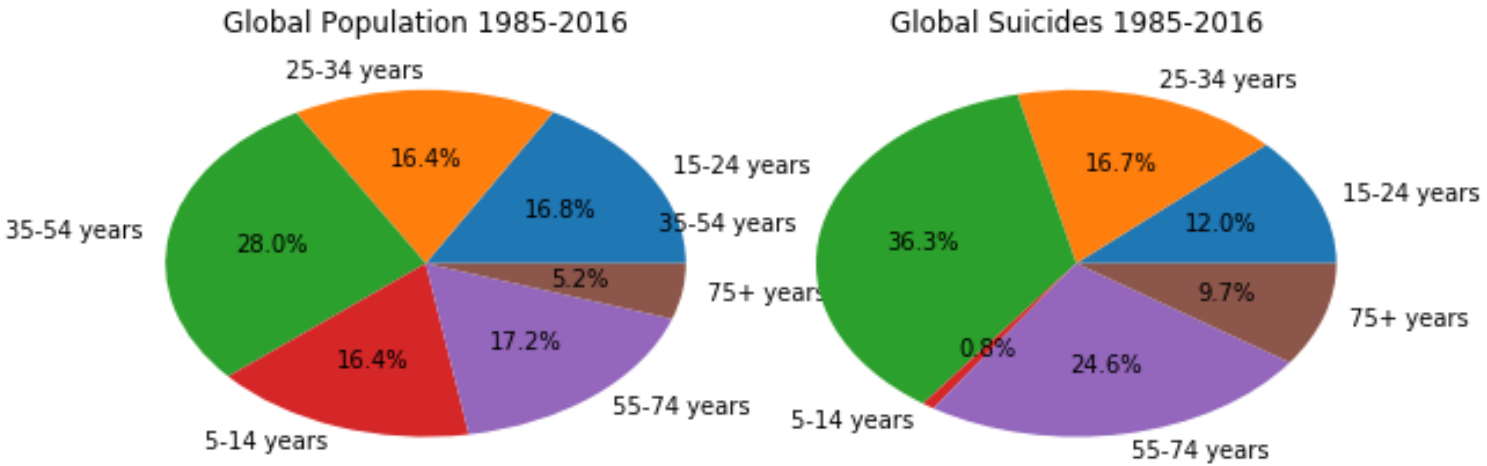
- Men are more likely than women to commit suicide





Findings

- Adults from ages 35-54 were the most likely to commit suicide



Limitations

The dataset is three years old now, so trends could have changed in relation to economic factors. Also, there appears to be an outlier year in 2016 where suicides across the board were down. If I were to re-analyze this dataset, I probably would have dropped all data from 2016 from the analysis.

All in all, this was an exploratory data analysis of an existing dataset. Further statistical analyses and proper statistical testing is required to make any conclusions based off of the data.

Conclusions

In conclusion, it was found from the data that men appeared more likely than women to commit suicide. It was also shown that adults ranging from ages 35-54 were found to be the most likely to commit suicide globally over any other age group. We also discovered that Lithuania had the highest amount of suicides per one hundred thousand population in our dataset. Finally, the rate of suicide appears to be steady since 1985 based off data visualization outside of a sharp drop in 2016 which could have been an outlier year, or improper data collection.

All in all, more statistical experimentation is required to make any hard conclusions on the relationships between different factors and suicide, but general trends in suicide rate found in the dataset were shown.

References

A number of exploratory data analysis studies were looked at for reference on this project. Due to the nature of the shared work that is data analysis, I do not know the names of the accounts that posted their exploratory analysis online. I can however link the corresponding URLs that helped me with the coding the most:

https://www.kaggle.com/canbugra/introduction-to-python-datai-team

https://www.kaggle.com/sway985/suicide-exploration/notebook

Acknowledgements

Data was taken from the Kaggle website, and other public informal analysis were studied to help with the coding portions of this final project. On the previous slide, these public informal analyses were listed with URL links.

Final Project Suicide Dataset

February 18, 2019

```
In [1]: import pandas as pd
        import seaborn as sns
        import numpy as np
        import matplotlib.pyplot as plt

In [22]: suicide = pd.read_csv("C:/Users/jeffr/Downloads/suicide-rates-overview-1985-to-2016/ma

In [3]: suicide.head

Out[3]: <bound method NDFrame.head of           country  year     sex          age  suicides_no
        0         Albania  1987    male  15-24 years           21      312900
        1         Albania  1987    male  35-54 years           16      308000
        2         Albania  1987  female  15-24 years           14      289700
        3         Albania  1987    male    75+ years            1       21800
        4         Albania  1987    male  25-34 years            9      274300
        5         Albania  1987  female    75+ years            1       35600
        6         Albania  1987  female  35-54 years            6      278800
        7         Albania  1987  female  25-34 years            4      257200
        8         Albania  1987    male  55-74 years            1      137500
        9         Albania  1987  female   5-14 years            0      311000
        10        Albania  1987  female  55-74 years            0      144600
        11        Albania  1987    male   5-14 years            0      338200
        12        Albania  1988  female    75+ years            2       36400
        13        Albania  1988    male  15-24 years           17      319200
        14        Albania  1988    male    75+ years            1       22300
        15        Albania  1988    male  35-54 years           14      314100
        16        Albania  1988    male  55-74 years            4      140200
        17        Albania  1988  female  15-24 years            8      295600
        18        Albania  1988  female  55-74 years            3      147500
        19        Albania  1988  female  25-34 years            5      262400
        20        Albania  1988    male  25-34 years            5      279900
        21        Albania  1988  female  35-54 years            4      284500
        22        Albania  1988  female   5-14 years            0      317200
        23        Albania  1988    male   5-14 years            0      345000
        24        Albania  1989    male    75+ years            2       22500
        25        Albania  1989    male  25-34 years           18      283600
        26        Albania  1989    male  35-54 years           15      318400
        27        Albania  1989    male  55-74 years            6      142100
```

28	Albania	1989	male	15-24 years	12	323500	
29	Albania	1989	female	35-54 years	7	288600	
...	...	...	...	...	...	...	
27790	Uzbekistan	2012	female	25-34 years	148	2556673	
27791	Uzbekistan	2012	female	35-54 years	89	3474788	
27792	Uzbekistan	2012	male	5-14 years	67	2701361	
27793	Uzbekistan	2012	female	55-74 years	25	1283060	
27794	Uzbekistan	2012	female	75+ years	4	338557	
27795	Uzbekistan	2012	female	5-14 years	16	2578408	
27796	Uzbekistan	2013	male	35-54 years	481	3346411	
27797	Uzbekistan	2013	male	25-34 years	328	2644648	
27798	Uzbekistan	2013	female	15-24 years	323	3039740	
27799	Uzbekistan	2013	male	15-24 years	320	3171202	
27800	Uzbekistan	2013	male	55-74 years	119	1202790	
27801	Uzbekistan	2013	male	75+ years	13	221002	
27802	Uzbekistan	2013	female	25-34 years	146	2647820	
27803	Uzbekistan	2013	female	35-54 years	99	3547895	
27804	Uzbekistan	2013	female	75+ years	8	345180	
27805	Uzbekistan	2013	male	5-14 years	61	2720938	
27806	Uzbekistan	2013	female	55-74 years	21	1356298	
27807	Uzbekistan	2013	female	5-14 years	31	2595000	
27808	Uzbekistan	2014	male	35-54 years	519	3421300	
27809	Uzbekistan	2014	male	25-34 years	318	2739150	
27810	Uzbekistan	2014	female	15-24 years	347	2992817	
27811	Uzbekistan	2014	male	55-74 years	144	1271111	
27812	Uzbekistan	2014	male	15-24 years	347	3126905	
27813	Uzbekistan	2014	male	75+ years	17	224995	
27814	Uzbekistan	2014	female	25-34 years	162	2735238	
27815	Uzbekistan	2014	female	35-54 years	107	3620833	
27816	Uzbekistan	2014	female	75+ years	9	348465	
27817	Uzbekistan	2014	male	5-14 years	60	2762158	
27818	Uzbekistan	2014	female	5-14 years	44	2631600	
27819	Uzbekistan	2014	female	55-74 years	21	1438935	

	suicides/100k pop	country-year	HDI for year	gdp_for_year ($)	\
0	6.71	Albania1987	NaN	2,156,624,900	
1	5.19	Albania1987	NaN	2,156,624,900	
2	4.83	Albania1987	NaN	2,156,624,900	
3	4.59	Albania1987	NaN	2,156,624,900	
4	3.28	Albania1987	NaN	2,156,624,900	
5	2.81	Albania1987	NaN	2,156,624,900	
6	2.15	Albania1987	NaN	2,156,624,900	
7	1.56	Albania1987	NaN	2,156,624,900	
8	0.73	Albania1987	NaN	2,156,624,900	
9	0.00	Albania1987	NaN	2,156,624,900	
10	0.00	Albania1987	NaN	2,156,624,900	
11	0.00	Albania1987	NaN	2,156,624,900	
12	5.49	Albania1988	NaN	2,126,000,000	

13	5.33	Albania1988	NaN	2,126,000,000
14	4.48	Albania1988	NaN	2,126,000,000
15	4.46	Albania1988	NaN	2,126,000,000
16	2.85	Albania1988	NaN	2,126,000,000
17	2.71	Albania1988	NaN	2,126,000,000
18	2.03	Albania1988	NaN	2,126,000,000
19	1.91	Albania1988	NaN	2,126,000,000
20	1.79	Albania1988	NaN	2,126,000,000
21	1.41	Albania1988	NaN	2,126,000,000
22	0.00	Albania1988	NaN	2,126,000,000
23	0.00	Albania1988	NaN	2,126,000,000
24	8.89	Albania1989	NaN	2,335,124,988
25	6.35	Albania1989	NaN	2,335,124,988
26	4.71	Albania1989	NaN	2,335,124,988
27	4.22	Albania1989	NaN	2,335,124,988
28	3.71	Albania1989	NaN	2,335,124,988
29	2.43	Albania1989	NaN	2,335,124,988
...	...	...	...	...
27790	5.79	Uzbekistan2012	0.668	51,821,573,338
27791	2.56	Uzbekistan2012	0.668	51,821,573,338
27792	2.48	Uzbekistan2012	0.668	51,821,573,338
27793	1.95	Uzbekistan2012	0.668	51,821,573,338
27794	1.18	Uzbekistan2012	0.668	51,821,573,338
27795	0.62	Uzbekistan2012	0.668	51,821,573,338
27796	14.37	Uzbekistan2013	0.672	57,690,453,461
27797	12.40	Uzbekistan2013	0.672	57,690,453,461
27798	10.63	Uzbekistan2013	0.672	57,690,453,461
27799	10.09	Uzbekistan2013	0.672	57,690,453,461
27800	9.89	Uzbekistan2013	0.672	57,690,453,461
27801	5.88	Uzbekistan2013	0.672	57,690,453,461
27802	5.51	Uzbekistan2013	0.672	57,690,453,461
27803	2.79	Uzbekistan2013	0.672	57,690,453,461
27804	2.32	Uzbekistan2013	0.672	57,690,453,461
27805	2.24	Uzbekistan2013	0.672	57,690,453,461
27806	1.55	Uzbekistan2013	0.672	57,690,453,461
27807	1.19	Uzbekistan2013	0.672	57,690,453,461
27808	15.17	Uzbekistan2014	0.675	63,067,077,179
27809	11.61	Uzbekistan2014	0.675	63,067,077,179
27810	11.59	Uzbekistan2014	0.675	63,067,077,179
27811	11.33	Uzbekistan2014	0.675	63,067,077,179
27812	11.10	Uzbekistan2014	0.675	63,067,077,179
27813	7.56	Uzbekistan2014	0.675	63,067,077,179
27814	5.92	Uzbekistan2014	0.675	63,067,077,179
27815	2.96	Uzbekistan2014	0.675	63,067,077,179
27816	2.58	Uzbekistan2014	0.675	63,067,077,179
27817	2.17	Uzbekistan2014	0.675	63,067,077,179
27818	1.67	Uzbekistan2014	0.675	63,067,077,179
27819	1.46	Uzbekistan2014	0.675	63,067,077,179

	gdp_per_capita ($)	generation
0	796	Generation X
1	796	Silent
2	796	Generation X
3	796	G.I. Generation
4	796	Boomers
5	796	G.I. Generation
6	796	Silent
7	796	Boomers
8	796	G.I. Generation
9	796	Generation X
10	796	G.I. Generation
11	796	Generation X
12	769	G.I. Generation
13	769	Generation X
14	769	G.I. Generation
15	769	Silent
16	769	G.I. Generation
17	769	Generation X
18	769	G.I. Generation
19	769	Boomers
20	769	Boomers
21	769	Silent
22	769	Generation X
23	769	Generation X
24	833	G.I. Generation
25	833	Boomers
26	833	Silent
27	833	G.I. Generation
28	833	Generation X
29	833	Silent
...	...	...
27790	1964	Millenials
27791	1964	Generation X
27792	1964	Generation Z
27793	1964	Boomers
27794	1964	Silent
27795	1964	Generation Z
27796	2150	Generation X
27797	2150	Millenials
27798	2150	Millenials
27799	2150	Millenials
27800	2150	Boomers
27801	2150	Silent
27802	2150	Millenials
27803	2150	Generation X
27804	2150	Silent

```
27805                2150     Generation Z
27806                2150          Boomers
27807                2150     Generation Z
27808                2309     Generation X
27809                2309       Millenials
27810                2309       Millenials
27811                2309          Boomers
27812                2309       Millenials
27813                2309           Silent
27814                2309       Millenials
27815                2309     Generation X
27816                2309           Silent
27817                2309     Generation Z
27818                2309     Generation Z
27819                2309          Boomers

[27820 rows x 12 columns]>
```

```
In [23]: suicide.isnull().any()
```

```
Out[23]: country               False
         year                  False
         sex                   False
         age                   False
         suicides_no           False
         population            False
         suicides/100k pop     False
         country-year          False
         HDI for year           True
          gdp_for_year ($)     False
         gdp_per_capita ($)    False
         generation            False
         dtype: bool
```

```
In [ ]: del suicide["HDI for year"]
```

```
In [4]: suicide.columns
```

```
Out[4]: 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 [5]: suicide.index
```

```
Out[5]: RangeIndex(start=0, stop=27820, step=1)
```

```
In [9]: suicide.dtypes
```

```
Out[9]: country               object
        year                   int64
        sex                   object
        age                   object
        suicides_no            int64
        population             int64
        suicides/100k pop    float64
        country-year          object
         gdp_for_year ($)     object
        gdp_per_capita ($)     int64
        generation            object
        dtype: object
```

```
In [12]: df_country = suicide.groupby(by=['country']).mean()[['suicides/100k pop','population']
         print('Top 10 countries with highest suicide averages per 100k population')
         print(df_country.head(10))
```

```
Top 10 countries with highest suicide averages per 100k population
              country  suicides/100k pop    population
0           Lithuania          40.415573  2.598672e+05
1           Sri Lanka          35.295152  1.382770e+06
2  Russian Federation          34.892377  1.139137e+07
3             Hungary          32.761516  8.020782e+05
4             Belarus          31.075913  7.832234e+05
5          Kazakhstan          30.511282  1.209980e+06
6              Latvia          29.259325  1.779867e+05
7            Slovenia          27.827857  1.597961e+05
8             Estonia          27.276905  1.075032e+05
9             Ukraine          26.582321  3.828777e+06
```

```
In [13]: # Create pie charts of suicide numbers and population by category
         def pie_chart(dataframe, group_col):
             columns = [group_col, 'suicides_no','population']
             grouped_sum = dataframe[columns].groupby(group_col).sum()
             display(grouped_sum)

             fig = plt.figure()

             ax1 = fig.add_axes([0, 0, .65, .65])
             ax1.pie(grouped_sum.population,
                     labels=grouped_sum.index,
                     autopct='%1.1f%%')
             ax1.set_title('Global Population 1985-2016')

             ax2 = fig.add_axes([.65, 0, .65, .65])
             ax2.pie(grouped_sum.suicides_no,
                     labels=grouped_sum.index,
```

```
                autopct='%1.1f%%')
        ax2.set_title('Global Suicides 1985-2016')

        plt.show()

    # Create plots of suicide numbers and population by category
    def plot_time_series(dataframe, group_col):
        categories = dataframe[group_col].unique()
        for category in categories:
            df = dataframe[dataframe[group_col] == category][
                [group_col,'year','suicides_no','population']]

            group_data = df.groupby('year').mean()
            group_data.apply(rescale).plot(figsize=(10,2))
            plt.title(category)
            plt.show()
```

In [16]:
```
# Group data by year
suicide_by_year = suicide.groupby('year').sum()

# Display first and last 5 rows
display(suicide_by_year.head())
display(suicide_by_year.tail())
```

year	suicides_no	population	suicides/100k pop	gdp_per_capita ($)
1985	116063	1008600086	6811.89	3508548
1986	120670	1029909613	6579.84	4104636
1987	126842	1095029726	7545.45	5645760
1988	121026	1054094424	7473.13	5870508
1989	160244	1225514347	8036.54	6068424

year	suicides_no	population	suicides/100k pop	gdp_per_capita ($)
2012	230160	1912812088	11101.91	26058300
2013	223199	1890161710	10663.64	26911368
2014	222984	1912057309	10306.73	25665252
2015	203640	1774657932	8253.99	19516008
2016	15603	132101896	2147.39	4106420

In [20]:
```
def rescale(values):
    max_val = max(values)
    min_val = min(values)
    scaled_values = []
    for val in values:
        new_val = (val - min_val) / (max_val - min_val)
        scaled_values.append(new_val)
```

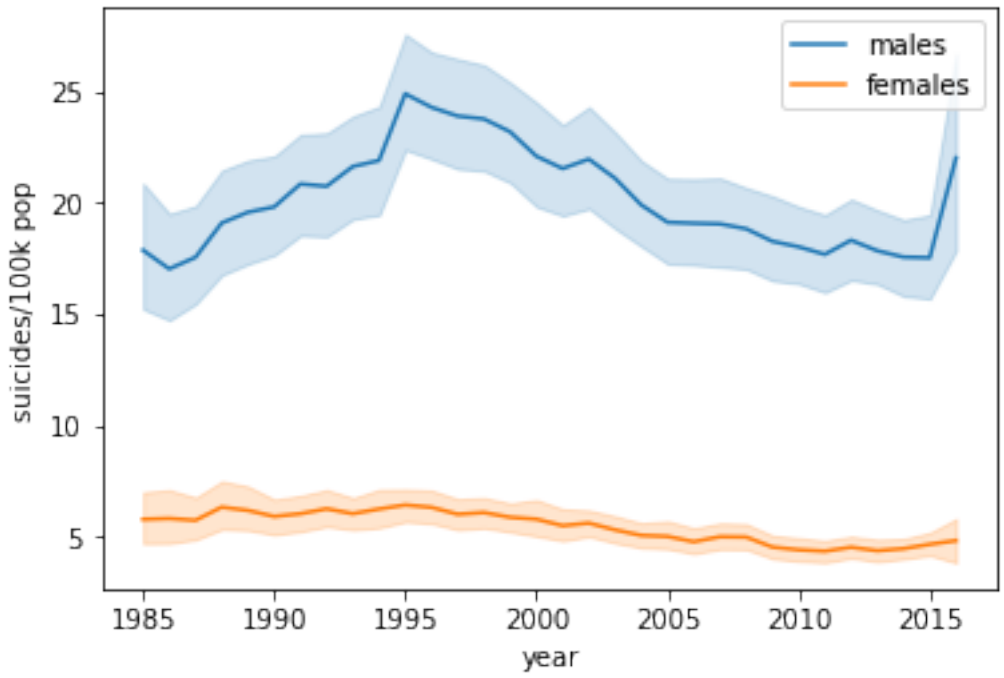
```
            return scaled_values

        rescaled = suicide_by_year.apply(rescale)

        display(rescaled.round(2).head())
        display(rescaled.tail())
```

year	suicides_no	population	suicides/100k pop	gdp_per_capita ($)
1985	0.42	0.47	0.37	0.00
1986	0.44	0.48	0.35	0.03
1987	0.46	0.52	0.43	0.09
1988	0.44	0.49	0.43	0.10
1989	0.60	0.59	0.47	0.11

year	suicides_no	population	suicides/100k pop	gdp_per_capita ($)
2012	0.892070	0.954704	0.715625	0.962527
2013	0.863128	0.942561	0.680599	0.998940
2014	0.862234	0.954300	0.652076	0.945750
2015	0.781807	0.880635	0.488026	0.683272
2016	0.000000	0.000000	0.000000	0.025520

In [24]:
```
male_population = suicide.loc[suicide.loc[:, 'sex']=='male',:]
female_population = suicide.loc[suicide.loc[:, 'sex']=='female',:]

p = sns.lineplot(x='year', y='suicides/100k pop', data=male_population)
q = sns.lineplot(x='year', y='suicides/100k pop', data=female_population)

_ = plt.legend(['males', 'females'])
```

```
C:\Users\jeffr\Anaconda3\lib\site-packages\scipy\stats\stats.py:1713: FutureWarning: Using a no
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```




```
In [25]: suicide.corr()
```

Out[25]:

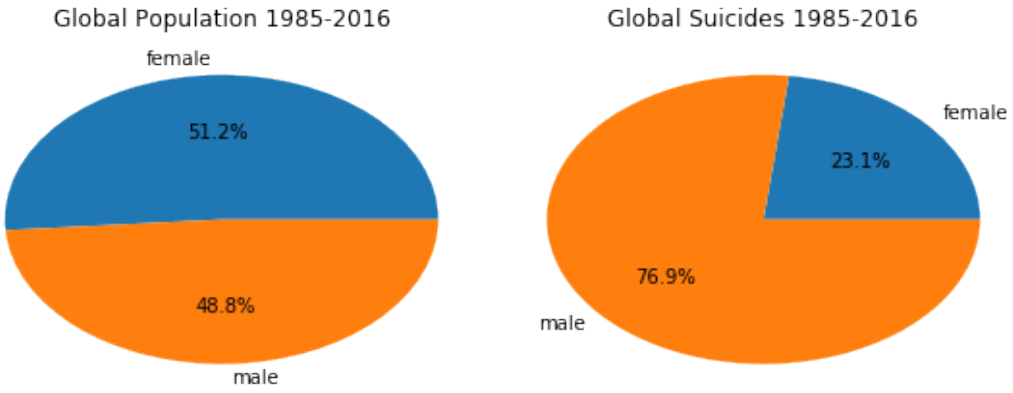
	year	suicides_no	population	suicides/100k pop	HDI for year	gdp_per_capita ($)
year	1.000000	-0.004546	0.008850	-0.039037	0.366786	0.339134
suicides_no	-0.004546	1.000000	0.616162	0.306604	0.151399	0.061330
population	0.008850	0.616162	1.000000	0.008285	0.102943	0.081510
suicides/100k pop	-0.039037	0.306604	0.008285	1.000000	0.074279	0.001785
HDI for year	0.366786	0.151399	0.102943	0.074279	1.000000	0.771228
gdp_per_capita ($)	0.339134	0.061330	0.081510	0.001785	0.771228	1.000000

```
In [27]: def pie_chart(dataframe, group_col):
             columns = [group_col, 'suicides_no','population']
             grouped_sum = dataframe[columns].groupby(group_col).sum()
             display(grouped_sum)

             fig = plt.figure()
```

```
    ax1 = fig.add_axes([0, 0, .65, .65])
    ax1.pie(grouped_sum.population,
            labels=grouped_sum.index,
            autopct='%1.1f%%')
    ax1.set_title('Global Population 1985-2016')

    ax2 = fig.add_axes([.65, 0, .65, .65])
    ax2.pie(grouped_sum.suicides_no,
            labels=grouped_sum.index,
            autopct='%1.1f%%')
    ax2.set_title('Global Suicides 1985-2016')

    plt.show()


def plot_time_series(dataframe, group_col):
    categories = dataframe[group_col].unique()
    for category in categories:
        df = dataframe[dataframe[group_col] == category][
            [group_col,'year','suicides_no','population']]

        group_data = df.groupby('year').mean()
        group_data.apply(rescale).plot(figsize=(10,2))
        plt.title(category)
        plt.show()
```

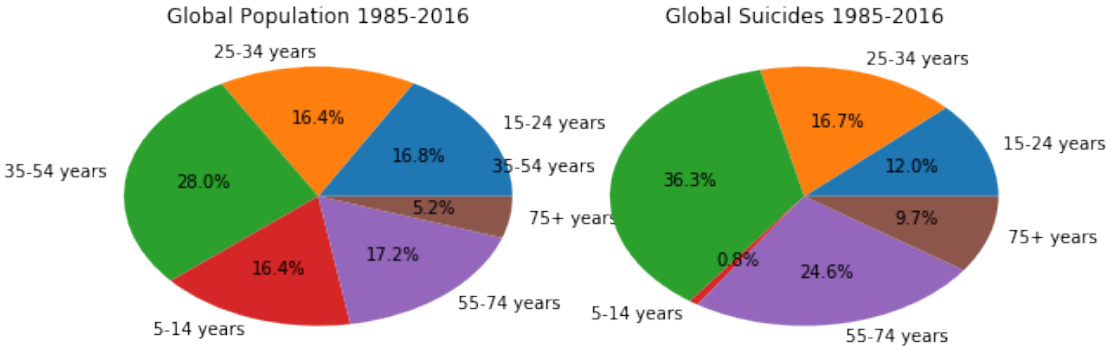
```
In [28]: pie_chart(suicide, "sex")
```

```
        suicides_no   population
sex
female      1559510  26272781857
male        5188910  25049376579
```

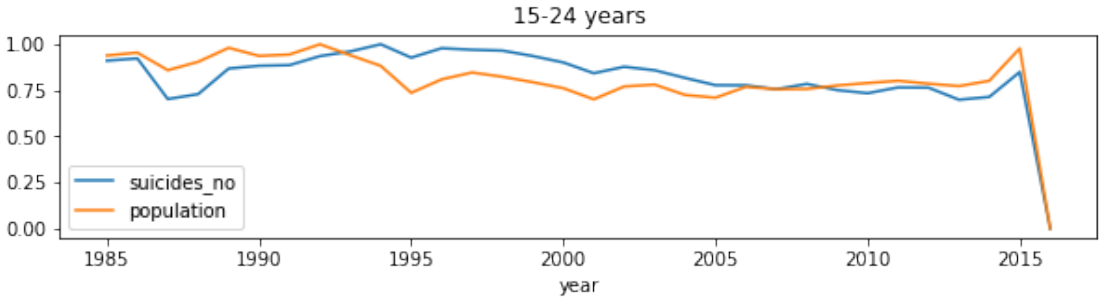


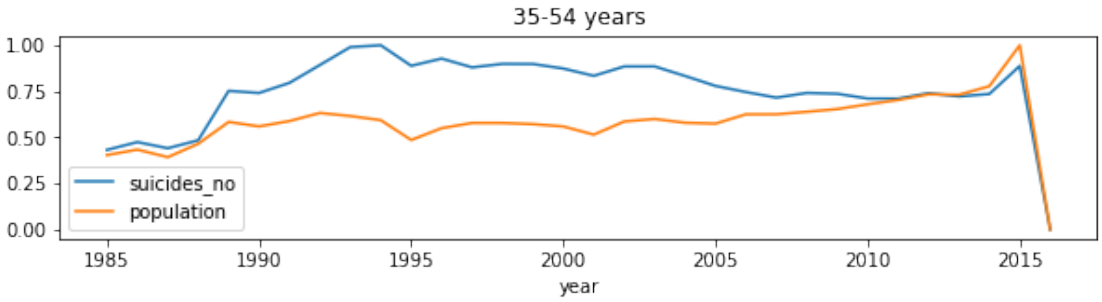
In [29]: pie_chart(suicide, "age")

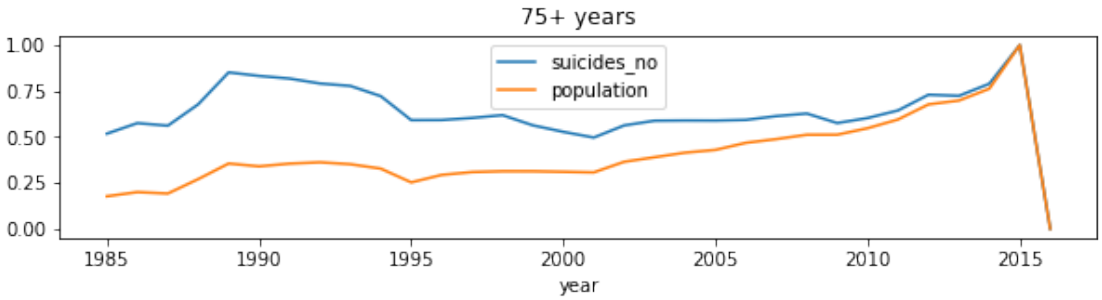
age	suicides_no	population
15-24 years	808542	8642946896
25-34 years	1123912	8438103587
35-54 years	2452141	14375888123
5-14 years	52264	8398693237
55-74 years	1658443	8803245340
75+ years	653118	2663281253

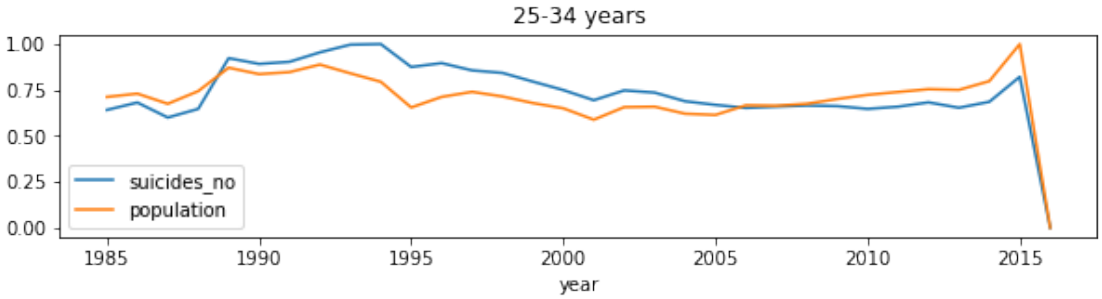


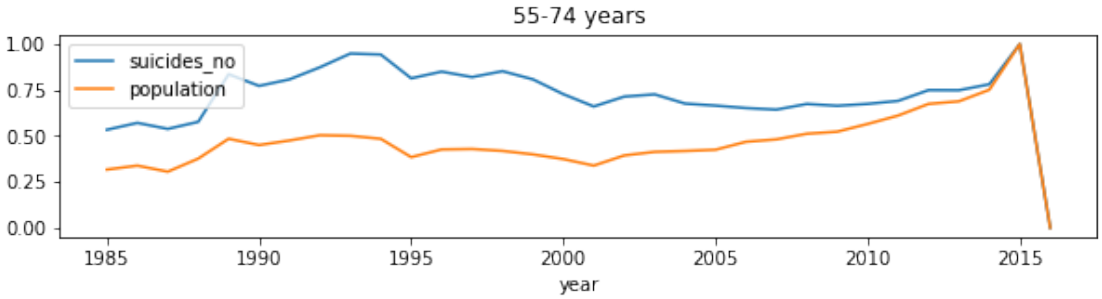
In [30]: plot_time_series(suicide, "age")

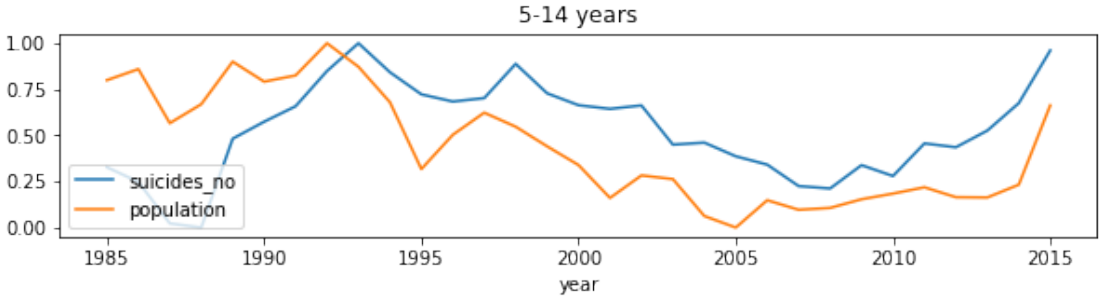





75+ years
suicides_no
population
1.00
0.75
0.50
0.25
0.00
1985
1990
1995
2000
2005
2010
2015
year


25-34 years
suicides_no
population
1.00
0.75
0.50
0.25
0.00
1985
1990
1995
2000
2005
2010
2015
year


55-74 years
suicides_no
population
1.00
0.75
0.50
0.25
0.00
1985
1990
1995
2000
2005
2010
2015
year


5-14 years
suicides_no
population
1.00
0.75
0.50
0.25
0.00
1985
1990
1995
2000
2005
2010
2015
year

```
In [ ]:
```