WHO Suicide Statistics Final

Jaquelin Martinez

December 17, 2018

Data

The dataset used for this analysis is merged data from WHO suicide statistics (https://www.kaggle.com/szamil/who-suicide-statistics) and World Bank data that was accessed through R by a package ("WDI"). The WHO suicide statistics had data on raw number of suicides and population, divided into age groups and gender, for certain countries for 1987-2016. The WDI provided data on GDP per capita, PPP (constant 2011 international $), income level, and region for countries in 1992-2016. The final merged dataset contains only countries and year common to both dataset.

My main question of exploration for merging the datasets is to see if there is a relationship between GDP per capita (PPP) and suicide rates (# suicides per 100,000 people in the population). I created the suicide rate to make the numbers comparable across age groups and countries with less population.

I would also like to explore differeces in suicide rates between man and women and through the years in our dataset.

```
source("final_processing.R")

source("final_exploring.R")
source("final_aggregate.R")
source("final_regression.R")

df <- merged_who_wdi()
```

What are the countries with the highest suicide rate (or the highest raw number of suicides) for each year in our dataset?

```
source("final_aggregate.R")

top_rate_year(df)
```

```
##    year        top_country
## 1  1990            Finland
## 2  1991            Hungary
## 3  1992            Hungary
## 4  1993 Russian Federation
## 5  1994 Russian Federation
## 6  1995          Lithuania
## 7  1996          Lithuania
## 8  1997          Lithuania
## 9  1998          Lithuania
## 10 1999          Lithuania
## 11 2000          Lithuania
## 12 2001          Lithuania
## 13 2002          Lithuania
## 14 2003          Lithuania
## 15 2004          Lithuania
## 16 2005          Lithuania
## 17 2006          Lithuania
## 18 2007          Lithuania
## 19 2008          Lithuania
## 20 2009          Lithuania
## 21 2010          Lithuania
## 22 2011          Lithuania
## 23 2012          Lithuania
## 24 2013          Lithuania
## 25 2014          Lithuania
## 26 2015          Lithuania
## 27 2016          Lithuania
```

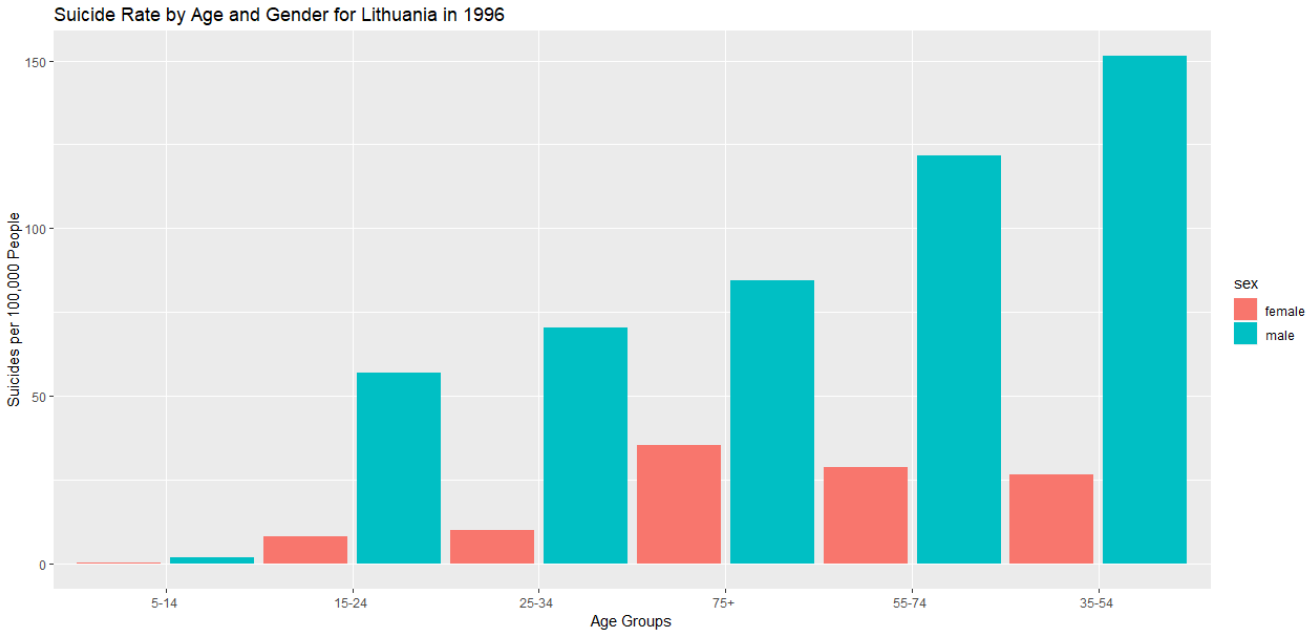
```
top_raw_year(df)
```

```
##    year              top_country
## 1  1990       Russian Federation
## 2  1991       Russian Federation
## 3  1992       Russian Federation
## 4  1993       Russian Federation
## 5  1994       Russian Federation
## 6  1995       Russian Federation
## 7  1996       Russian Federation
## 8  1997       Russian Federation
## 9  1998       Russian Federation
## 10 1999       Russian Federation
## 11 2000       Russian Federation
## 12 2001       Russian Federation
```

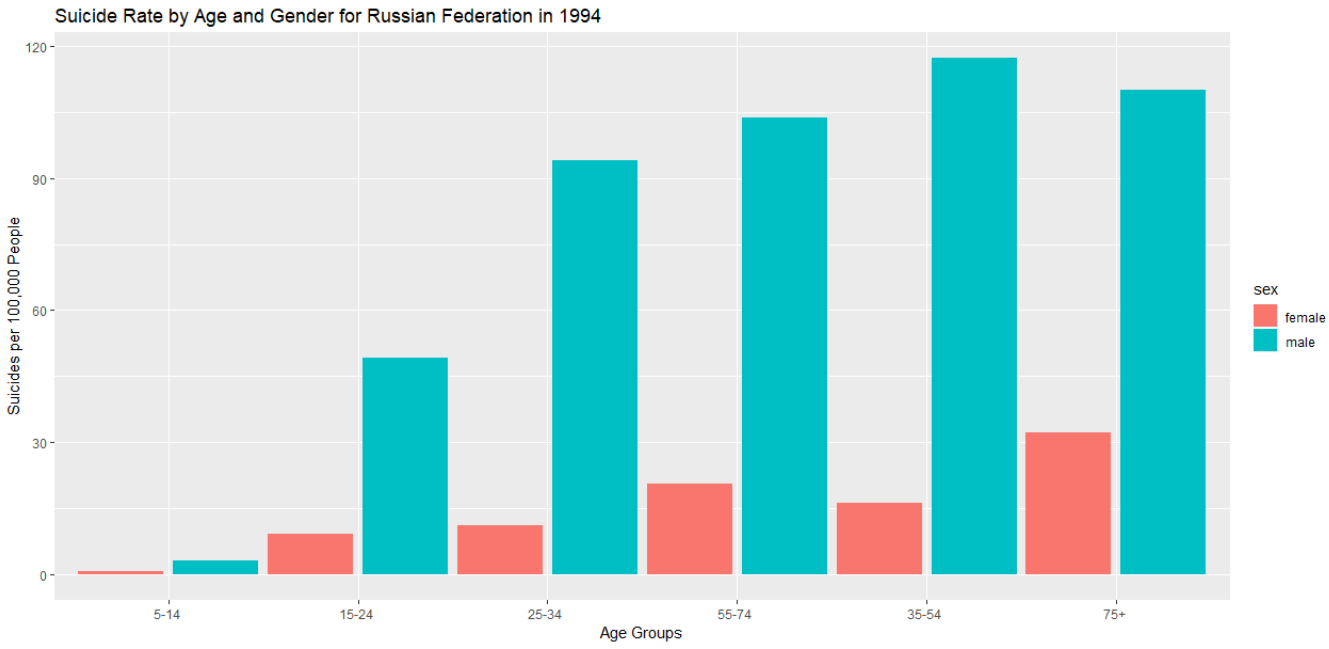
```
## 13 2002       Russian Federation
## 14 2003       Russian Federation
## 15 2004       Russian Federation
## 16 2005       Russian Federation
## 17 2006       Russian Federation
## 18 2007       Russian Federation
## 19 2008       Russian Federation
## 20 2009       Russian Federation
## 21 2010 United States of America
## 22 2011 United States of America
## 23 2012 United States of America
## 24 2013 United States of America
## 25 2014 United States of America
## 26 2015 United States of America
## 27 2016                 Thailand
```

Given these results, some countries of interest are: Findland, Hungary, Russian Federation, Lithuania, United States, Thailand.

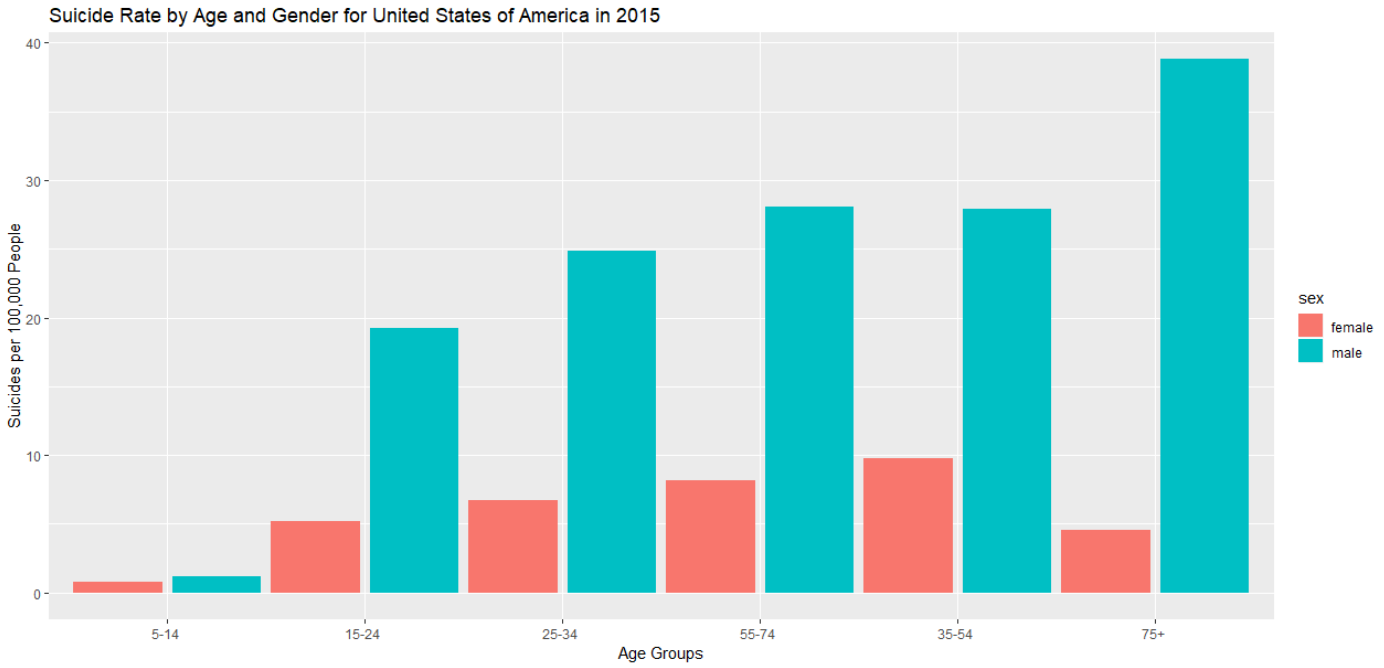
```
graph_by_country(df, "Lithuania", 1996)
```



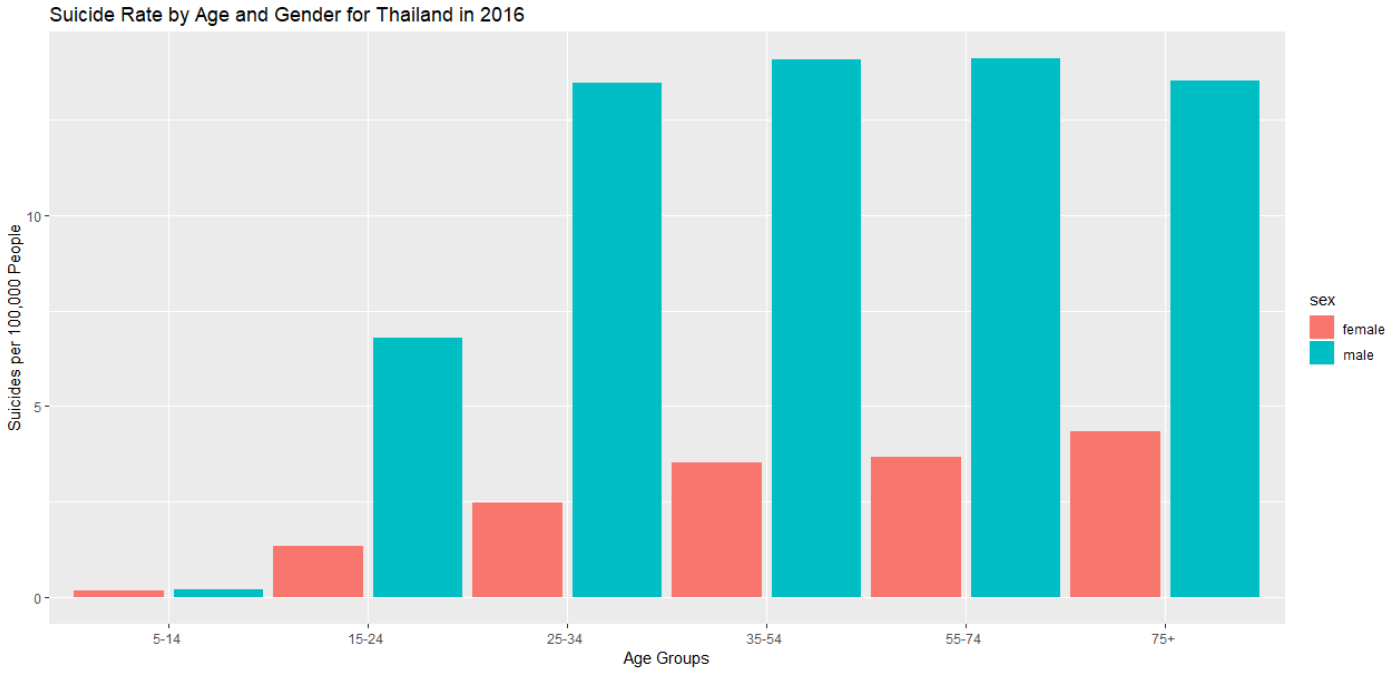

```
graph_by_country(df, "Russian Federation", 1994)
```



```
graph_by_country(df, "United States of America", 2015)
```

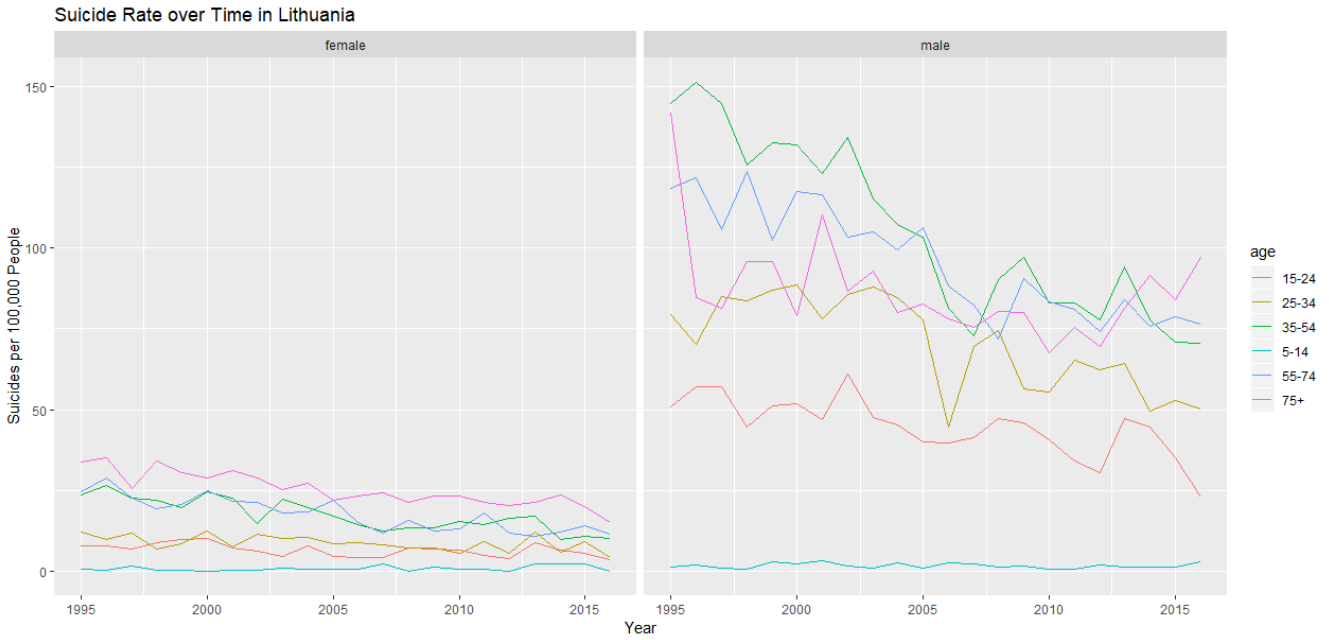



```
graph_by_country(df, "Thailand", 2016)
```

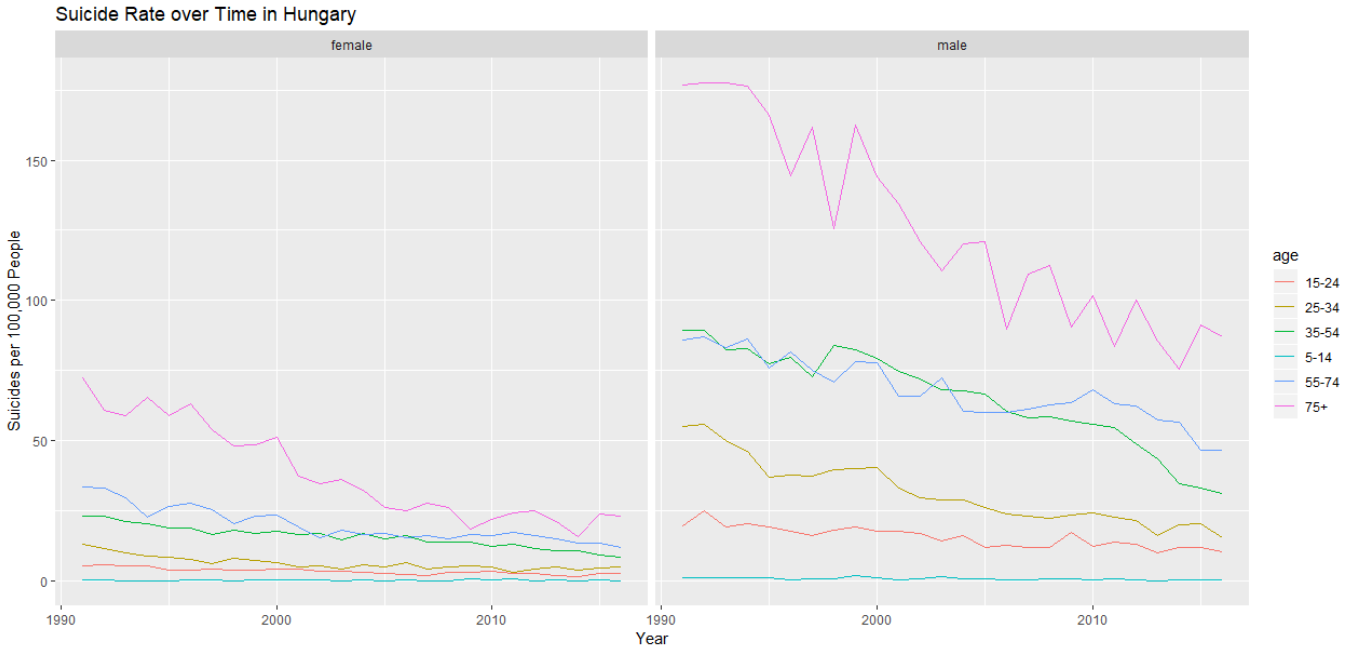


These graphs show the great disparities in suicide rates between men and women.

```
time_trend(df, "Lithuania")
```

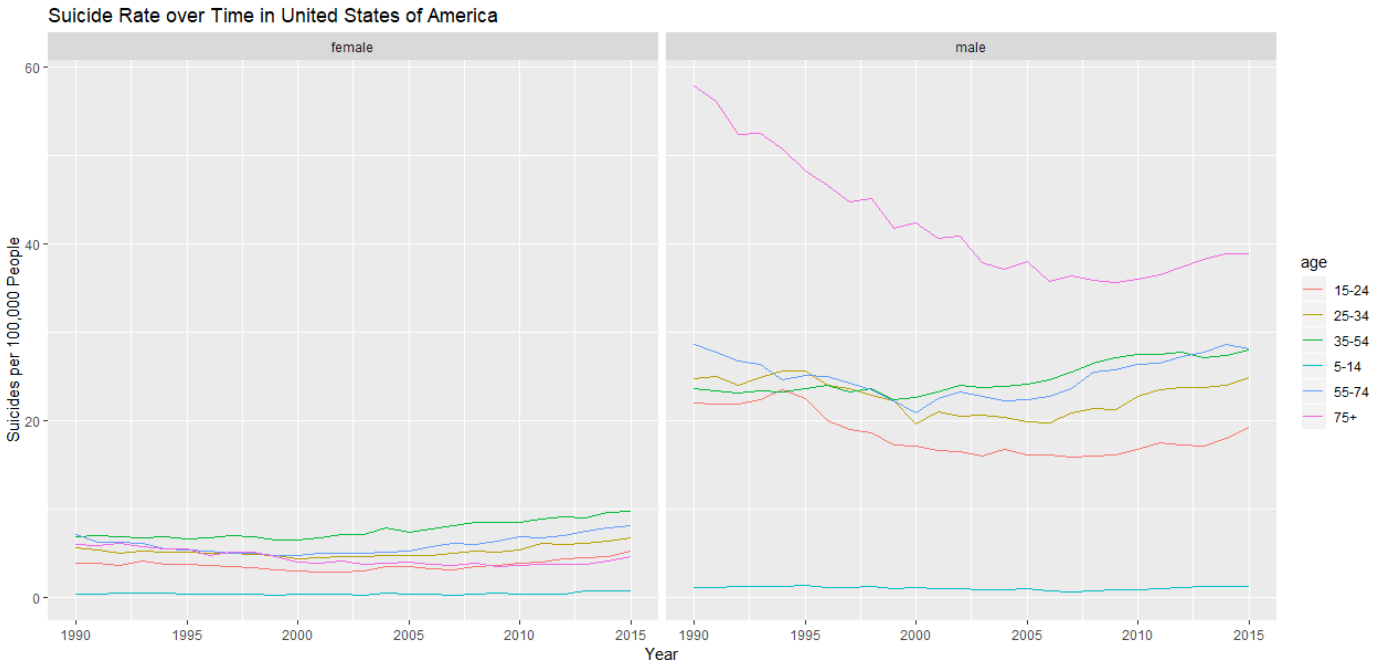



```
time_trend(df, "Hungary")
```

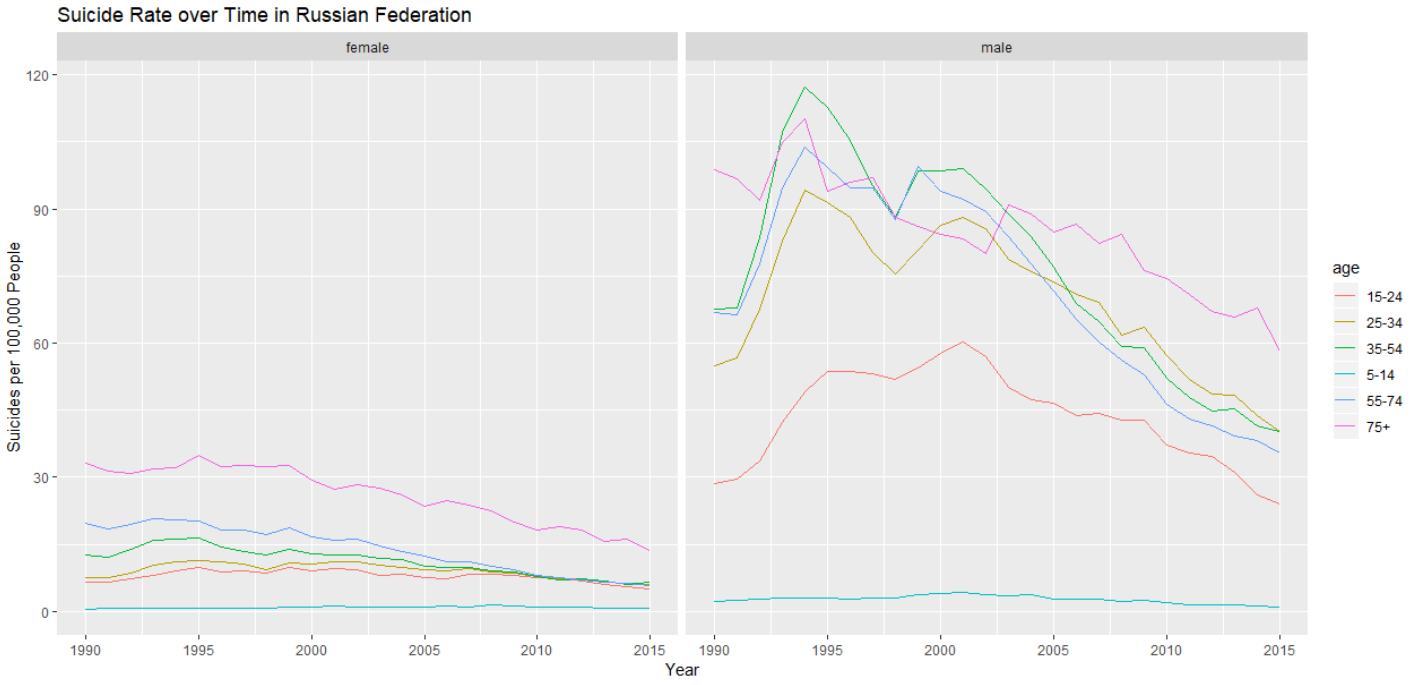


```
time_trend(df, "Finland")
```

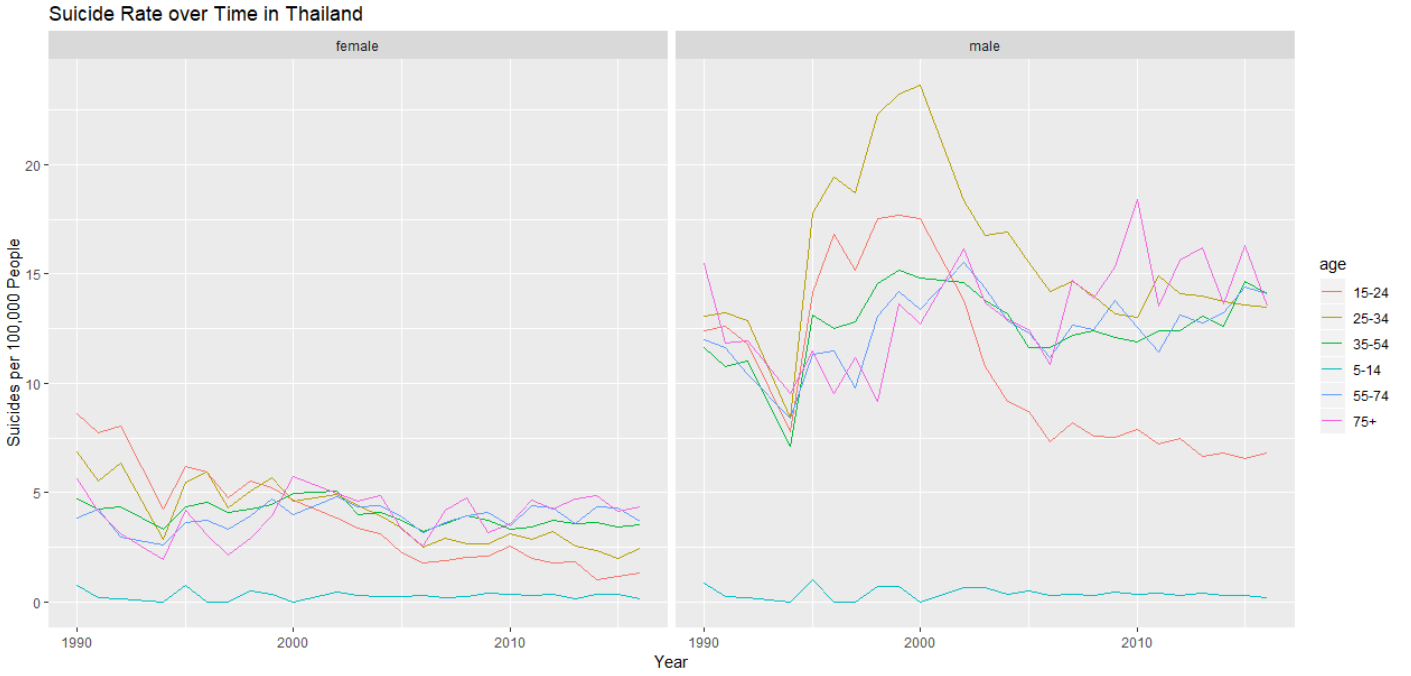

```
time_trend(df, "United States of America")
```




```
time_trend(df, "Russian Federation")
```



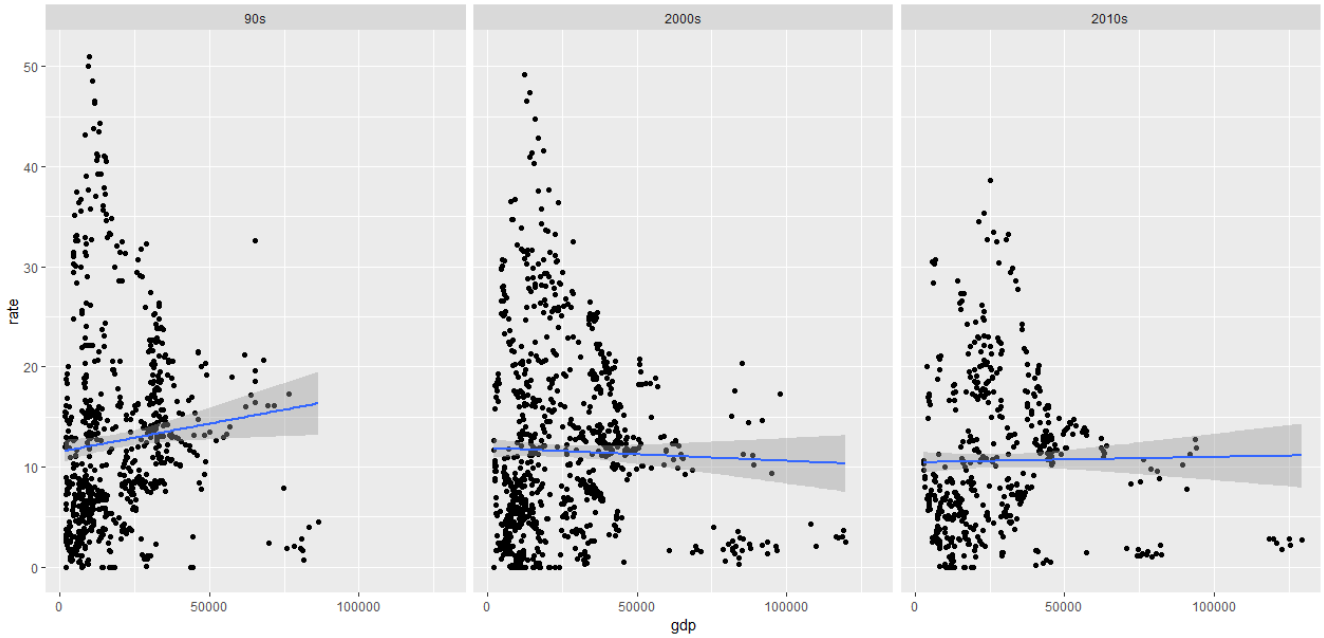

```
time_trend(df, "Thailand")
```



These time trends, divided by gender and age group, reflect that different countries have different age groups most vulnerable to suicide. In the United States, men 75+ have the highest suicide rates for all years, while in Lithuania men 34-54 have the highest incidence of suicide. They also show if the rates have increased or stayed steady over time.

To check if there is a visible relationship between GDP and suicide rates, we can plot country-level suicide rates and GDP(PPP) data.

```
plot_aggregate(df)
```



There is no visible relationship, given this data, between GDP(PPP- constant 2011 int $) and suicide rates. The plots are divided by decades. The 90s reflect a slightly positive relationship between GDP and suicide rates.

We can verify that with a simple regression on GDP (PPP) on the suicide rate.

Our second regression model, a multivarate_regression, is much better model for creating a predicted suicide rate, as it adjusts for country, age group, gdp, and gender. Though the coefficients are not shown, we can see that the adjusted R-squared is: 0.5213.

```
simple_regression(df)
```

```
##
## Call:
## lm(formula = rate ~ gdp, data = aggregate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.693  -6.776  -1.769   4.621  39.337
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.168e+01  3.082e-01   37.90   <2e-16 ***
## gdp         2.960e-07  1.003e-05    0.03    0.976
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.035 on 2134 degrees of freedom
## Multiple R-squared:  4.082e-07,  Adjusted R-squared:  -0.0004682
## F-statistic: 0.0008712 on 1 and 2134 DF,  p-value: 0.9765
```

```
multivarate_regression(df)
```

```
##
## Call:
## lm(formula = rate ~ year + as.factor(age) + NY.GDP.PCAP.PP.KD +
##     as.factor(sex) + as.factor(country), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.240  -6.973  -1.278   4.617 264.090
##
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.06 on 25519 degrees of freedom
## Multiple R-squared:  0.5234, Adjusted R-squared:  0.5213
## F-statistic: 250.2 on 112 and 25519 DF,  p-value: < 2.2e-1
```