An examination of the correlation between Gross Domestic Product and Relative Suicide Rates between 1985 and 2016.

February 2021

CDL7508-2021

Carl Wilson [U0370630]

u0370630@unimail.hud.ac.uk

1. INTRODUCTION

1.1 – Background

Suicide rates have recently seen an increase in visibility with the media with a strong focus on mental health and reducing suicide rates (Bulman, 2020). Mental health specifically has been quite a prominent topic throughout 2020 when considering the level of isolation and financial hardships that many people have had to endure during national and local lockdowns to manage the spread of the COVID19 virus (Reuters, 2021). Data is available on Kaggle including variables in suicide numbers and Gross Domestic Product [GDP] by country, which can be used to explore the link between the two – as well as other potential statistical influences such as gender and age between the years of 1985 and 2016.

1.2 – Purpose Statement

The purpose of this report is to determine if relative suicide rates [suicides per 100,00 population] are correlated to the Gross Domestic Product of a country for a given year.

1.3 – Data Source

The *Suicide Rates Overview 1985 to 2016* dataset (Rusty, 2018) is available on *kaggle.com* and contains a number of interesting categorical variables which may or may not be related to suicide rates. The variables include country, year, sex, age, GDP and suicides per 100k population.

1.4 – The plan

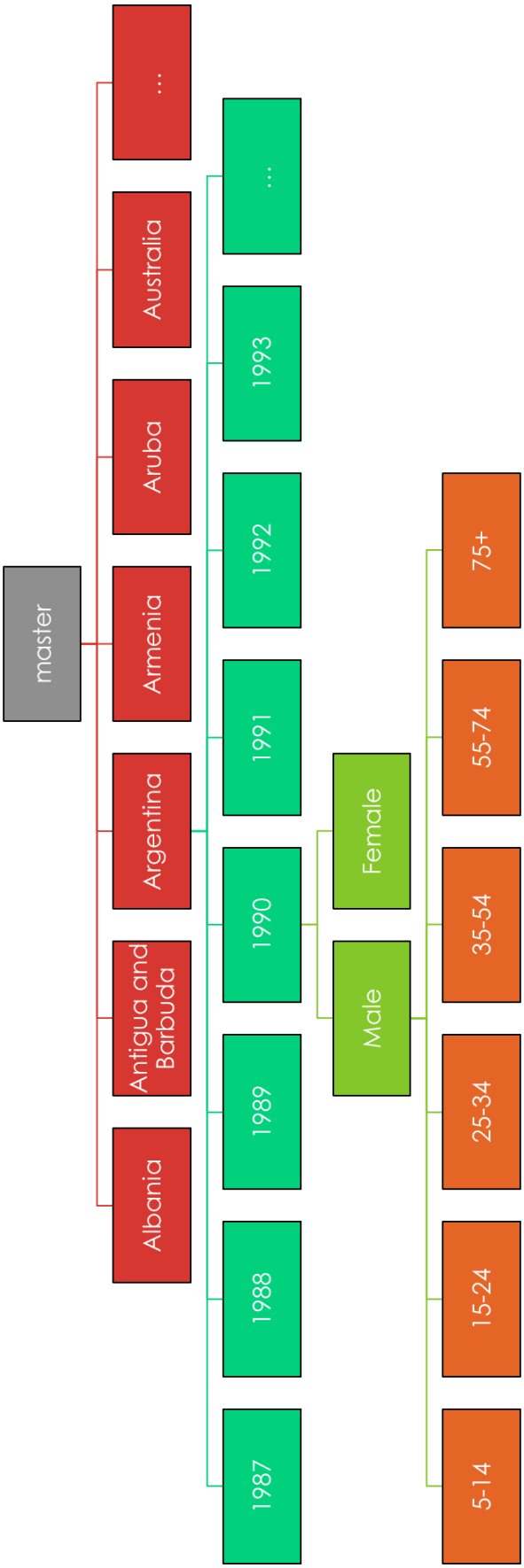
In order to determine the correlation between suicide and GDP, several other categorical variables need to be understood to ensure they are not influencing each other and ultimately the correlation. Table 1 details the analyses that will be completed in order to understand these variables before completing the correlation analysis.

Analysis	Hypothesis [h0]	Alternate [h1]	Test	Test Statistic	Significance Level
1	Male and Female Suicide Rates are equal	Male and Female Suicide Rates are **not** equal	Two Sample T-Test	T	P>0.05
2	Suicide rates are equal across all age groups	Suicide rates are **not** equal across all age groups	One way ANOVA	F	P>0.05
3	Suicide rates are equal across all years	Suicide rates are **not** equal across all years	One way ANOVA	F	P>0.05
4	Suicide rates are equal across all countries	Suicide rates are **not** equal across all countries	One way ANOVA	F	P>0.05
5	Suicide rates are **not** correlated to GDP per capita	Suicide rates are correlated to GDP per capita	Pearson's Correlation	R	P>0.05

Table 1: proposed method and plan of analyses.

2. THE DATA

2.1 – Data Structure



The master dataset contains 12 variables, six continuous and six categorical. Four of the categorical variables are of specific interest, as outlined on the previous page. These four variables filter down into a specific observation for a given country, year, sex and age. Each observation then has a suicides per 100k population continuous variable, that is the population of said category for country, year, sex and age. Each country and year also have a GDP per capita continuous variable. These are the six variables of interest in this study.

2.2 – Data Preparation & Cleaning

The initial dataset contained variables that were not required as they offered no value to the hypotheses to be tested in this report, as such: *country-year*, *HDI for year* and *generation* were all removed. Additionally, the *suicides/100k pop* variable was renamed to *suicides_100k_pop* to make it more "code friendly". The dataset was then sorted initially by year, then country, sex and age in ascending order. This would allow plots to be created consistently and in an order that made sense. Finally, whitespace within variables was removed and replaced with an underscore.

2.3 – Data Visualisation

A histogram of all the data, split only on *sex*, for *suicides per 100k population* was created in order to observe the distribution and make an initial assessment. Due to the large volume of data a *bin* count of 100 was used. As can be observed in *Figure 1* the data does not appear to fit a *normal* or *Gaussian* distribution; the data is heavily skewed towards zero. A log-normal transformation could potentially correct this. It can also be observed, now, that there are differences between *male* and *female* suicide rates; with significantly more observations in the "zero" bin to the far left in the female distribution.

In examining the data across different age groups, the data is presented in a boxplot for all countries split out by each age group on the x-axis and suicides per 100k population on the y-axis. The boxes are colour coded for sex and one plot was produced per year. The boxplot for 2001 can be observed in *Figure 2*. The other 31 years can be found in appendix 6.3. Visually, two trends can be observed: males generally have higher rates than females, and suicide rates for both males and females increase with age.

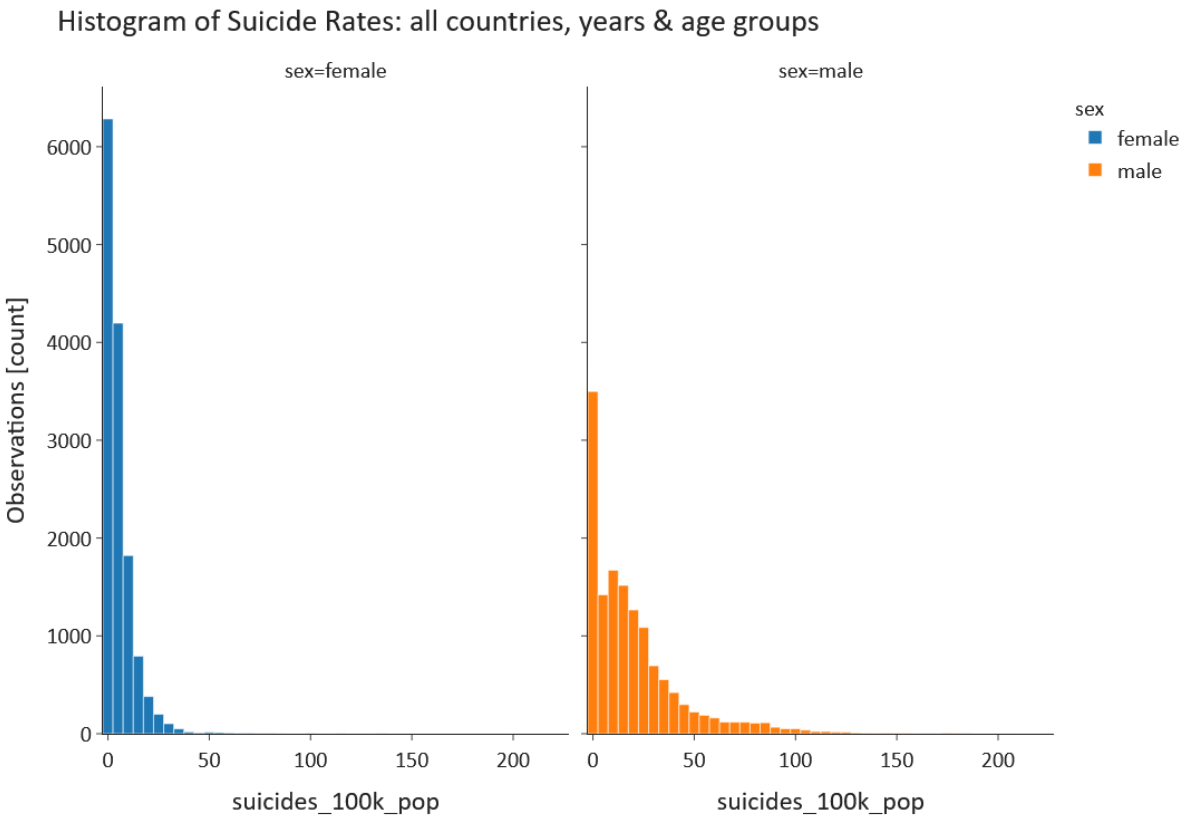


Figure 1: Histograms for male and female suicide rates for all countries, years and age groups.

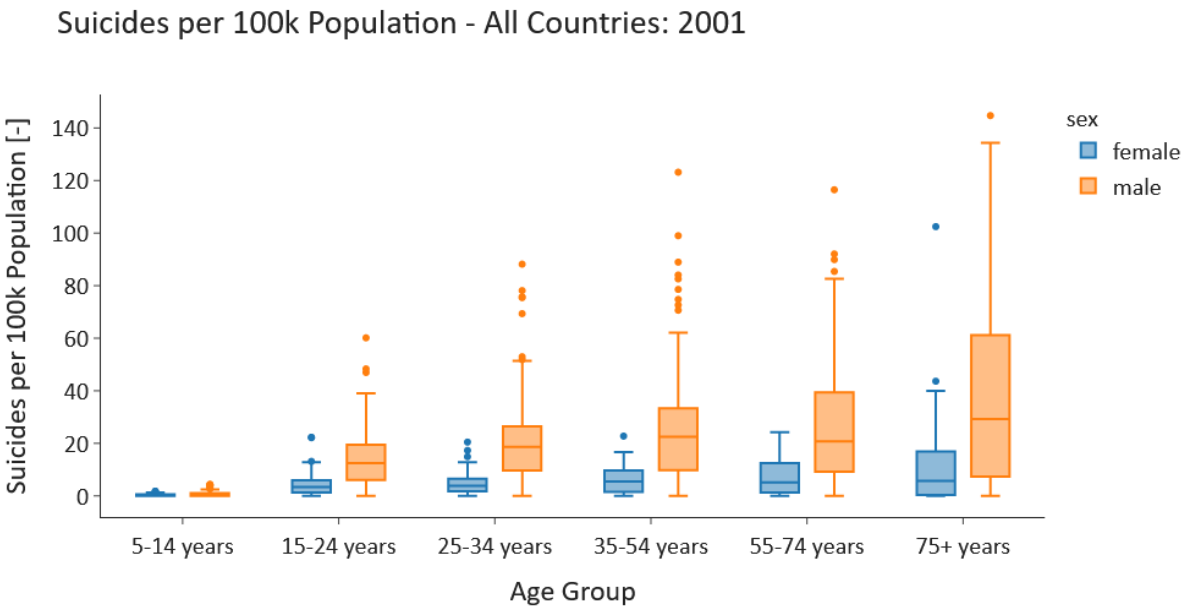


Figure 2: Boxplot of suicide rates across different age groups, between male and females for all countries in 2001.

Additionally, a box plot was created to illustrate suicide rates across the year category, see *Figure 3*. Relatively speaking, the rates look consistent year to year, with some small fluctuations – likely caused by outliers.

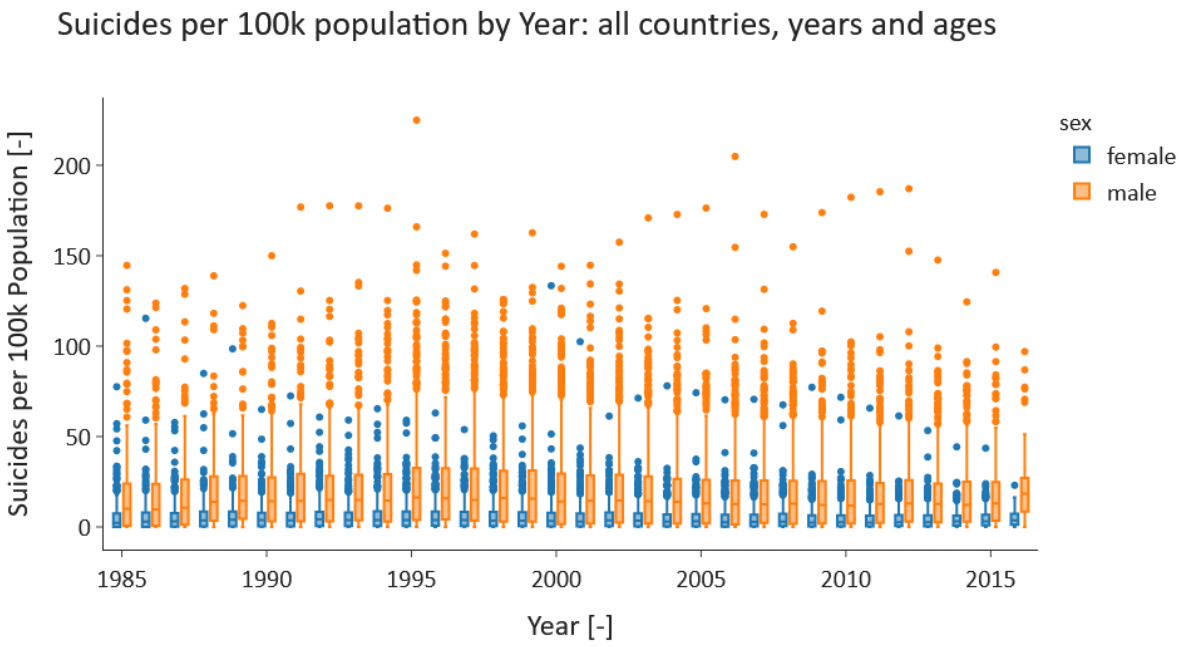


Figure 3: Suicides per 100k population by year, all countries, all age groups – colour coded by sex. The mean and upper quartile of the boxplots appear relatively constant across each year for each respective sex.

Examining the data across the country category using a boxplot in *figure 4*, it can be observed there is significantly more variety in the distributions within the male groups than the female groups.

The final plot created to initially visualise the data was a scatter plot of suicides per 100k population against GDP, with a boxplot of the distributions in the margin, see *Figure 5.* An initial observation here is that the data appears quite noisy, likely due to the number of categorical variables included.

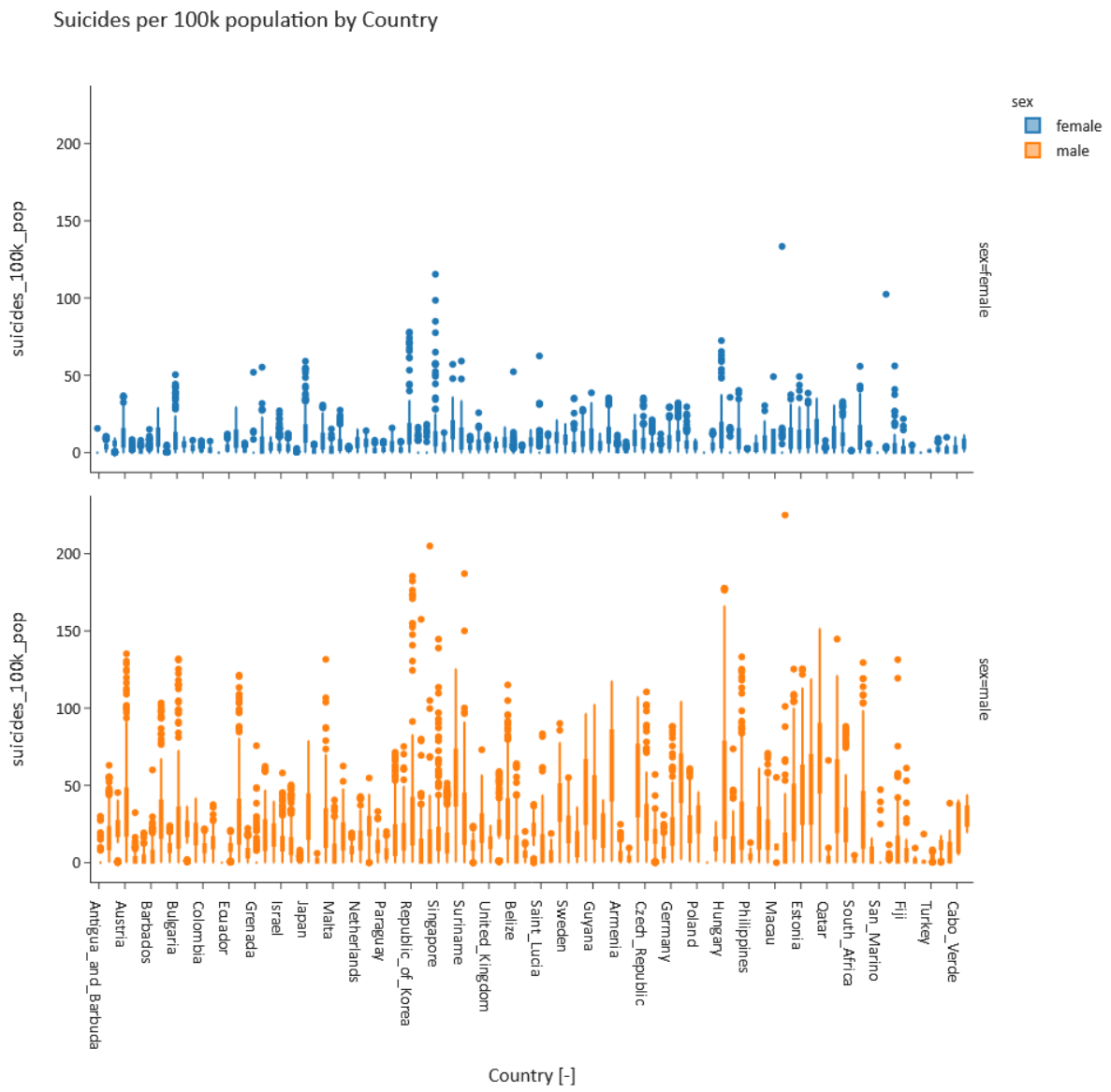


Figure 4: Suicides per 100k population, all ages and years, split by sex.

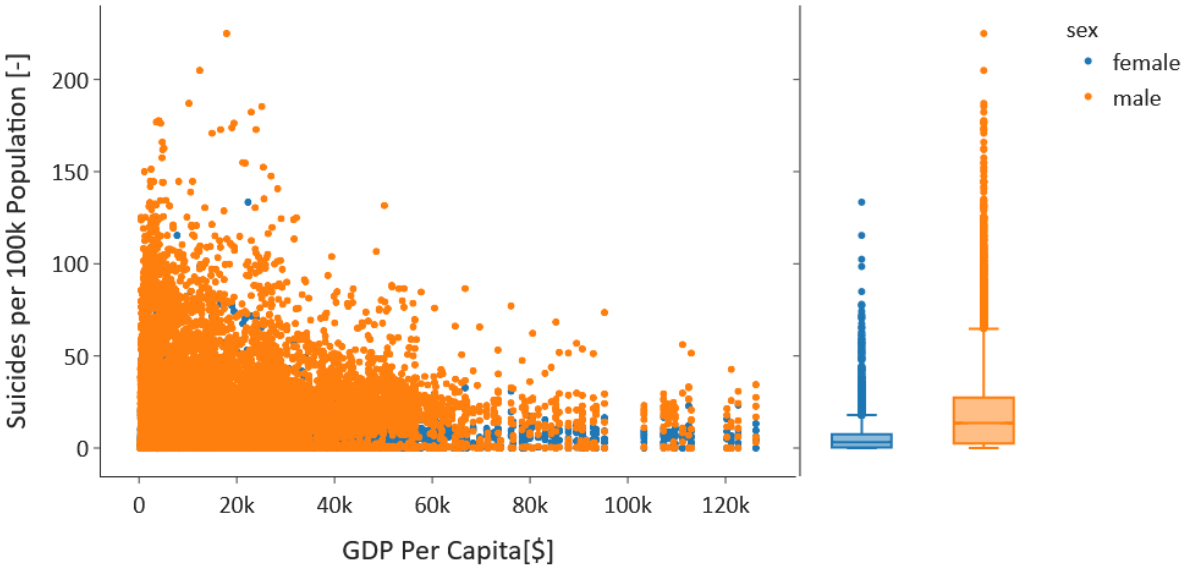


Figure 5: scatterplot of Suicides per 100k population vs. GDP per capital for all data, colour coded by sex.

2.3 – Statistical Analysis

Visual analysis of the data suggested the distributions would be non-normal. In order to test this the Kolmogorov-Smirnov test was used, due to the sample size being great than 5000 in count (Mason et al., 2003), to compare the sample distribution with a Gaussian distribution.

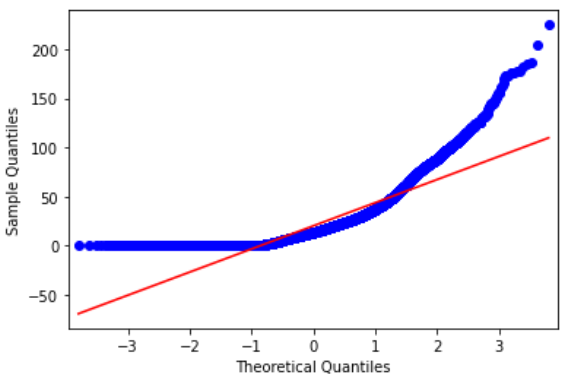


Figure 6: QQ plot for all male suicide rates

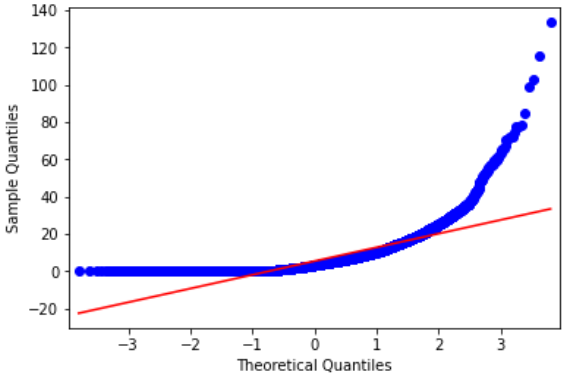


Figure 7: QQ plot for all female suicide rates

The QQ plots in figures 6 and 7 also strongly suggest the data is not normal and is finally backed up by a test statistic of 0.195 and 0.232 and p-values of 0.000 and 0.000 for males and females, respectively. Similarly, the Kolmogorov-Smirnov test was used to determine if any of the smaller groups, such as age [Table 4; appendix 6.2] and year [Table 5; appendix 6.2] fitted a normal distribution when split out by sex. All results returned a p-value of 0.000.

When each individual countries data was examined, split by sex, there were several interesting observations. Firstly, there were six instances where the data range equalled zero – so the test statistic could not be calculated. These were countries that typically had little data available [Dominica, Saint Kitts and Nevis, Cabo Verde and Oman]. There were also four instances where the p-value was greater or equal to 0.05. In all but one instance this was due to a small sample size of 5 – 6 observations. For the remaining statistical analysis, it was assumed that all data is not normal.

Due to the data not following a Gaussian distribution, non-parametric tests were required to be employed in place of those originally planned. In place of the two-sample t-test, the Mann-Whitney U test was used (Currell & Dowman, 2009). For the analysis of variance, in place of the one way ANOVA, the Kruskal-Wallis H test was used (Currell & Dowman, 2009). For the correlation test, Pearson's product moment correlation was replaced with a Spearman Rank-Order correlation (Corder & Foreman, 2014). The original analysis table can be updated to read as per Table 2.

Due to the significance of the results of analysis 1, indicating that the distributions of male and female suicide rates are not equal, multiple permutations of the remaining analyses were completed to understand the impact. For example, analysis 2 considered ages for all data, then split out by sex giving three sub-analyses for this test [all data, male, female].

Analysis	H_0	H_1	Test	Test Statistic	Significance Level
1	Male suicide rates = Female suicide rates	Male suicide rates ≠ Female suicide rates	Mann-Whitney U	U	P > 0.05
2	Distributions of age **are** equal	Distribution of ages **are not** equal	Kruskal-Wallis	H	P > 0.05
3	Distributions of country **are** equal	Distribution of count **are not** equal	Kruskal-Wallis	H	P > 0.05
4	Distributions of countries **are** equal	Distribution of counties **are not** equal	Kruskal-Wallis	H	P > 0.05
5	Suicide rates **are not** correlated to GDP	Suicide rates **are** correlated to GDP	Spearman's Rank Order	R	P > 0.05

Table 2: plan of analysis, with hypothesis, test statistics and significance levels.

The same methodology was applied to the year and country variables. In each instance, while the distributions and test statistics did change – the p-value remained below the significance level.

Due to each variable having individual distributions and not being able to consider them as a group, the correlation analysis needed to be completed on each individual sex, age and year. As GDP is provided as per capita, one value is available for each country and year and is the same for both sexes and all age groups for a given country and year. This meant that at total of 384 correlation analyses were required to be completed.

3. THE RESULTS

3.1 – Results discussion

The Mann-Whitney U test indicated that male and female suicide rates [for all countries, age groups and years] are different to each other, confirming the observations made within the visualisation of the dataset.

Extensive use of the Kruskal-Wallis test across age groups, years and country revealed that each variable did not have equal distributions of suicide rates. Each of the three categorical variables were also compared using only the male and female samples and this did not change the conclusions. It is worth noting however, that while there were 32 different year groups compared and only 6 age groups, the test statistic ***H*** was nearly an order of magnitude lower for the year groups – indicating that the age distributions are more dissimilar to each other than the year distributions are, as observed when visualising the data.

Analysis	Test	Test Statistic	P - Value	Result	Conclusion
1	Mann-Whitney U	U = 53248013	0.000	Reject H_0	Distribution of suicide rates between males and females are not equal
2a	Kruskal-Wallis H Test	H = 7008	0.000	Reject H_0	Distribution of suicide rates are not equal across age groups
2b	Kruskal-Wallis H Test	H = 4589	0.000	Reject H_0	Distribution of suicide rates are not equal across age groups in males
2c	Kruskal-Wallis H Test	H = 3012	0.000	Reject H_0	Distribution of suicide rates are not equal across age groups in females
3a	Kruskal-Wallis H Test	H = 112.5	0.000	Reject H_0	Distribution of suicide rates are not equal across year groups
3b	Kruskal-Wallis H Test	H = 87.3	0.000	Reject H_0	Distribution of suicide rates are not equal across year groups in males
3c	Kruskal-Wallis H Test	H = 85.9	0.000	Reject H_0	Distribution of suicide rates are not equal across year groups in females
4a	Kruskal-Wallis H Test	H = 8708	0.000	Reject H_0	Country does impact suicide rates
4b	Kruskal-Wallis H Test	H = 5592	0.000	Reject H_0	Country does impact suicide rates in males
4c	Kruskal-Wallis H Test	H = 6070	0.000	Reject H_0	Country does impact suicide rates in females
5	Spearman's Rank Order	See Fig 8	0.000	Reject H_0	GDP is correlated to suicide rates

Table 3: summary results table of completed statistical analyses.

It has been shown that sex, age, year and country all have an influence on suicide rates. The suicide rates for each of these categorical variables have been broken out and correlated with GDP per capita and show that there is a correlation between suicides per 100k population and GDP per capita, with a p-value of 0.000 in each instance using Spearman's correlation.

Due to the number of analyses completed, a histogram of correlation coefficients was created, see *Figure 8*. The histograms of the resulting correlations suggest a weak positive correlation – with a mean rho correlation of 0.126. This correlation would suggest that suicide rates are higher when GDP is higher.

Creating a heatmap of the correlation coefficients for year vs. age, split out by males [*Figure 9]* and females *[Figure 10]* shows where the stronger positive and negative correlations are. The strongest positive correlations are ages 35-54 and 55-74 years in 1986 and 1987 for both males and females; this is where suicide rates increase with GDP per capita increasing. The strongest negative correlations, which were weaker than the strongest positive correlations were all in the youngest age group of 5-14 years – particularly in females.

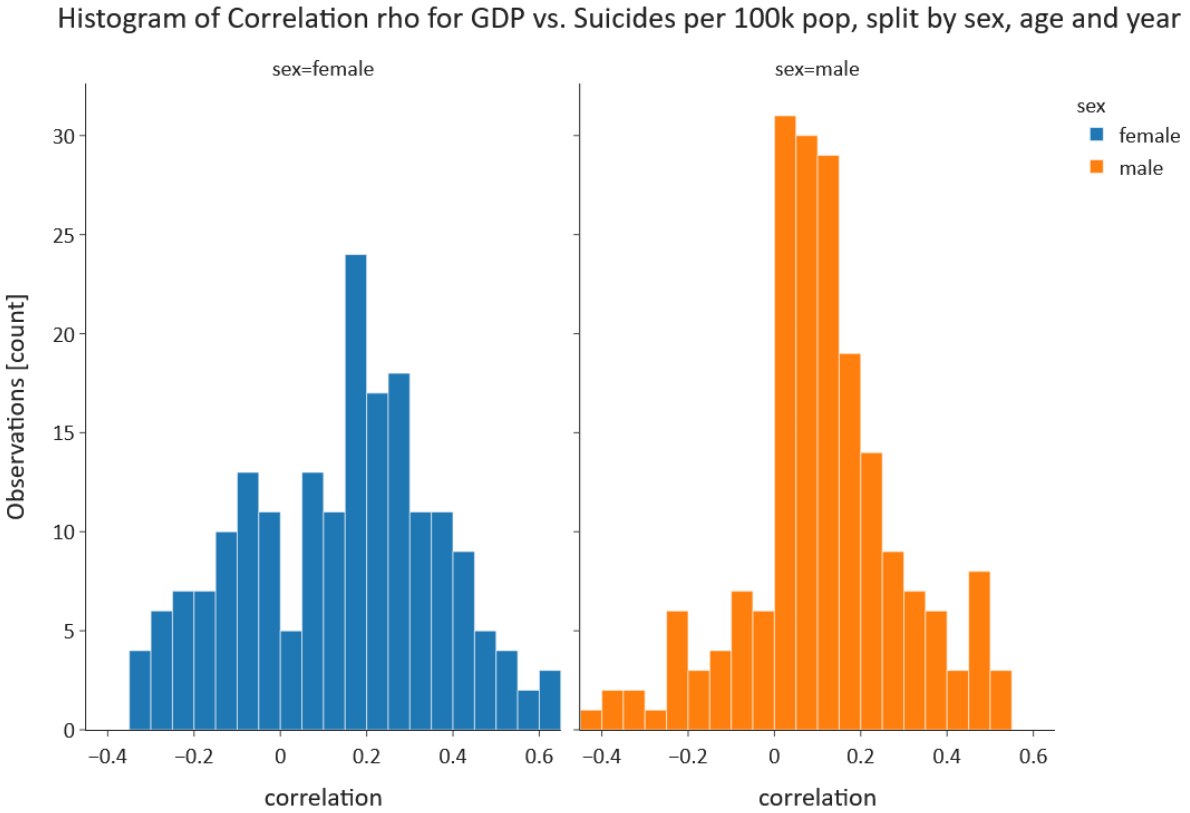


Figure 8: Histogram of "correlation statistics" for GDP vs. Suicides split by sex, age and year.

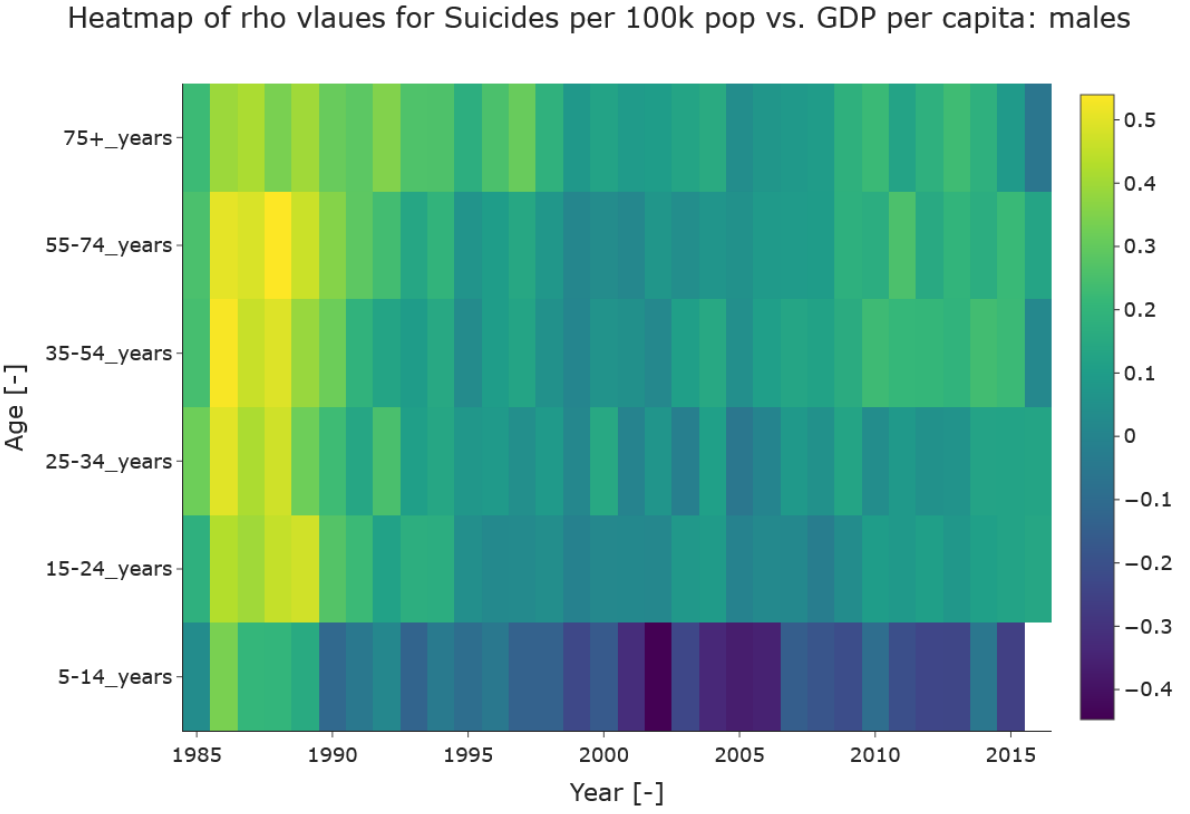


Figure 9: Heatmap of correlation coefficients by year and age for males

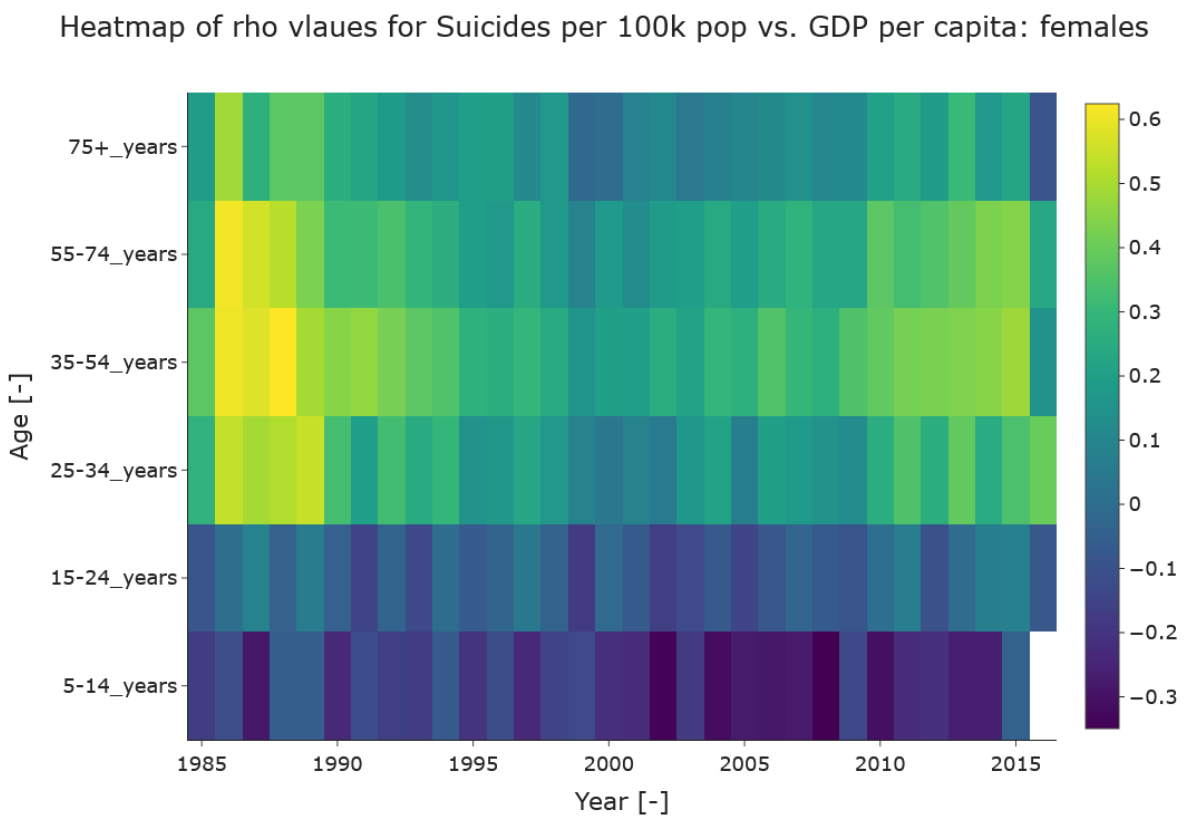


Figure 10: Heatmap of correlation coefficients by year and age for females.

4. CONCLUSION

It is concluded that suicides rates are correlated to GDP per capita, with a mean correlation of 0.114 and a p-value of 0.000 when tested using Spearman's rho correlation statistic.

5. REFERENCES

Bulman, M. (2020, 08/04/2020). Suicides in England hit record high in 2019, figures show. *The Independent*. https://www.independent.co.uk/news/uk/home-news/suicide-rate-england-record-high-figures-latest-a9454446.html

Corder, G. W., & Foreman, D. I. (2014). *Nonparametric statistics: a step-by-step approach* (2nd;2nd; ed.). Wiley. http://hud.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwjR3vS8JA9GH6RSEqK9RKJMIvpey2m3NBhJnihzSIVfTpmNutJN3CqVB_Yn9V7-amU4r8MrjdG4979_Z-3b33ABS5KlXWZIJwhHRet3RVk7hMLMV0JLtvy32Jo86atyXTe139rkGN5wR8R6kxb1O7ag5H5i-5DiJrVdxy_Iz6TF0PwuN1EWWitPN4y9ABNv7Y4WAz0am-GoiSp2oN-ZSKtBUqy-RiESts9cr-zArgmsZ9eVodvLremIsXnAXoBdRkPOXlGeJYYr2MfvUtSMmqImy37eUsazaMWCQIzRCiBm2MSB0dLkUmywHqi7AyVTQprYypMI4zpv-OMhHl5cQXnZDM8fSLuytGc8b3TKQmpfiIadD2LqS4SKvYgwR3s7ATGr-lULT4Wch0FwVkcZQWRvC8hvQ-5HqeK0qUj0T3L6vkL6YO4LzdMpqdSgwrCwNSbJ3cyiEkXc_lOSgpsmZbjk4t00a1qddNyyGq4yCT6DXNUaU8nMaWymbD4EjZZyv0-R8IiZaHQkgm9jGv68GUQHmrlJA8lCLKseDj8KYsa900SU2roRmdh7NN1lbYDOwI0miJ0Xls5xiSgqNOYAs_Ka7zTRFSjdbDy9MPRTP96w

Currell, G., & Dowman, A. (2009). *Essential mathematics and statistics for science* (2nd ed.). Wiley-Blackwell. http://hud.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwjV1LS8NAEB7aeDEg2KpofbAXRcHK2myT7LH0QQ7WU8Rj2CS7eKgpdI3-fWfy0NJTL4GF2SUsycw3O_N9C-CNnvhwxyfkJlUZ-kqufMTPMg-CsZ-mQiOgkNpXRE6eydelfJmI-L0D7e2p1GT5varKhrXbxsEPNUvXaqjVDVObquWX6qlRSaIb1qzyuzZB4iLAGEmS513My7hw4EiI6G2WYGocJ9NJXJHUG5uwFeRp57jg2rX6KHMh8LEVdxbHcKCjJNCDji764C7_NFZtH3rNb2nZfaMd_XACfG6JTYQfFfv8N2aqyBlRh2pVZoZAlTWh7xQeF_N4Gg23XiFpznQShEkkRsYJr5CK17N3Bk6xLvQ5MJMFmVa4tejohExlqLQygRcaz1eYX3kXcLvXkoM97S7hsC6x0LnEFThfm1JfQxfn3Ozu9i_dCJZb

Mason, R. L., Gunst, R. F., & Hess, J. L. (2003). *Statistical design and analysis of experiments: with applications to engineering and science* (2nd ed.). J. Wiley. http://hud.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwjV3dS8MwED_28aIguqk4P_sgQ1FHkyadFUTmnAzU-VLRt5KtqQ60A-cG_ov-VebSZt2Gwh5auCa9a0kuuUvufgFwaM0-mxsTGLN7LIwiyrpePRJUel3nvOcoVeQiYjqt-8brPHj3DeY_5-DHpMa8jcKaeP8Qf-Q6YNYqRjl-m3Omrvrp9jquMjHWfroJlAPs_9PCujGVU33ZR8hT5UZwylxbQ2KR08laYatTHY57ul7Tf6yOav3XePAp8YEMtHis9fU5ktWxkpFJvTCqnoci5Q7abitZadBsTLSOOExZERnonYNTtHGiEiNeaSHlOp1e6acyeeyXhCBMGfu2M0UQgslreOKU61J9phVxEdUNUc4VoSbjpKRusIdShu6ETnjas7SpP-HMU9owpwmd8Z8x2JeHA6FakjF1m5q9b9egKDGlowQ5GZdhNTW8rXRYG5ZheQqVsQwlU2AdpaDcx-tQQ6tcg1qLdyvUYS-WiEN1JaAu1iCystMShhtwctvym-2zqW8K0qWyYL4jOJtQiAex3AJLuB4VHpchPydMkkhwGkp037punfEwrMCB-eFA73OnwbVB67pJECWRV-BwEZnbi1XbgSUdV6hXg3ahgH1wD_Lqlf35nrYP-Tuf_QLp7PmT

Reuters. (2021, 16/01/2021). Japan's suicides jump 16% in COVID-19 2nd wave after fall in 1st wave - study. *The Guardian Online*. https://www.theguardian.com/world/2021/jan/16/japans-suicide-rate-rises-16-in-second-wave-of-covid-study-finds

Rusty. (2018). *Suicide Rates Overview 1985 to 2016* Version Version 1). https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016/metadata

6. APPENDIX

6.1 Python Code

6.1 Python Code

```
#import libraries for manipulating and plotting data
import numpy as np
import pandas as pd

import scipy.stats as stats
from statsmodels.graphics.gofplots import qqplot
from matplotlib import pyplot

import plotly.io as pio
import plotly.graph_objects as go
import plotly.express as px
pio.templates.default="simple_white"
```

Import and clean the dataset

```
#import dataset from csv into a Pandas Dataframe
MASTER=pd.read_csv("master.csv")

#rename "suicides/100k pop" to be more code friendly
MASTER=MASTER.rename(columns={"suicides/100k pop":"suicides_100k_pop"})

#remove columns that aren't required
MASTER=MASTER.drop(columns=["country-year","HDI for year","generation"])

#create a category data type for years in order to correct the order
CAT_AGE=pd.CategoricalDtype(["5-14 years","15-24 years","25-34 years",
                             "35-54 years","55-74 years","75+ years"],ordered=True)
#apply the category data tyep
MASTER["age"]=MASTER["age"].astype(CAT_AGE)

#sort the dataframe by year > country > sex > age into ascending order
MASTER=MASTER.sort_values(["year","country","sex","age"],ascending=[True,True,True,True])

#remove whitespace from variables

MASTER["country"]=MASTER["country"].str.replace(" ","_")
MASTER["age"]=MASTER["age"].str.replace(" ","_")

#review the updated dataframe
MASTER
```

Visualise the Data

```
#create histograms of suicides per 100k population for all groups, split by sex
FIG1 = px.histogram(MASTER,x="suicides_100k_pop",color="sex",nbins=100,template="simple_whi
te",
                    title="Histogram of Suicide Rates: all countries, years & age groups",
                    width=1080,height=760,log_y=False,facet_col="sex")
FIG1.update_layout(font=dict(family="Calibri",size=20),
                     yaxis_title="Observations [count]")
FIG1.show()

#create a series of bar charts for suicides per 100k population vs. age group, all countrie
s, colourcoded by sex,
#individual plots per year
for i in MASTER.year.unique():
```

```
    WORKING_DATA=MASTER[MASTER["year"]==i]
    FIG2=px.box(WORKING_DATA,
                template="simple_white",
                y="suicides_100k_pop",
                x="age",
                color="sex",
                title="Suicides per 100k Population - All Countries: {}".format(i))
    FIG2.update_layout(font=dict(family="Calibri",size=20),
                       xaxis_title="Age Group",
                       yaxis_title="Suicides per 100k Population [-]")
    FIG2.show()

#create boxplot of suicides per 100k population vs year for all groups, split by sex.
FIG3 = px.box(MASTER,
              y="suicides_100k_pop",x="year",color="sex",
              title="Suicides per 100k population by Year: all countries, years and ages",
              template="simple_white",
              width=960,height=540)

FIG3.update_layout(font=dict(family="Calibri",size=20),
                   xaxis_title="Year [-]",
                   yaxis_title="Suicides per 100k Population [-]")
FIG3.show()

#create boxplot of suicides per 100k population vs. country for all groups, split by sex.
FIG4 = px.box(MASTER,
              y="suicides_100k_pop",x="country",color="sex",
              title="Suicides per 100k population by Country",
              template="simple_white",
              width=960,height=960,facet_row="sex")

FIG4.update_layout(font=dict(family="Calibri",size=14),
                   xaxis_title="Country [-]")
FIG4.show()

#create scatterplot of suicides per 100k population vs. GDP per capital for all groups, spl
it by sex.
FIG5 = px.scatter(MASTER,x="gdp_per_capita",y="suicides_100k_pop",color="sex",marginal_y="b
ox",
                  template="simple_white")

FIG5.update_layout(font=dict(family="Calibri",size=20),
                   xaxis_title="GDP Per Capita[$]",
                   yaxis_title="Suicides per 100k Population [-]")

FIG5.show()
```

Check the distribution of the data vs. Gaussian distribution

Males vs. Females

```
#short function to detemine data length and then compare the  sample vs. a normal distribut
ion with the
#Kolmogorov-Smirinov test if the sample size is >=5000 or use the Shapiro-Wilkes test if lo
wer.
def normalityTester(DATA):
    if DATA.size>=5000:
        DISTRIBUTION = getattr(stats,"norm")
        PARAMS=DISTRIBUTION.fit(DATA)
        TEST_STAT,PVAL = stats.kstest((DATA),"norm",PARAMS)
    else:
        TEST_STAT,PVAL = stats.shapiro(DATA)
    return(TEST_STAT,PVAL)

#create two new DataFrames for all male and female data to be kept seperate and compared.
ALL_MALES=MASTER[(MASTER["sex"]=="male")]
ALL_FEMALES=MASTER[(MASTER["sex"]=="female")]
```

```
#QQ plot of all male data
ALL_MALE_QQPLOT=qqplot(ALL_MALES.suicides_100k_pop,line="s")

#QQ plot of all female data
ALL_FEMALE_QQPLOT=qqplot(ALL_FEMALES.suicides_100k_pop,line="s")

#normal test for all male data
TEST_STAT, PVAL = normalityTester(ALL_MALES.suicides_100k_pop)

print("Test Statistic=%.3f, p-value=%.3f"%(TEST_STAT,PVAL))

#normal test for all female data
normalityTester(ALL_FEMALES.suicides_100k_pop)

print("Test Statistic=%.3f, p-value=%.3f"%(TEST_STAT,PVAL))
```

Normality test for age groups, split by sex

```
#create an empty dataframe for the results
RESULTS = pd.DataFrame(columns=["sex","age","stat","p-value"])

for s in MASTER.sex.unique():
    for a in MASTER.age.unique():
        WORKING_DATA=MASTER[(MASTER["sex"]==s)&(MASTER["age"]==a)]
        TEST_STAT, PVAL = normalityTester(WORKING_DATA.suicides_100k_pop)
        TEST_STAT=round(TEST_STAT,5)
        PVAL=round(PVAL,5)
        RESULTS=RESULTS.append({"sex":s,"age":a,"stat":TEST_STAT,"p-value":PVAL},ignore_ind
ex=True)

#export results to csv for capturing in report
RESULTS.to_csv("normality_age_sex.csv")

#display results in Jupyter
RESULTS
```

Normality test for years, split by sex

```
RESULTS = pd.DataFrame(columns=["sex","year","stat","p-value"])

for s in MASTER.sex.unique():
    for y in MASTER.year.unique():
        WORKING_DATA=MASTER[(MASTER["sex"]==s)&(MASTER["year"]==y)]
        TEST_STAT, PVAL = normalityTester(WORKING_DATA.suicides_100k_pop)
        TEST_STAT=round(TEST_STAT,5)
        PVAL=round(PVAL,5)
        RESULTS=RESULTS.append({"sex":s,"year":y,"stat":TEST_STAT,"p-value":PVAL},ignore_in
dex=True)

RESULTS.to_csv("normality_year_sex.csv")
RESULTS
```

Normality test for country, split by sex

```
RESULTS = pd.DataFrame(columns=["sex","country","stat","p-value"])

for s in MASTER.sex.unique():
    for c in MASTER.country.unique():
        WORKING_DATA=MASTER[(MASTER["sex"]==s)&(MASTER["country"]==c)]
        TEST_STAT, PVAL = normalityTester(WORKING_DATA.suicides_100k_pop)
        TEST_STAT=round(TEST_STAT,5)
        PVAL=round(PVAL,5)
        RESULTS=RESULTS.append({"sex":s,"country":c,"stat":TEST_STAT,"p-value":PVAL},ignore
_index=True)

RESULTS.to_csv("normality_country_sex.csv")
RESULTS
```

```
#check for any results with p-value greater than 0.05
RESULTS[(RESULTS["p-value"]>=0.05)]

#QQ plot of females in Turkey
qqplot(MASTER[(MASTER["sex"]=="female")&(MASTER["country"]=="Turkey")].suicides_100k_pop,li
ne="s");

#QQ plot of females in Mongolia
qqplot(MASTER[(MASTER["sex"]=="female")&(MASTER["country"]=="Mongolia")].suicides_100k_pop,
line="s");

#QQ plot of males in Cabo Verde
qqplot(MASTER[(MASTER["sex"]=="male")&(MASTER["country"]=="Cabo_Verde")].suicides_100k_pop,
line="s");

#QQ plot of males in Mongolia
qqplot(MASTER[(MASTER["sex"]=="male")&(MASTER["country"]=="Mongolia")].suicides_100k_pop,li
ne="s");
```

Analysis 1: Mann-Whitney U test - are diistributions of suicides for males and females equal

```
TEST_STAT, PVAL = stats.mannwhitneyu(ALL_MALES.suicides_100k_pop,ALL_FEMALES.suicides_100k_
pop)
print("Test Statistic=%.3f, p-value=%.6f"%(TEST_STAT,PVAL))
```

Analysis 2: Kruskal-Wallis H test - Are distributions of suicide rates equal across age groups?

2a) First consider both male and female

```
#create a new dictionary for age data
AGE_DICT={}

#loop through each country and add suicides_100k_pop data to dictionary
for A in MASTER["age"].unique():
    AGE_DICT[A]=MASTER["suicides_100k_pop"][MASTER["age"]==A].values

#run the Kruskal-Wallis test
TEST_STAT, PVALUE = stats.kruskal(*AGE_DICT.values())
print("test statistic: %.4f"%TEST_STAT)
print("p-value: %.10f"%PVALUE)

#manual test to check the dictionary loop is working
TEST_STAT, PVALUE = stats.kruskal(MASTER[(MASTER["age"]=="5-14_years")].suicides_100k_pop,
                                  MASTER[(MASTER["age"]=="15-24_years")].suicides_100k_pop,
                                  MASTER[(MASTER["age"]=="25-34_years")].suicides_100k_pop,
                                  MASTER[(MASTER["age"]=="35-54_years")].suicides_100k_pop,
                                  MASTER[(MASTER["age"]=="55-74_years")].suicides_100k_pop,
                                  MASTER[(MASTER["age"]=="75+_years")].suicides_100k_pop)
print("test statistic: %.4f"%TEST_STAT)
print("p-value: %.4f"%PVALUE)
```

2b) Now consider only male distributions

```
AGE_DICT_MALE={}

for A in ALL_MALES["age"].unique():
    AGE_DICT_MALE[A]=ALL_MALES["suicides_100k_pop"][ALL_MALES["age"]==A].values

TEST_STAT, PVALUE = stats.kruskal(*AGE_DICT_MALE.values())
print("test statistic: %.4f"%TEST_STAT)
print("p-value: %.10f"%PVALUE)
```

2c) Finally consier only female distibutions

```
AGE_DICT_FEMALE={}

for A in ALL_FEMALES["age"].unique():
    AGE_DICT_FEMALE[A]=ALL_FEMALES["suicides_100k_pop"][ALL_FEMALES["age"]==A].values

TEST_STAT, PVALUE = stats.kruskal(*AGE_DICT_FEMALE.values())
```

```
print("test statistic: %.4f"%TEST_STAT)
print("p-value: %.10f"%PVALUE)
```

Analysis 3: Are distributions of suicide rates equal across different years?

3a) Both sexes

```
YEAR_DICT={}

for Y in MASTER["year"].unique():
    YEAR_DICT[Y]=MASTER["suicides_100k_pop"][MASTER["year"]==Y].values

TEST_STAT, PVALUE = stats.kruskal(*YEAR_DICT.values())
print("test statistic: %.4f"%TEST_STAT)
print("p-value: %.10f"%PVALUE)
```

3b) Males only

```
YEAR_DICT_MALES={}

for Y in ALL_MALES["year"].unique():
    YEAR_DICT_MALES[Y]=ALL_MALES["suicides_100k_pop"][ALL_MALES["year"]==Y].values

TEST_STAT, PVALUE = stats.kruskal(*YEAR_DICT_MALES.values())
print("test statistic: %.4f"%TEST_STAT)
print("p-value: %.10f"%PVALUE)
```

3c) Females only

```
YEAR_DICT_FEMALES={}

for Y in ALL_FEMALES["year"].unique():
    YEAR_DICT_FEMALES[Y]=ALL_FEMALES["suicides_100k_pop"][ALL_FEMALES["year"]==Y].values

TEST_STAT, PVALUE = stats.kruskal(*YEAR_DICT_FEMALES.values())
print("test statistic: %.4f"%TEST_STAT)
print("p-value: %.10f"%PVALUE)
```

Analysis 4: Are distributions of suicide rates equal across different countries?

4a) Both sexes

```
COUNTRY_DICT={}

for C in MASTER["country"].unique():
    COUNTRY_DICT[C]=MASTER["suicides_100k_pop"][MASTER["country"]==C].values

TEST_STAT, PVALUE = stats.kruskal(*COUNTRY_DICT.values())
print("test statistic: %.4f"%TEST_STAT)
print("p-value: %.10f"%PVALUE)
```

4b) Males only

```
COUNTRY_DICT_MALES={}

for C in ALL_MALES["country"].unique():
    COUNTRY_DICT_MALES[C]=ALL_MALES["suicides_100k_pop"][ALL_MALES["country"]==C].values

TEST_STAT, PVALUE = stats.kruskal(*COUNTRY_DICT_MALES.values())
print("test statistic: %.4f"%TEST_STAT)
print("p-value: %.10f"%PVALUE)
```

4c) Females only

```
COUNTRY_DICT_FEMALES={}

for C in ALL_FEMALES["country"].unique():
    COUNTRY_DICT_FEMALES[C]=ALL_FEMALES["suicides_100k_pop"][ALL_FEMALES["country"]==C].val
ues

TEST_STAT, PVALUE = stats.kruskal(*COUNTRY_DICT_FEMALES.values())
print("test statistic: %.4f"%TEST_STAT)
print("p-value: %.10f"%PVALUE)
```

Regression Analysis of GDP vs. Suicide Rates

5a) Correlation split out by sex and age

```
#create an empty dataframe to store results
RESULTS = pd.DataFrame(columns=["sex","age","correlation","p-value"])

#iterate over combinations of sex and age, calculating correlation and p-value and store in
results dataframe
for s in MASTER.sex.unique():
    for a in MASTER.age.unique():
        WORKING_DATA=MASTER[(MASTER["sex"]==s)&(MASTER["age"]==a)]
        CORR,PVALUE = stats.spearmanr(WORKING_DATA.gdp_per_capita,WORKING_DATA.suicides_100
k_pop)

        RESULTS=RESULTS.append({"sex":s,"age":a,"correlation":CORR,"p-value":PVAL},ignore_i
ndex=True)

#export to csv to copying into report
RESULTS.to_csv("spearmanr_results.csv")

#display results
RESULTS
```

5b) Correlation split out by sex, age and year

```
RESULTS = pd.DataFrame(columns=["sex","age","year","correlation","p-value"])

for s in MASTER.sex.unique():
    for a in MASTER.age.unique():
        for y in MASTER.year.unique():
            WORKING_DATA=MASTER[(MASTER["sex"]==s)&(MASTER["age"]==a)&(MASTER["year"]==y)]
            CORR,PVALUE = stats.spearmanr(WORKING_DATA.gdp_per_capita,WORKING_DATA.suicides
_100k_pop)

            RESULTS=RESULTS.append({"sex":s,"age":a,"year":y,"correlation":CORR,"p-value":P
VAL},ignore_index=True)

RESULTS.to_csv("spearmanr_year_results.csv")

RESULTS

#check if any rows have p-values greater than or equal to the significance level
RESULTS[RESULTS["p-value"]>=0.05]

#create histograms of correlation values observed for each correlation analysis completed f
or sex, age and year.
FIG8 = px.histogram(RESULTS,x="correlation",color="sex",nbins=50,template="simple_white",
                    title="Histogram of Correlation rho for GDP vs. Suicides per 100k pop, s
plit by sex, age and year",
                    width=1080,height=760,log_y=False,facet_col="sex")
FIG8.update_layout(font=dict(family="Calibri",size=20),
                     yaxis_title="Observations [count]")
FIG8.show()

#mean correlation of all 384 observations
RESULTS["correlation"].mean()

#mean correlation for all males
RESULTS[RESULTS["sex"]=="male"].correlation.mean()

#mean correlation for all females
RESULTS[RESULTS["sex"]=="female"].correlation.mean()

RESULTS[RESULTS["sex"]=="male"].nlargest(5,columns="correlation")

RESULTS[RESULTS["sex"]=="female"].nlargest(5,columns="correlation")
```

```
#create a heatmap for the correlation values at all years and ages for males
FIG9 = go.Figure(data=go.Heatmap(z=RESULTS[RESULTS["sex"]=="male"].correlation,
                                 x=RESULTS[RESULTS["sex"]=="male"].year,
                                 y=RESULTS[RESULTS["sex"]=="male"].age))

FIG9.update_layout(autosize=False,width=1080,height=760,
                   title="Heatmap of rho vlaues for Suicides per 100k pop vs. GDP per ca
pita: males",
                   title_xanchor="center",title_x=0.5,
                   xaxis_title="Year [-]",
                   yaxis_title="Age [-]",
                   font_size=18,
                   legend=dict(x=0.65,y=0.995,bordercolor="#C4C3D0",borderwidth=2))

FIG9.show()

#create a heatmap for the correlation values at all years and ages for females
FIG10 = go.Figure(data=go.Heatmap(z=RESULTS[RESULTS["sex"]=="female"].correlation,
                                 x=RESULTS[RESULTS["sex"]=="female"].year,
                                 y=RESULTS[RESULTS["sex"]=="female"].age))

FIG10.update_layout(autosize=False,width=1080,height=760,
                   title="Heatmap of rho vlaues for Suicides per 100k pop vs. GDP per ca
pita: females",
                   title_xanchor="center",title_x=0.5,
                   xaxis_title="Year [-]",
                   yaxis_title="Age [-]",
                   font_size=18,
                   legend=dict(x=0.65,y=0.995,bordercolor="#C4C3D0",borderwidth=2))

FIG10.show()
```

6.2 Tables

sex	age	stat	p-value
female	5-14_years	0.50294	0.000
female	15-24_years	0.75009	0.000
female	25-34_years	0.88212	0.000
female	35-54_years	0.92366	0.000
female	55-74_years	0.89278	0.000
female	75+_years	0.74163	0.000
male	5-14_years	0.64343	0.000
male	15-24_years	0.89351	0.000
male	25-34_years	0.86907	0.000
male	35-54_years	0.84056	0.000
male	55-74_years	0.89119	0.000
male	75+_years	0.89020	0.000

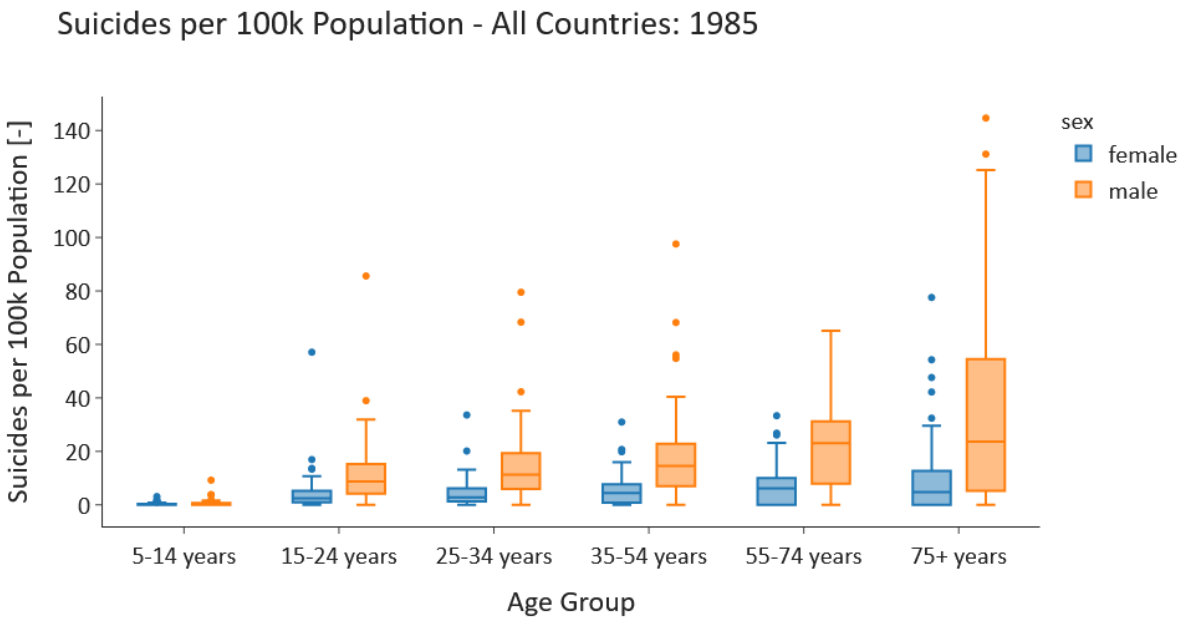
Table 4: Kolmogorov-Smirnov test vs. normal distribution for different age groups, split by sex.

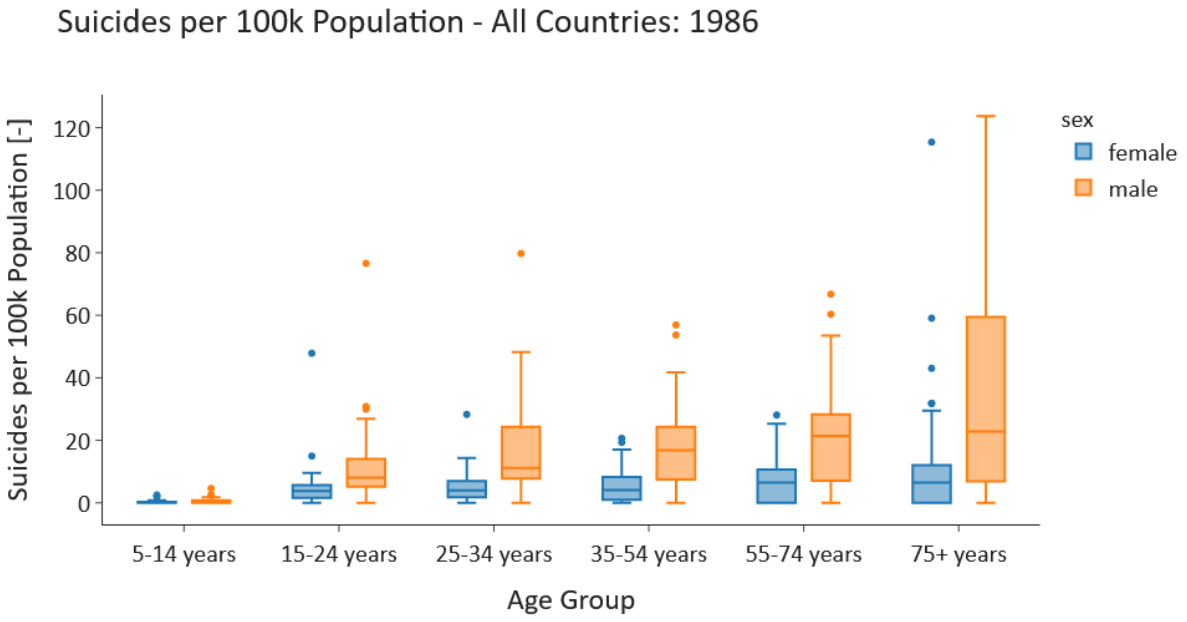
sex	year	stat	p-value
female	1985	0.61259	0.000
female	1986	0.53691	0.000
female	1987	0.67511	0.000
female	1988	0.63948	0.000
female	1989	0.62869	0.000
female	1990	0.72309	0.000
female	1991	0.70387	0.000
female	1992	0.72641	0.000
female	1993	0.72237	0.000
female	1994	0.69402	0.000
female	1995	0.73539	0.000
female	1996	0.7455	0.000
female	1997	0.76659	0.000
female	1998	0.75486	0.000
female	1999	0.75646	0.000
female	2000	0.57617	0.000
female	2001	0.64296	0.000
female	2002	0.75923	0.000
female	2003	0.72669	0.000
female	2004	0.70418	0.000
female	2005	0.70562	0.000
female	2006	0.69968	0.000
female	2007	0.70287	0.000
female	2008	0.6978	0.000
female	2009	0.68054	0.000
female	2010	0.64429	0.000
female	2011	0.70877	0.000
female	2012	0.73561	0.000
female	2013	0.7381	0.000
female	2014	0.8046	0.000
female	2015	0.80405	0.000
female	2016	0.88504	0.000

male	1985	0.72317	0.000
male	1986	0.75206	0.000
male	1987	0.78133	0.000
male	1988	0.80779	0.000
male	1989	0.81297	0.000
male	1990	0.80927	0.000
male	1991	0.79614	0.000
male	1992	0.78958	0.000
male	1993	0.78541	0.000
male	1994	0.78133	0.000
male	1995	0.76302	0.000
male	1996	0.79213	0.000
male	1997	0.78973	0.000
male	1998	0.80473	0.000
male	1999	0.80216	0.000
male	2000	0.78942	0.000
male	2001	0.80321	0.000
male	2002	0.79141	0.000
male	2003	0.79018	0.000
male	2004	0.7897	0.000
male	2005	0.78533	0.000
male	2006	0.76098	0.000
male	2007	0.78312	0.000
male	2008	0.80124	0.000
male	2009	0.79077	0.000
male	2010	0.78438	0.000
male	2011	0.78318	0.000
male	2012	0.76765	0.000
male	2013	0.81118	0.000
male	2014	0.8286	0.000
male	2015	0.80605	0.000
male	2016	0.81966	0.000

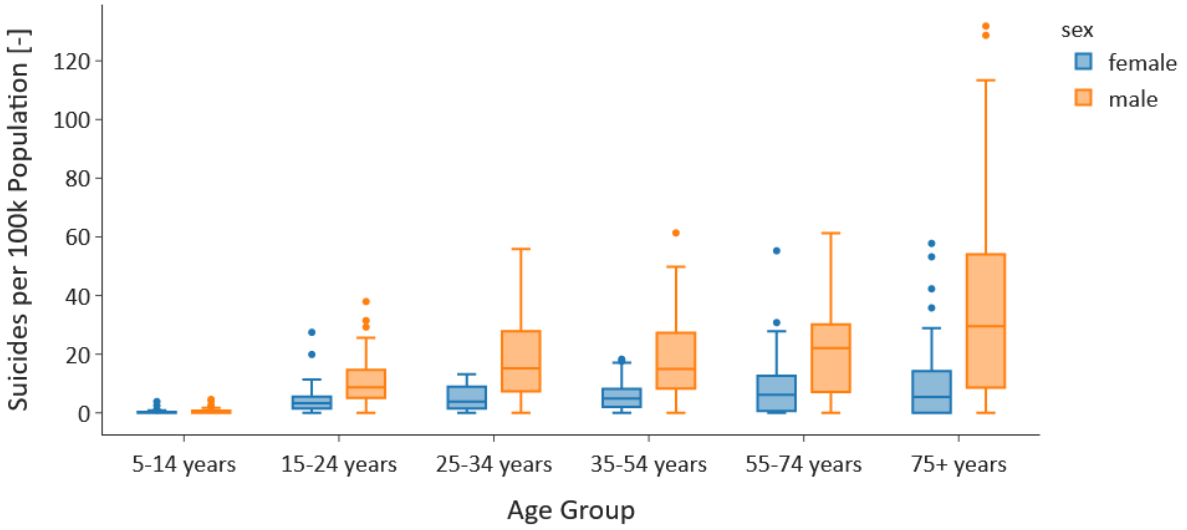
Table 5: Kolmogorov-Smirnov test vs. normal distribution for different years, split by sex.

6.3 Additional Plots

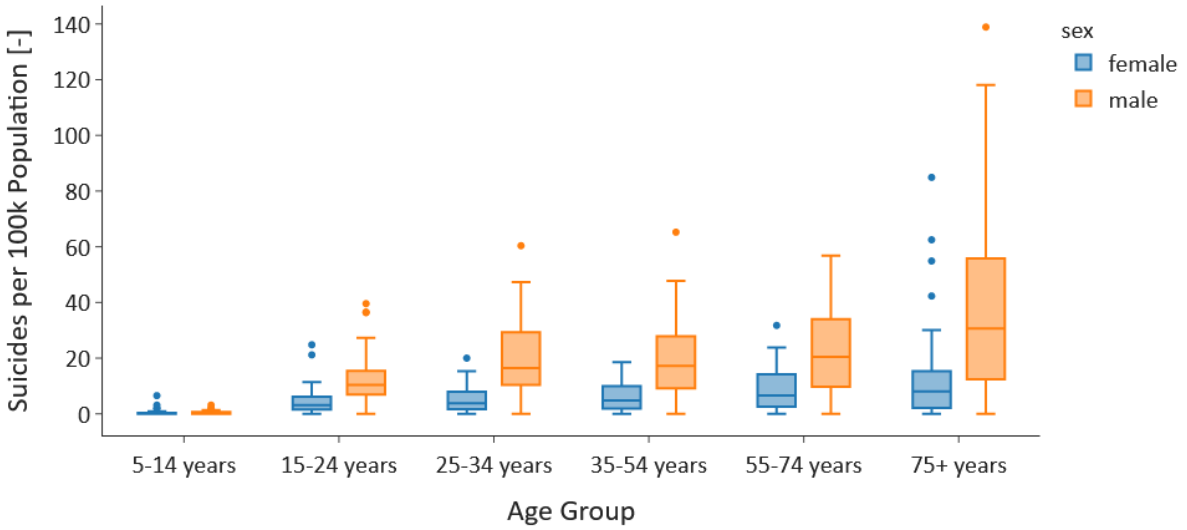




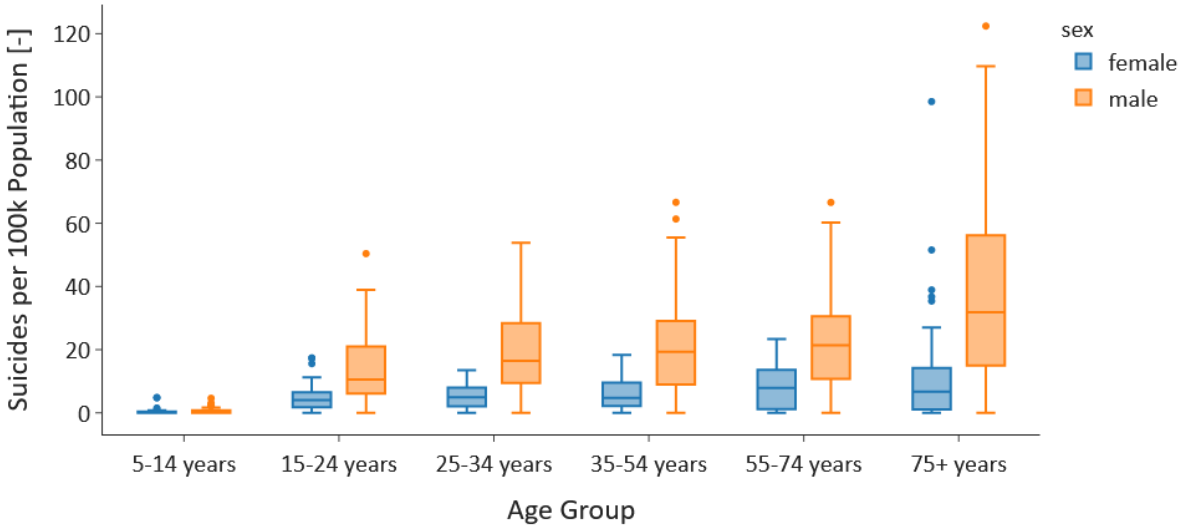
Suicides per 100k Population - All Countries: 1987



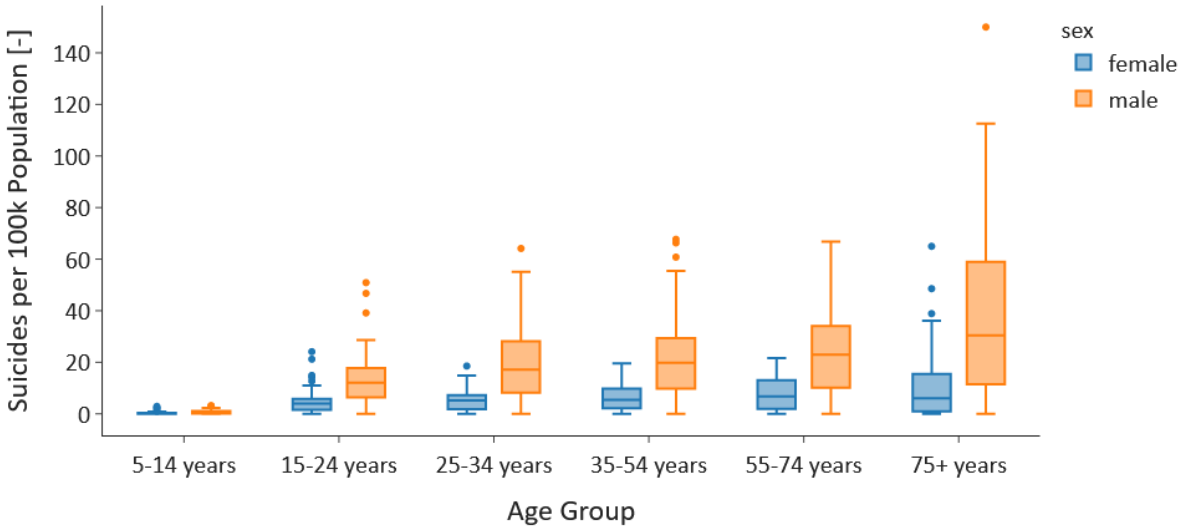
Suicides per 100k Population - All Countries: 1988



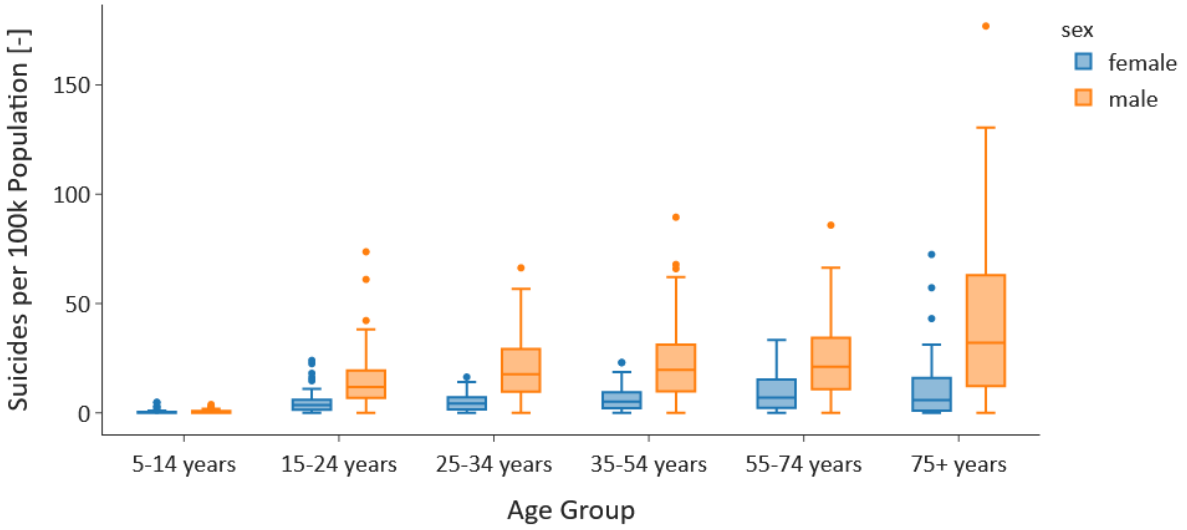
Suicides per 100k Population - All Countries: 1989



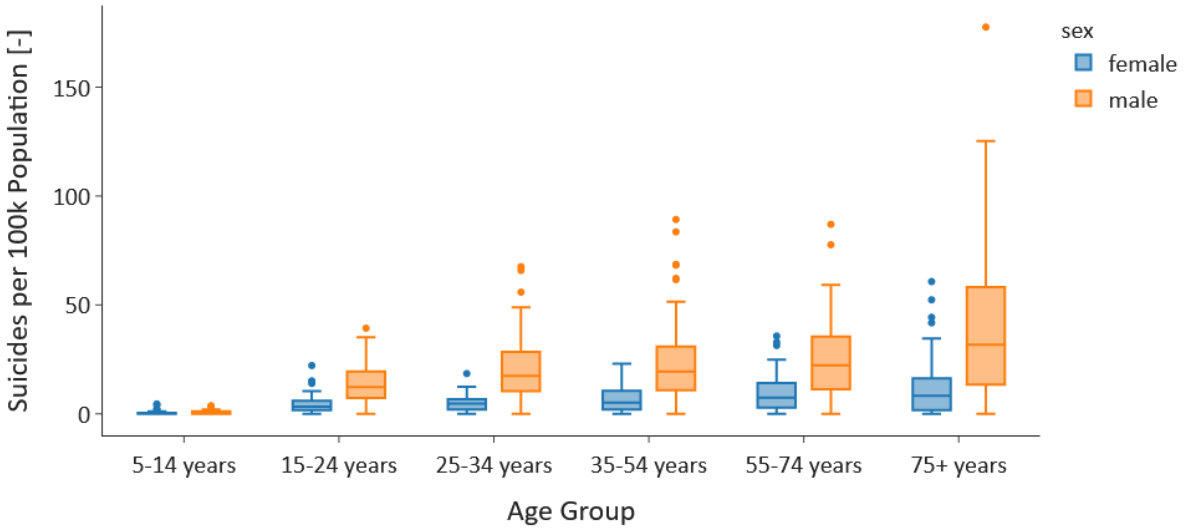
Suicides per 100k Population - All Countries: 1990



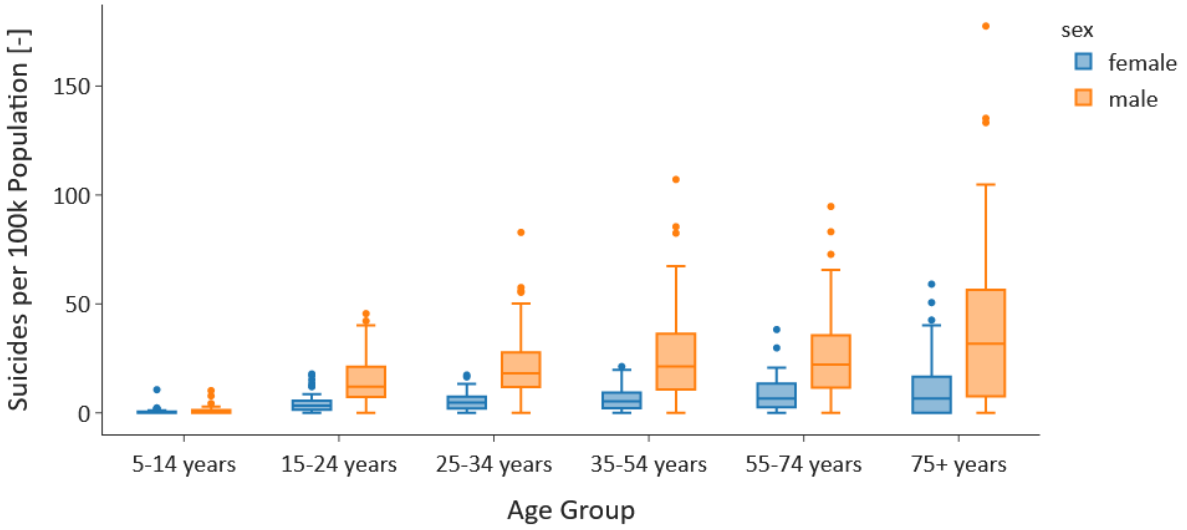
Suicides per 100k Population - All Countries: 1991



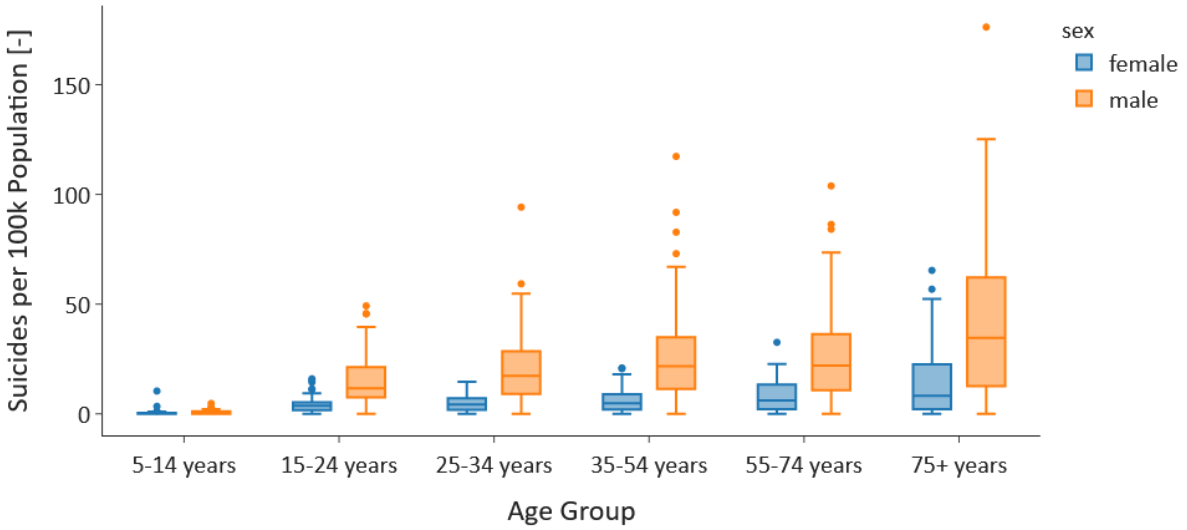
Suicides per 100k Population - All Countries: 1992



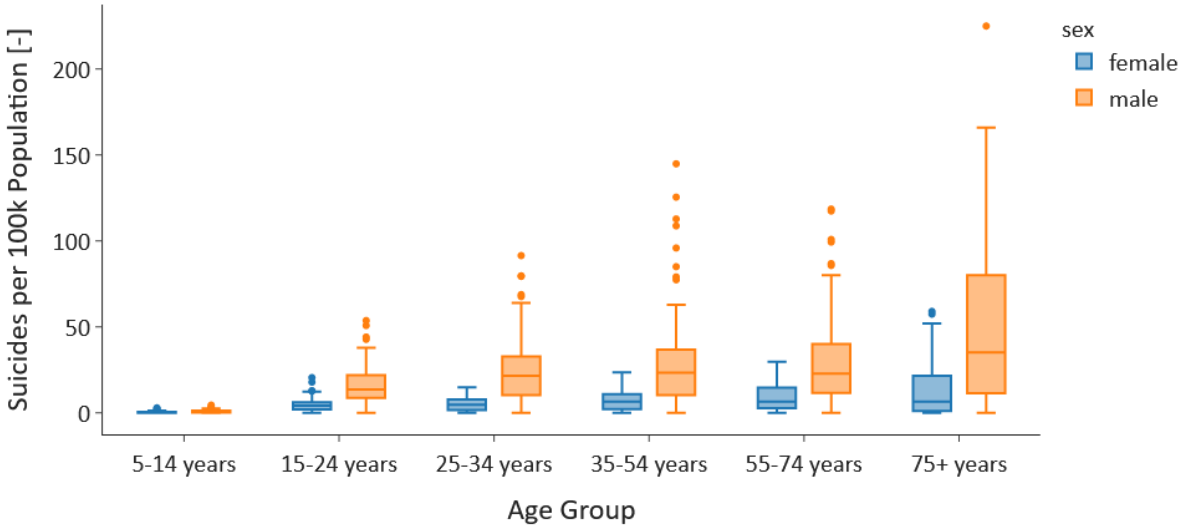
Suicides per 100k Population - All Countries: 1993



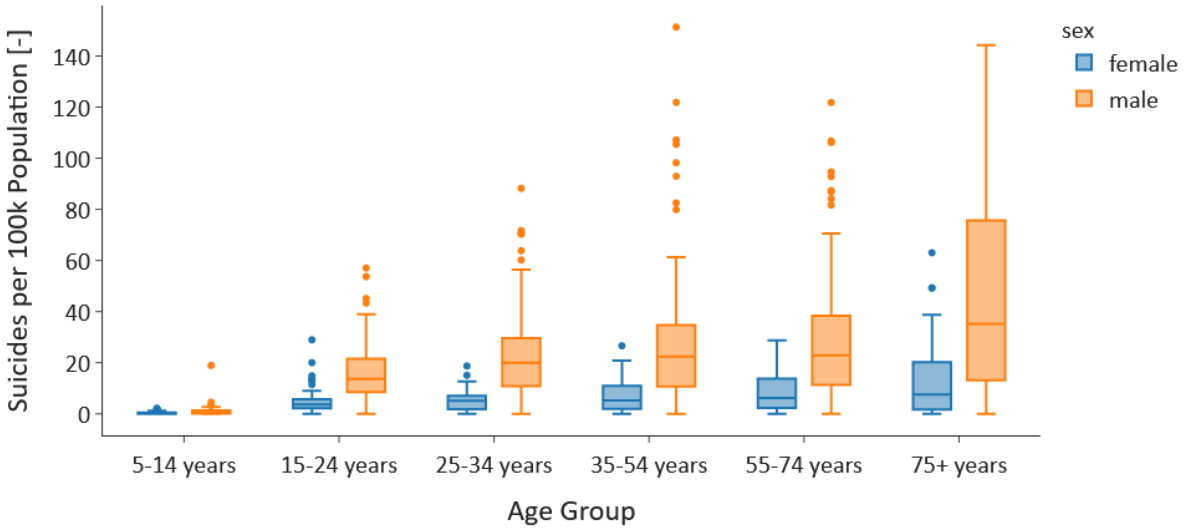
Suicides per 100k Population - All Countries: 1994



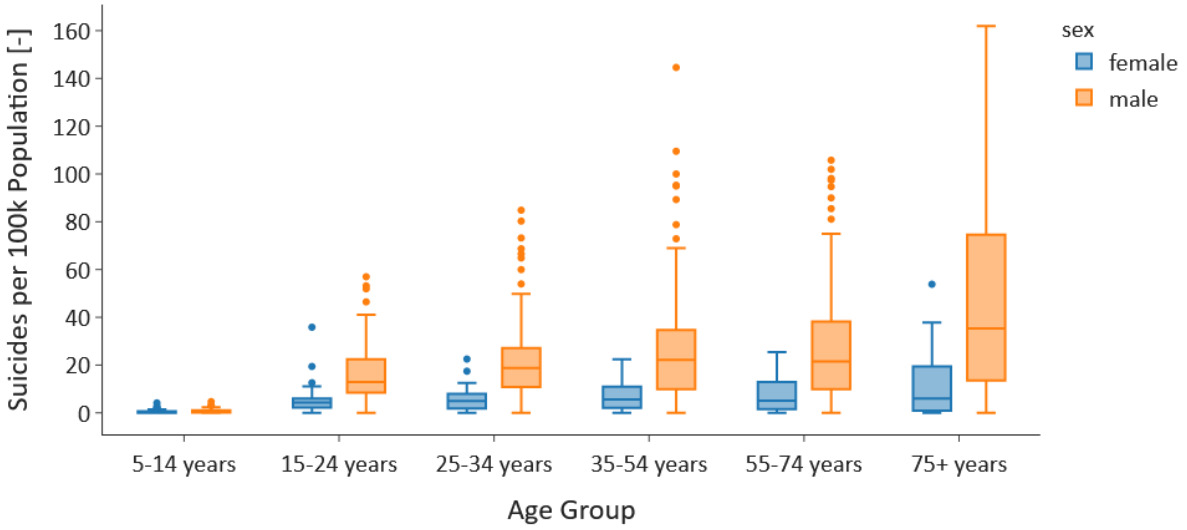
Suicides per 100k Population - All Countries: 1995



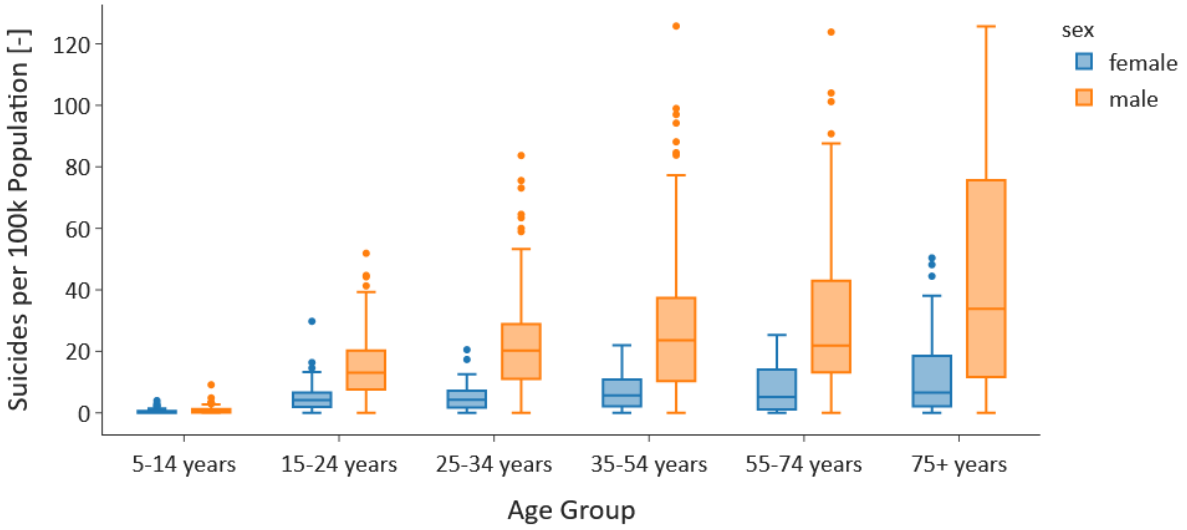
Suicides per 100k Population - All Countries: 1996



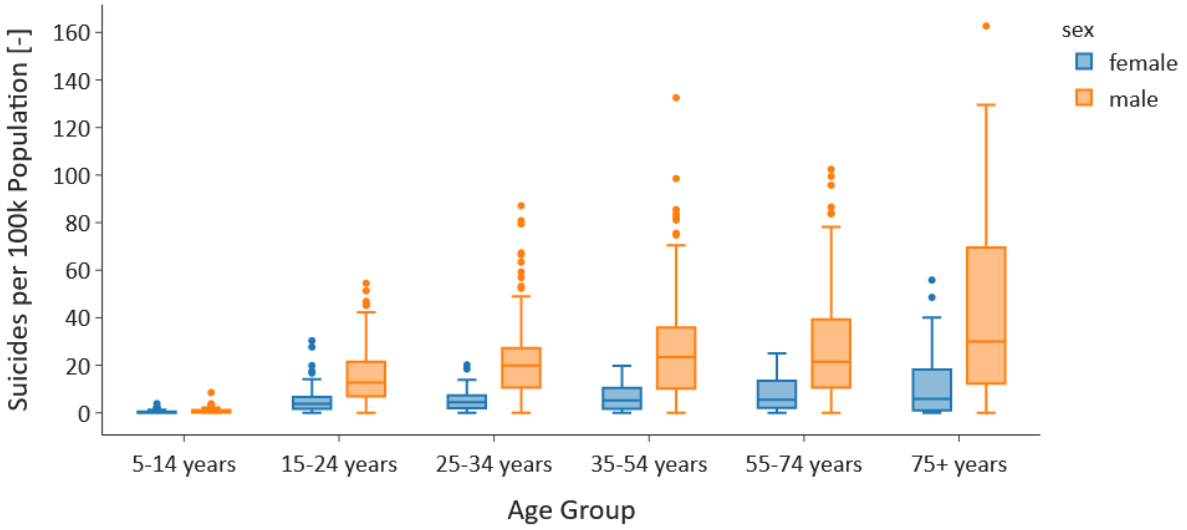
Suicides per 100k Population - All Countries: 1997



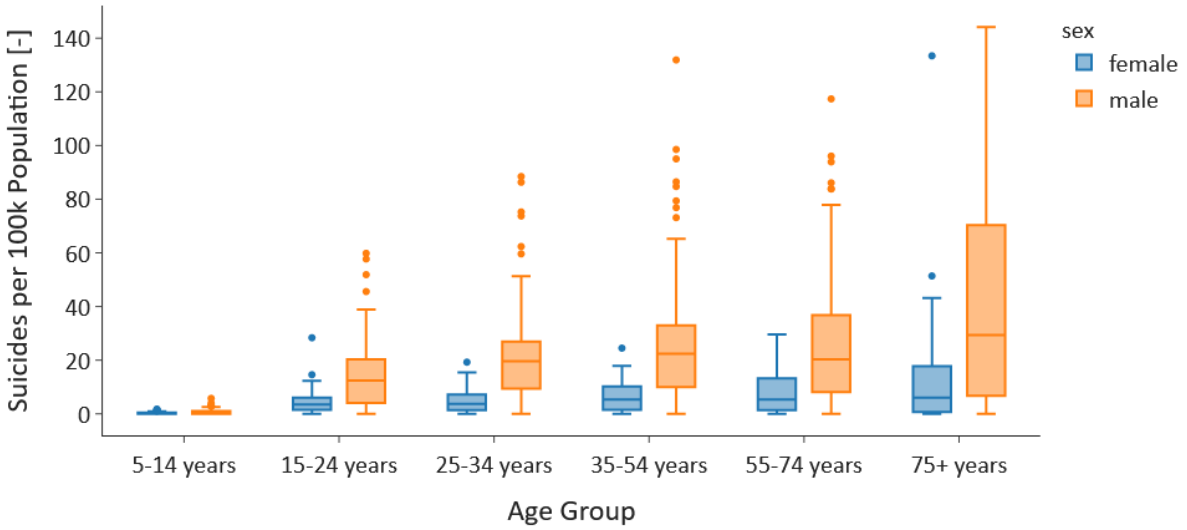
Suicides per 100k Population - All Countries: 1998

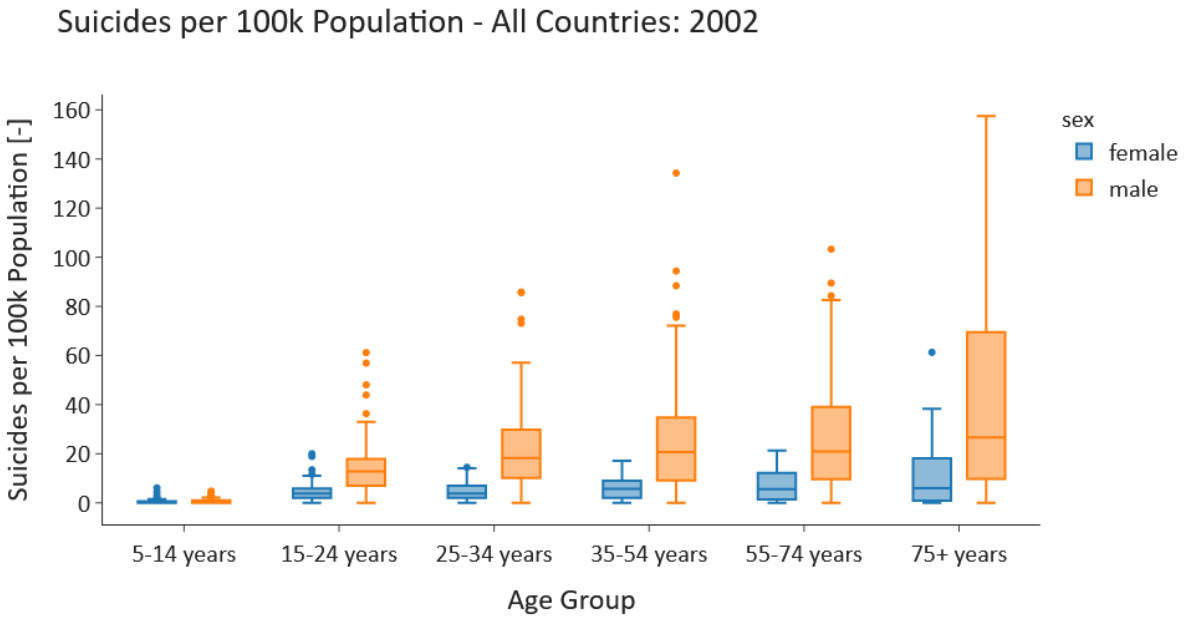


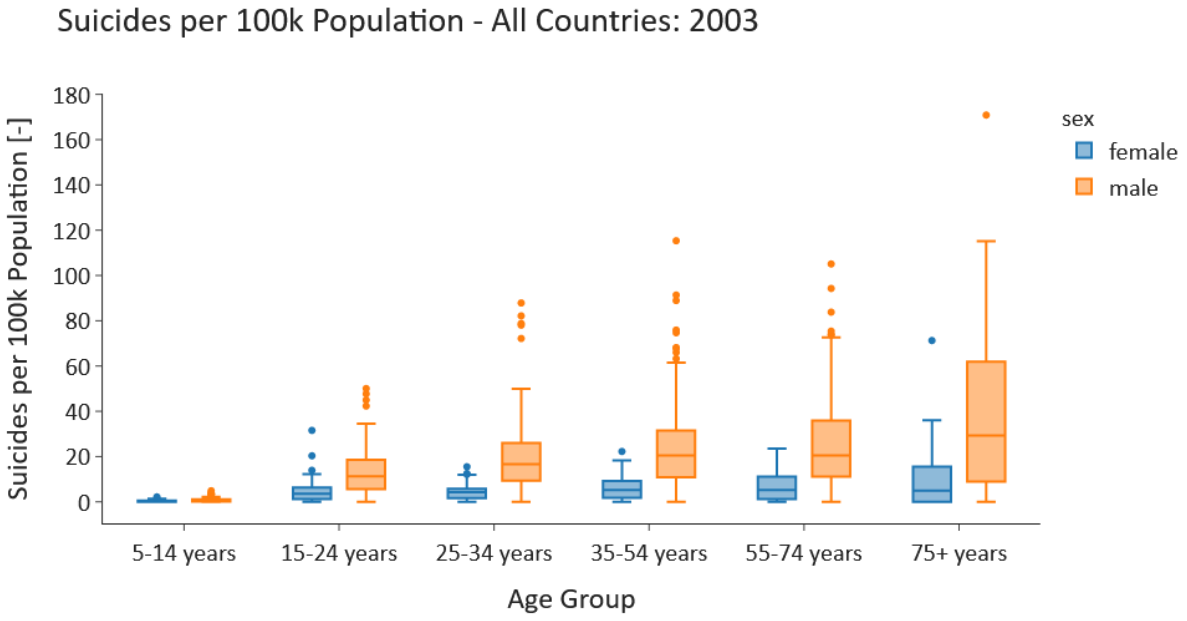
Suicides per 100k Population - All Countries: 1999

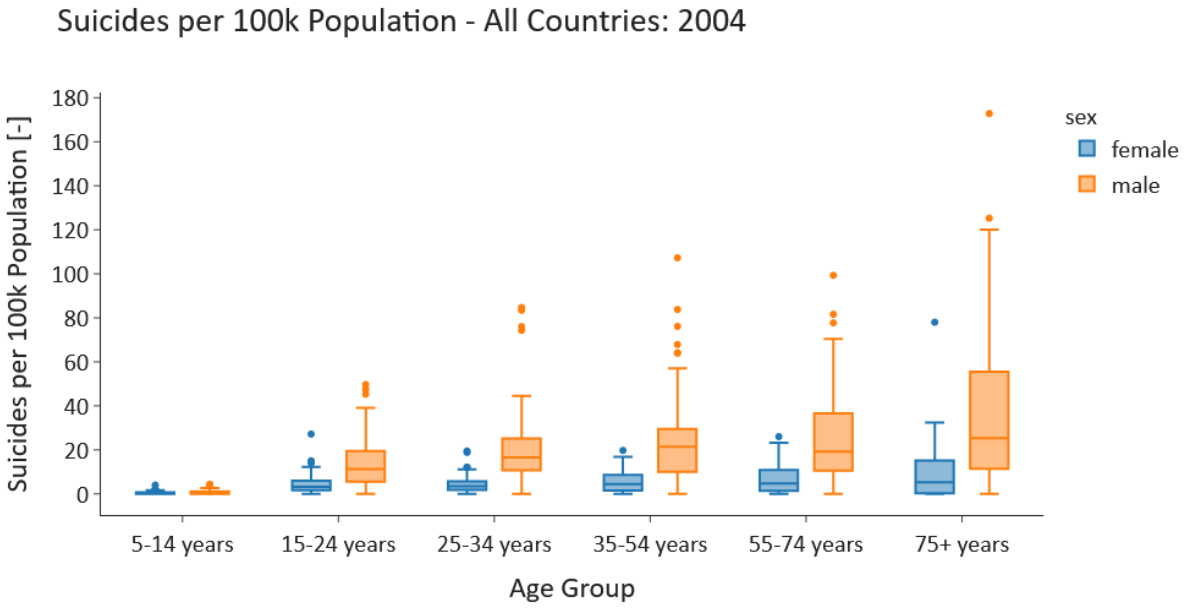


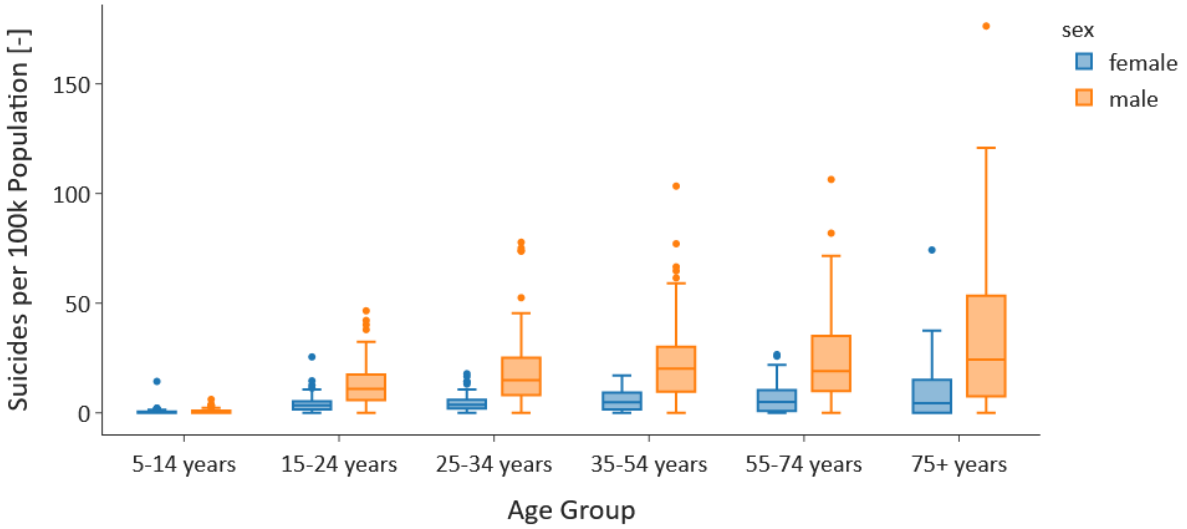
Suicides per 100k Population - All Countries: 2000

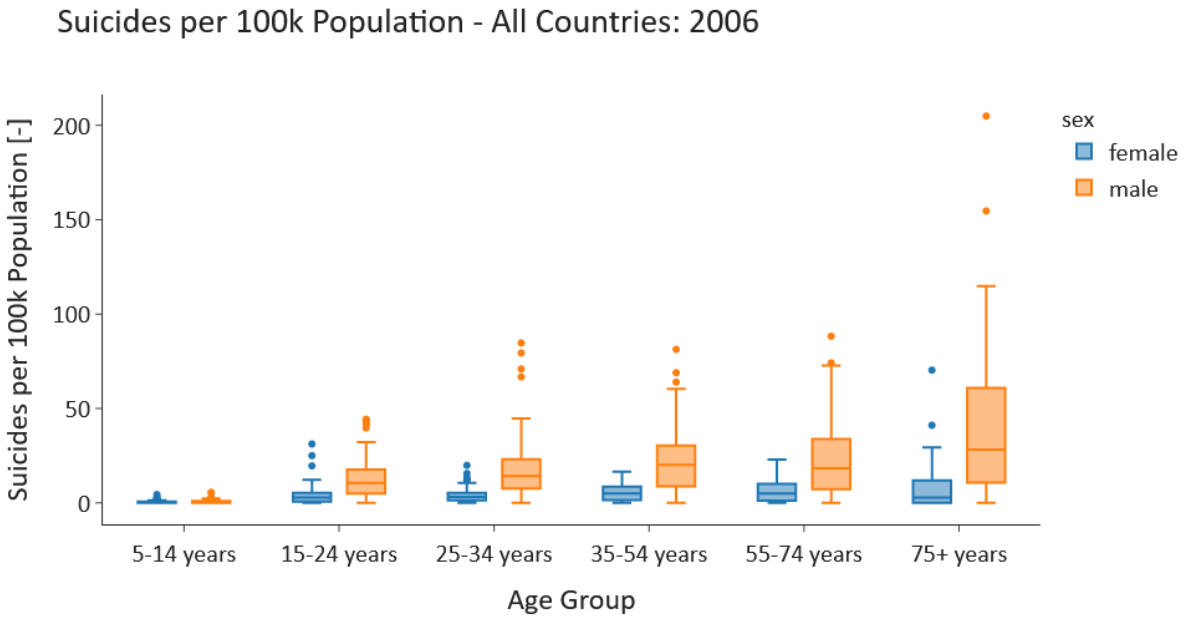


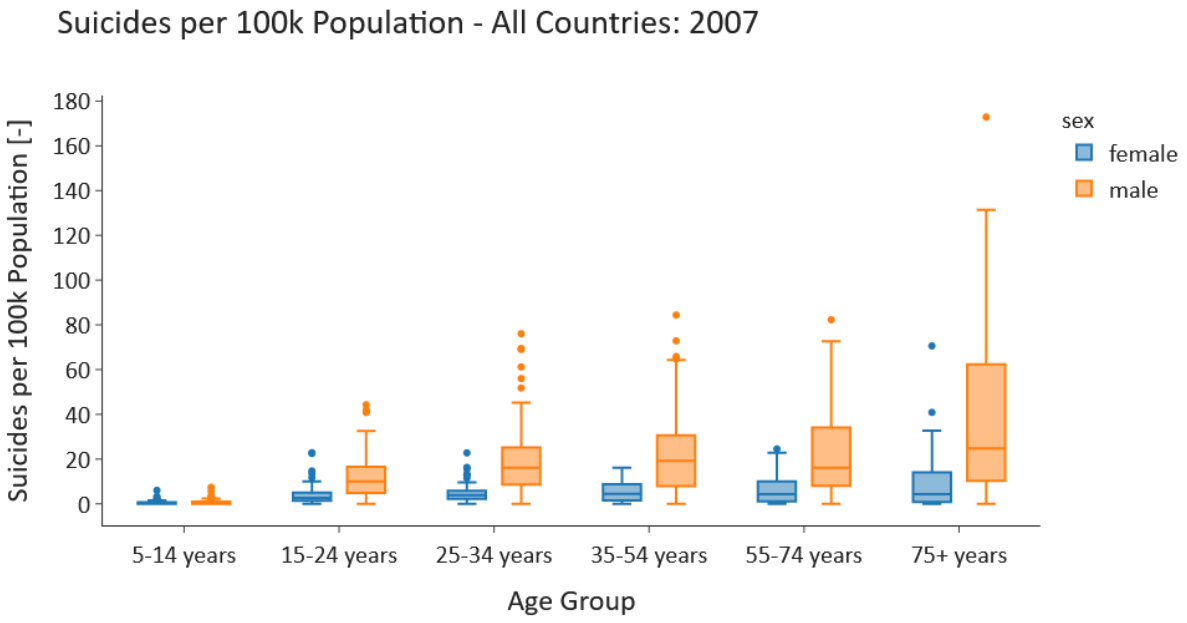

Suicides per 100k Population - All Countries: 2002
Suicides per 100k Population [-]
160
140
120
100
80
60
40
20
0
5-14 years
15-24 years
25-34 years
35-54 years
55-74 years
75+ years
Age Group
sex
female
male


Suicides per 100k Population - All Countries: 2003
Suicides per 100k Population [-]
180
160
140
120
100
80
60
40
20
0
5-14 years
15-24 years
25-34 years
35-54 years
55-74 years
75+ years
Age Group
sex
female
male

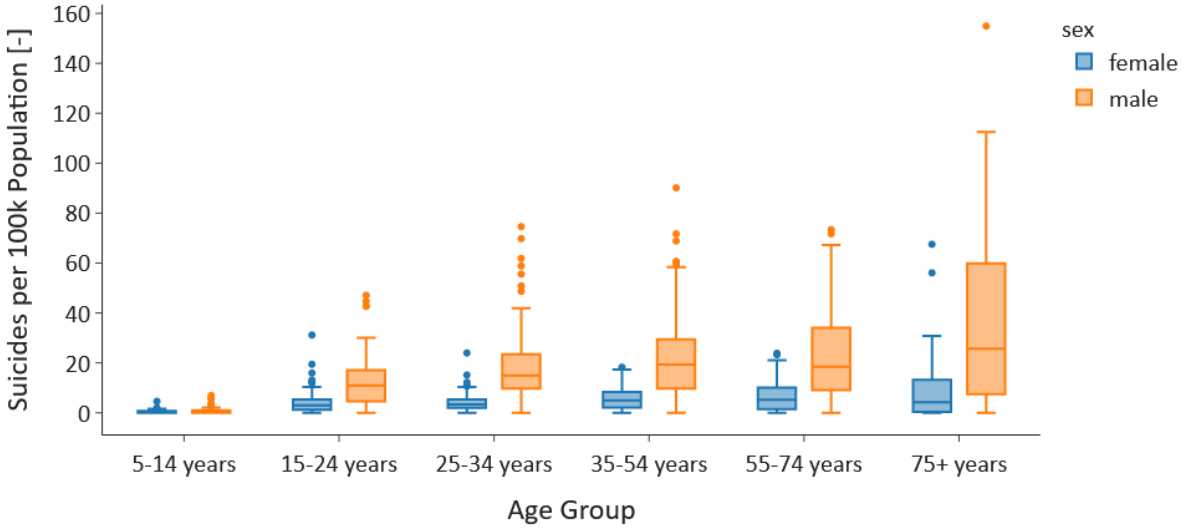

Suicides per 100k Population - All Countries: 2004
Suicides per 100k Population [-]
180
160
140
120
100
80
60
40
20
0
5-14 years
15-24 years
25-34 years
35-54 years
55-74 years
75+ years
Age Group
sex
female
male


Suicides per 100k Population - All Countries: 2005
Suicides per 100k Population [-]
150
100
50
0
5-14 years
15-24 years
25-34 years
35-54 years
55-74 years
75+ years
Age Group
sex
female
male

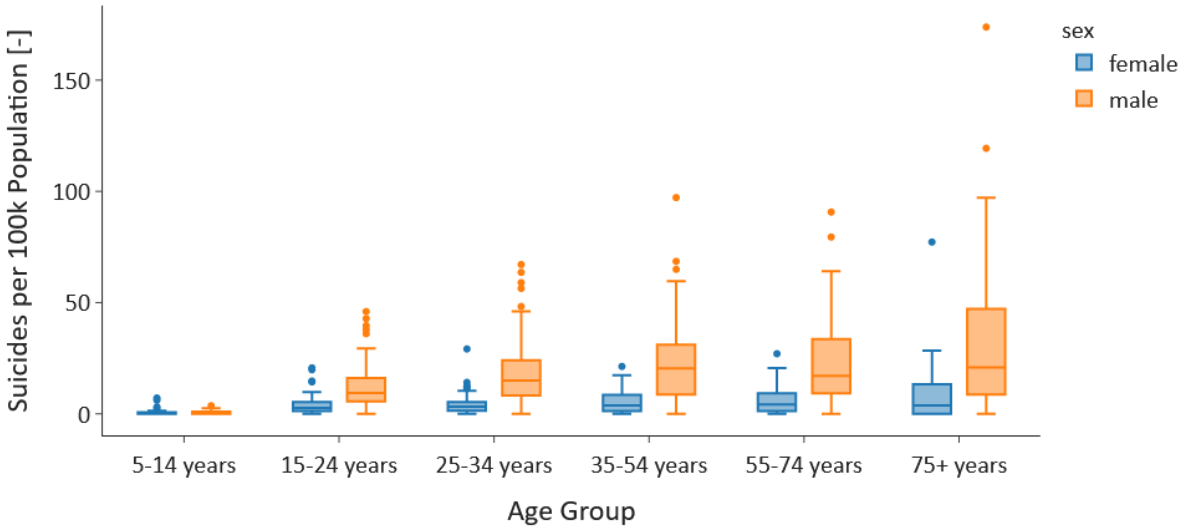

Suicides per 100k Population - All Countries: 2006
Suicides per 100k Population [-]
200
150
100
50
0
5-14 years
15-24 years
25-34 years
35-54 years
55-74 years
75+ years
Age Group
sex
female
male


Suicides per 100k Population - All Countries: 2007
Suicides per 100k Population [-]
180
160
140
120
100
80
60
40
20
0
5-14 years
15-24 years
25-34 years
35-54 years
55-74 years
75+ years
Age Group
sex
female
male

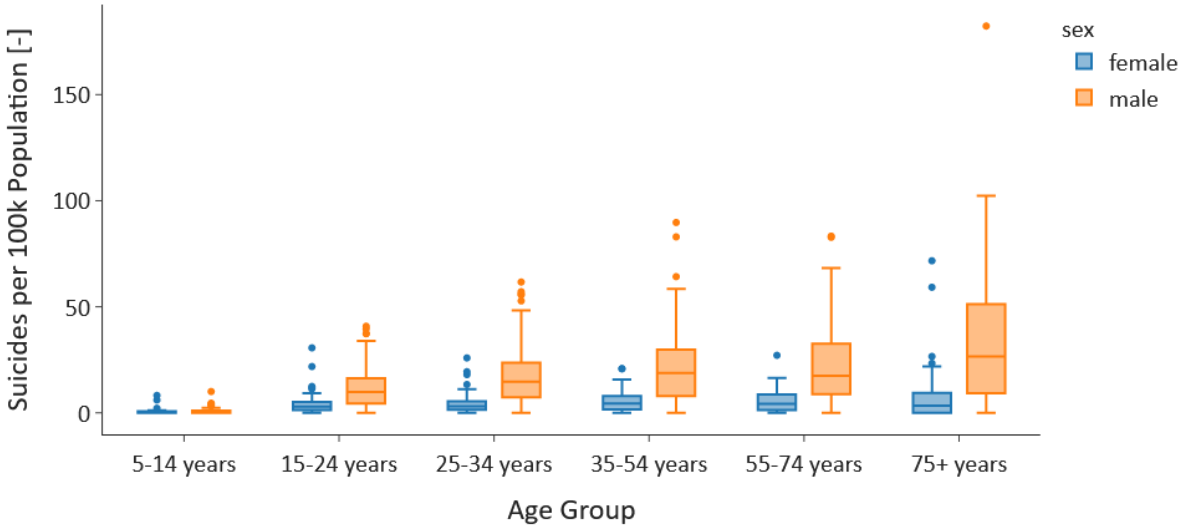
Suicides per 100k Population - All Countries: 2008



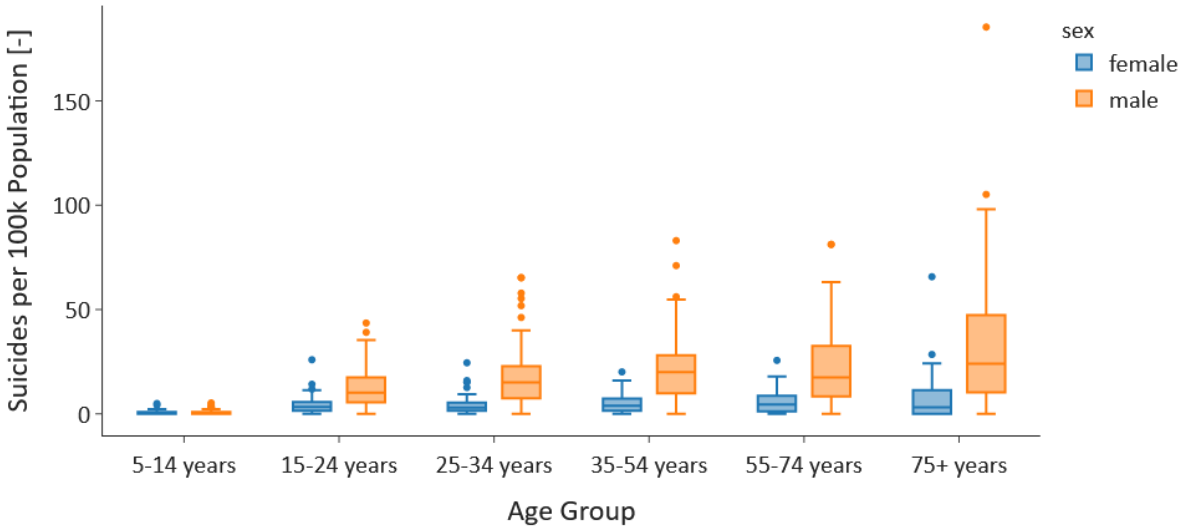
Suicides per 100k Population - All Countries: 2009

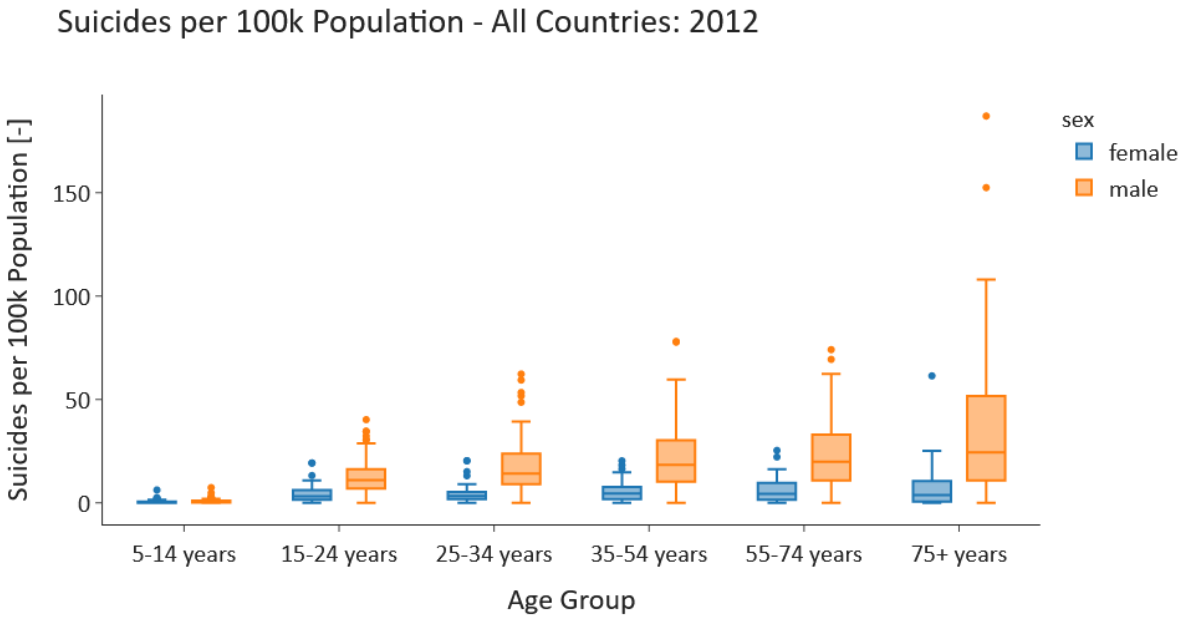


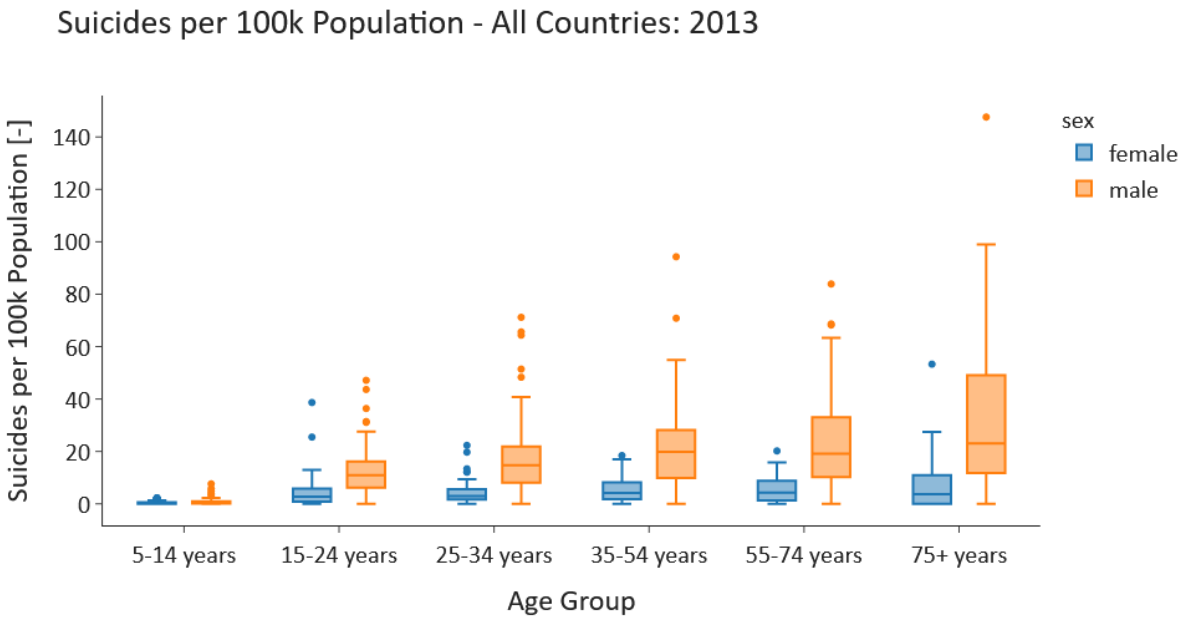
Suicides per 100k Population - All Countries: 2010



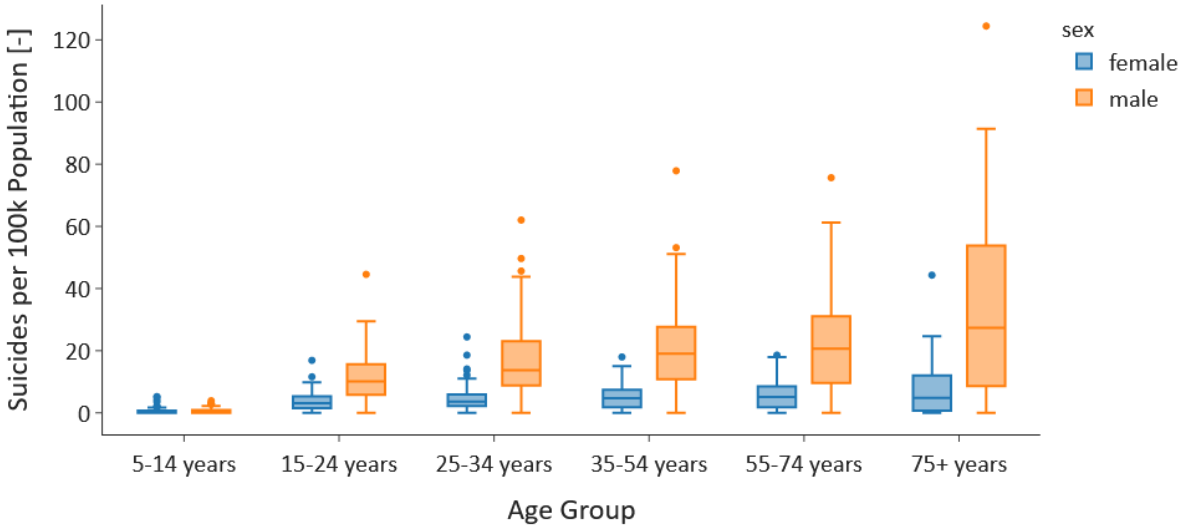
Suicides per 100k Population - All Countries: 2011



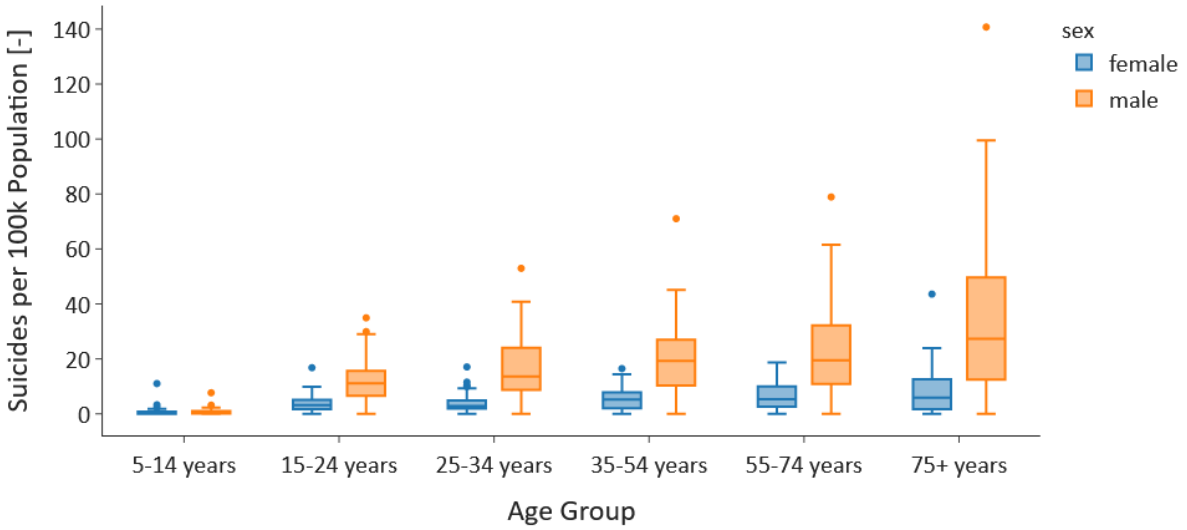

Suicides per 100k Population - All Countries: 2012
Suicides per 100k Population [-]
150
100
50
0
5-14 years
15-24 years
25-34 years
35-54 years
55-74 years
75+ years
Age Group
sex
female
male

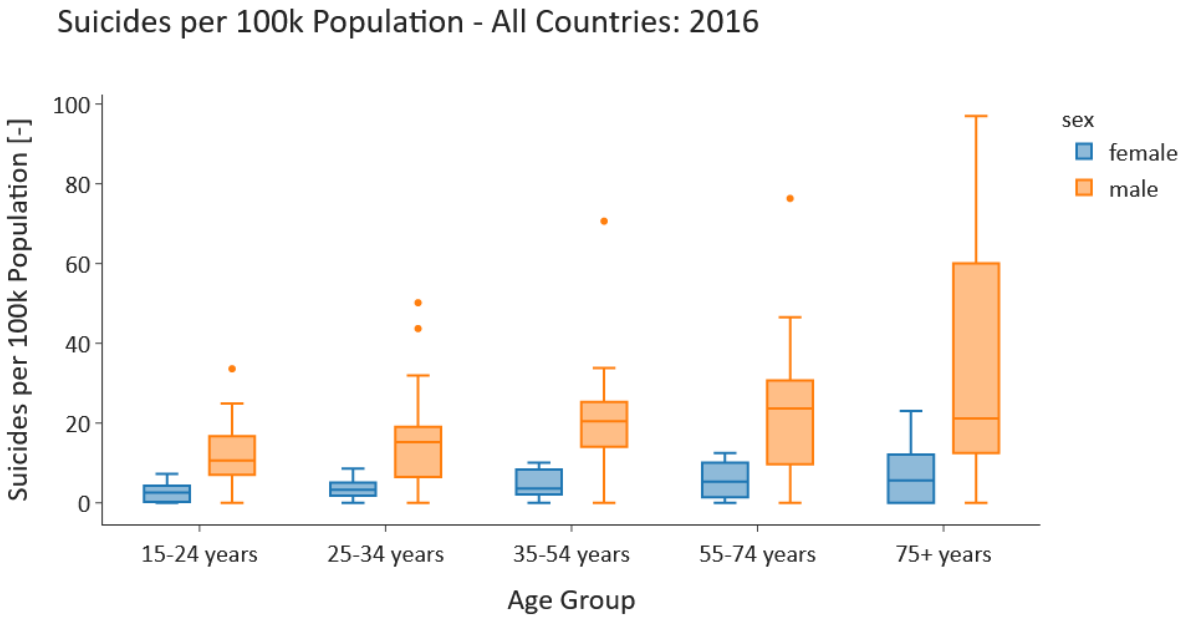

Suicides per 100k Population - All Countries: 2013
Suicides per 100k Population [-]
140
120
100
80
60
40
20
0
5-14 years
15-24 years
25-34 years
35-54 years
55-74 years
75+ years
Age Group
sex
female
male

Suicides per 100k Population - All Countries: 2014



Suicides per 100k Population - All Countries: 2015




Suicides per 100k Population - All Countries: 2016
Suicides per 100k Population [-]
100
80
60
40
20
0
15-24 years
25-34 years
35-54 years
55-74 years
75+ years
Age Group
sex
female
male