Motivation: Suicide is unfortunate and I want to learn what makes people give up their lives, may be we can find how to prevent suicide by analyzing the data

```
library(readr)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────────────────────
──────────────────────── tidyverse 1.2.1 ──
```

```
## ✔ ggplot2 3.1.0       ✔ purrr   0.3.0
## ✔ tibble  2.0.1       ✔ dplyr   0.8.0.1
## ✔ tidyr   0.8.3       ✔ stringr 1.3.1
## ✔ ggplot2 3.1.0       ✔ forcats 0.3.0
```

```
## ── Conflicts ────────────────────────────────────────────────────────────
──────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
suicide <- read_csv("/Users/zhijiewu/Downloads/master.csv")
```

```
## Parsed with column specification:
## cols(
##   country = col_character(),
##   year = col_double(),
##   sex = col_character(),
##   age = col_character(),
##   suicides_no = col_double(),
##   population = col_double(),
##   `suicides/100k pop` = col_double(),
##   `country-year` = col_character(),
##   `HDI for year` = col_double(),
##   `gdp_for_year ($)` = col_number(),
##   `gdp_per_capita ($)` = col_double(),
##   generation = col_character()
## )
```

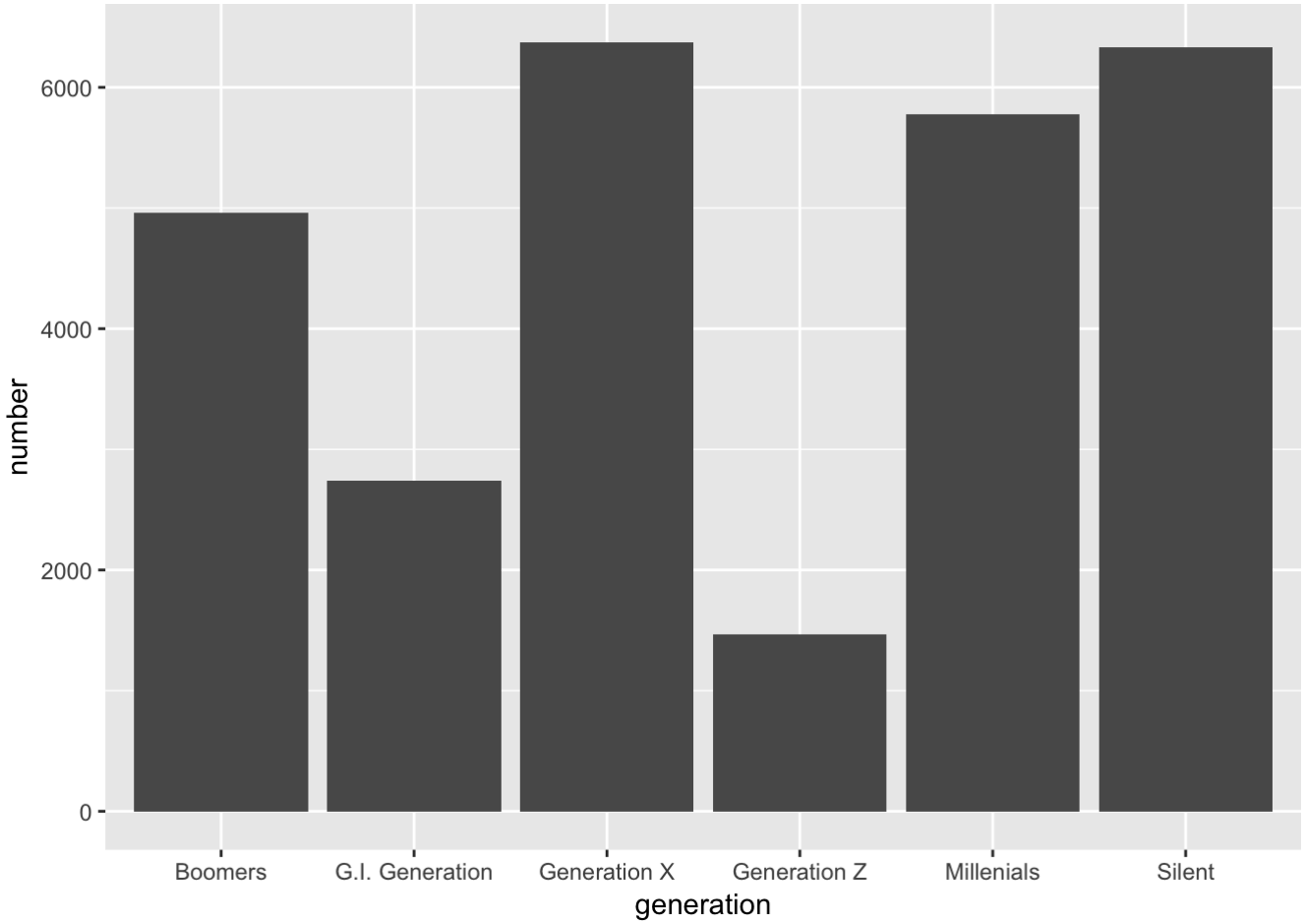
```
colnames(suicide)[colnames(suicide) == "gdp_for_year ($)"] <- "gdp"
colnames(suicide)[colnames(suicide) == "gdp_per_capita ($)"] <- "gdp_per_capita"
colnames(suicide)[colnames(suicide) == "suicides/100k pop"] <-"suicides_per_100k"


suicide <- suicide %>% filter(year < 2016) %>% select(-c('HDI for year'))
suicide
```

```
## # A tibble: 27,660 x 11
##    country  year sex   age   suicides_no population suicides_per_10…
##    <chr>   <dbl> <chr> <chr>       <dbl>      <dbl>            <dbl>
##  1 Albania  1987 male  15-2…          21     312900             6.71
##  2 Albania  1987 male  35-5…          16     308000             5.19
##  3 Albania  1987 fema… 15-2…          14     289700             4.83
##  4 Albania  1987 male  75+ …           1      21800             4.59
##  5 Albania  1987 male  25-3…           9     274300             3.28
##  6 Albania  1987 fema… 75+ …           1      35600             2.81
##  7 Albania  1987 fema… 35-5…           6     278800             2.15
##  8 Albania  1987 fema… 25-3…           4     257200             1.56
##  9 Albania  1987 male  55-7…           1     137500             0.73
## 10 Albania  1987 fema… 5-14…           0     311000             0
## # … with 27,650 more rows, and 4 more variables: `country-year` <chr>,
## #   gdp <dbl>, gdp_per_capita <dbl>, generation <chr>
```

we change some type of the data, see that data in 2016 is incomplete, so just remove that, also remove the HDI for year column because of too much missing data

let first see how generation affect

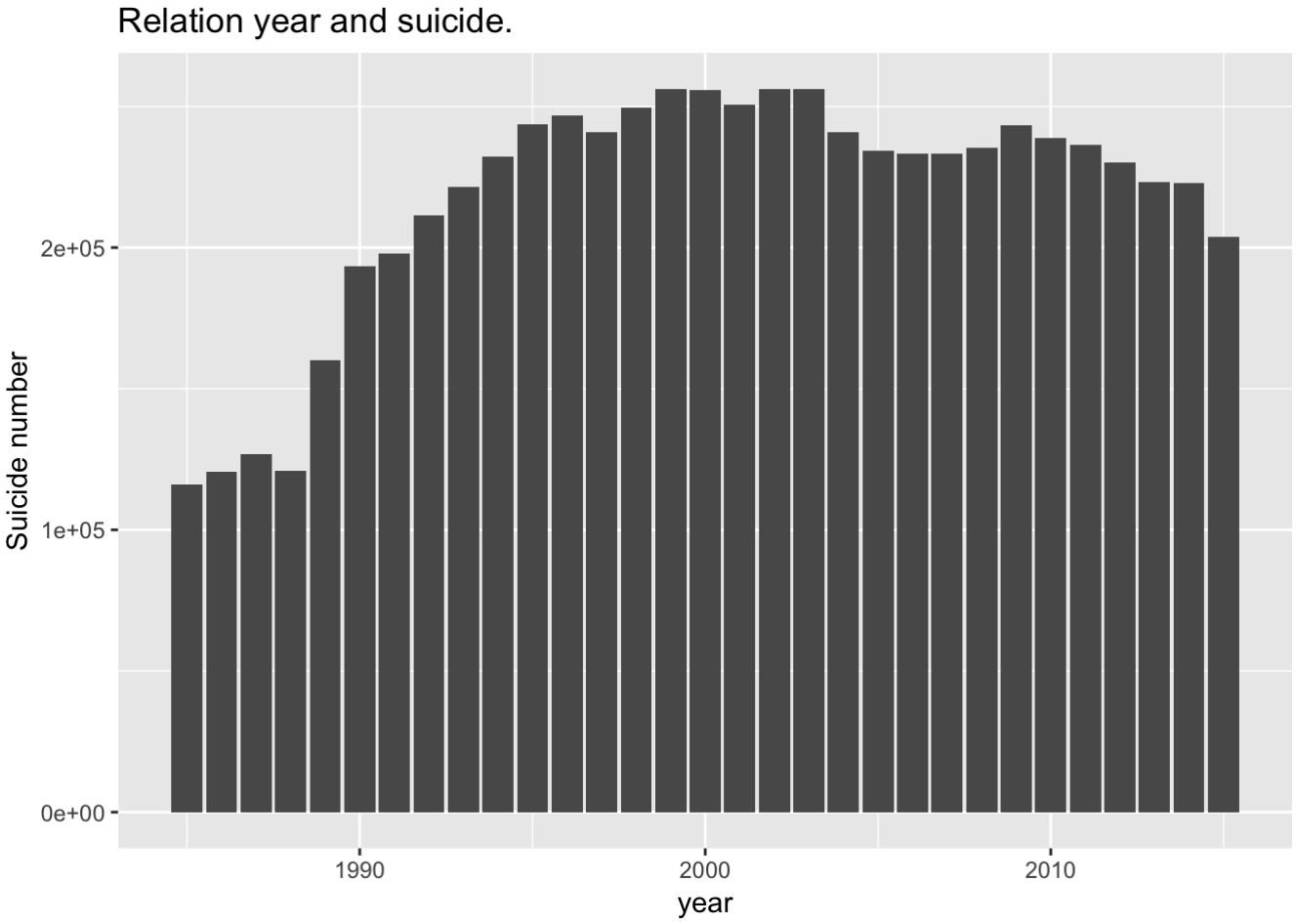
```
suicide %>% group_by(generation) %>% summarize(number = n())  %>% ggplot(mapping=aes(x=
generation, y = number)) +  geom_bar(stat="identity")
```



we find that Generation Z and G.I Generation is particularly low, other generation almost the same

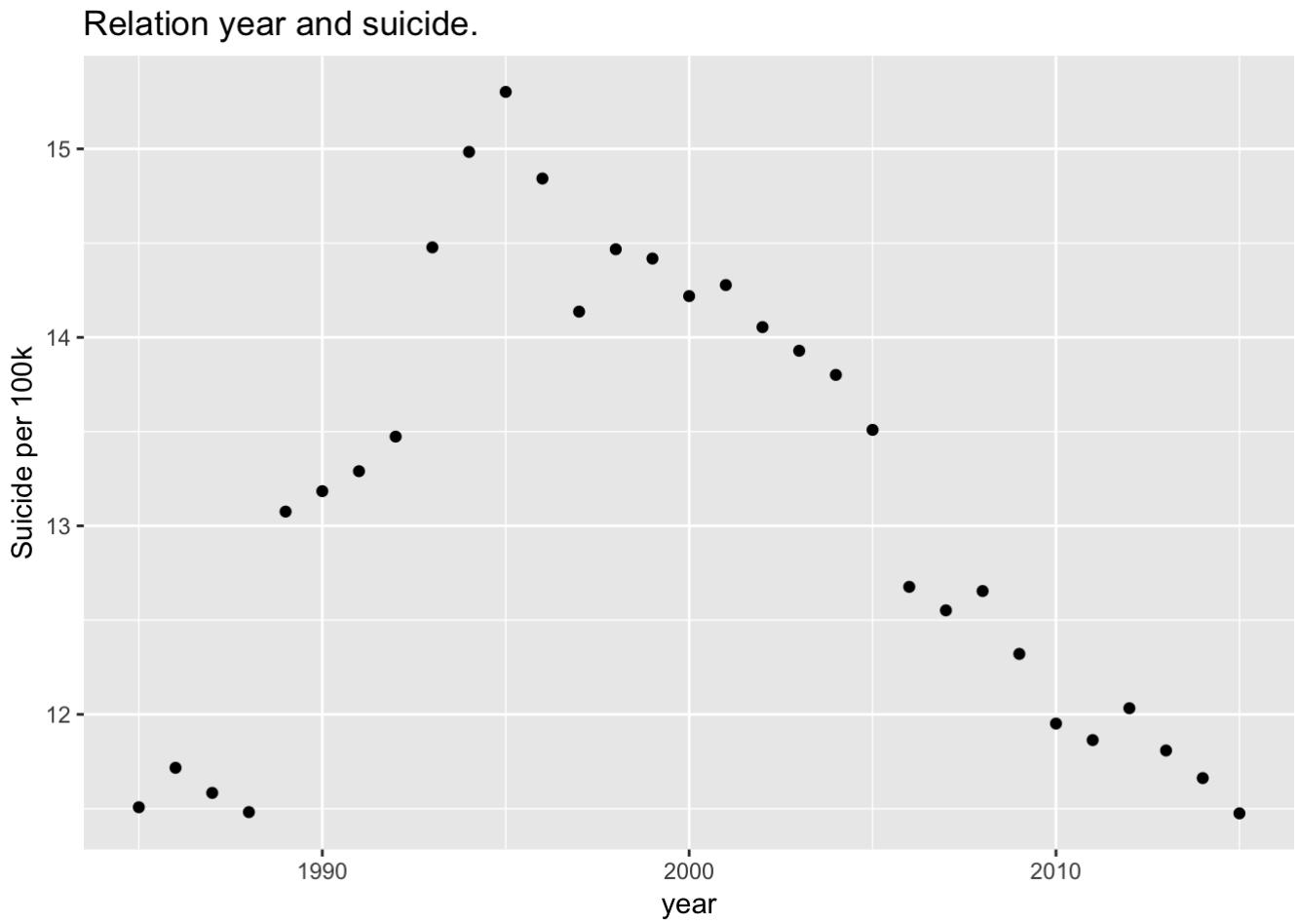
Then what about year with suicide?

```
suicide %>% group_by(year) %>% summarize(number = sum(suicides_no)) %>% ggplot(mapping=a
es(x= year, y = number)) +  geom_bar(stat="identity") +  labs(
       title = "Relation year and suicide.",
       x = "year",
       y = "Suicide number")
```



we see that there is no obvious relation with year and suicide number, but consider that population also increase, use suicide number per 100k may be more accurate.

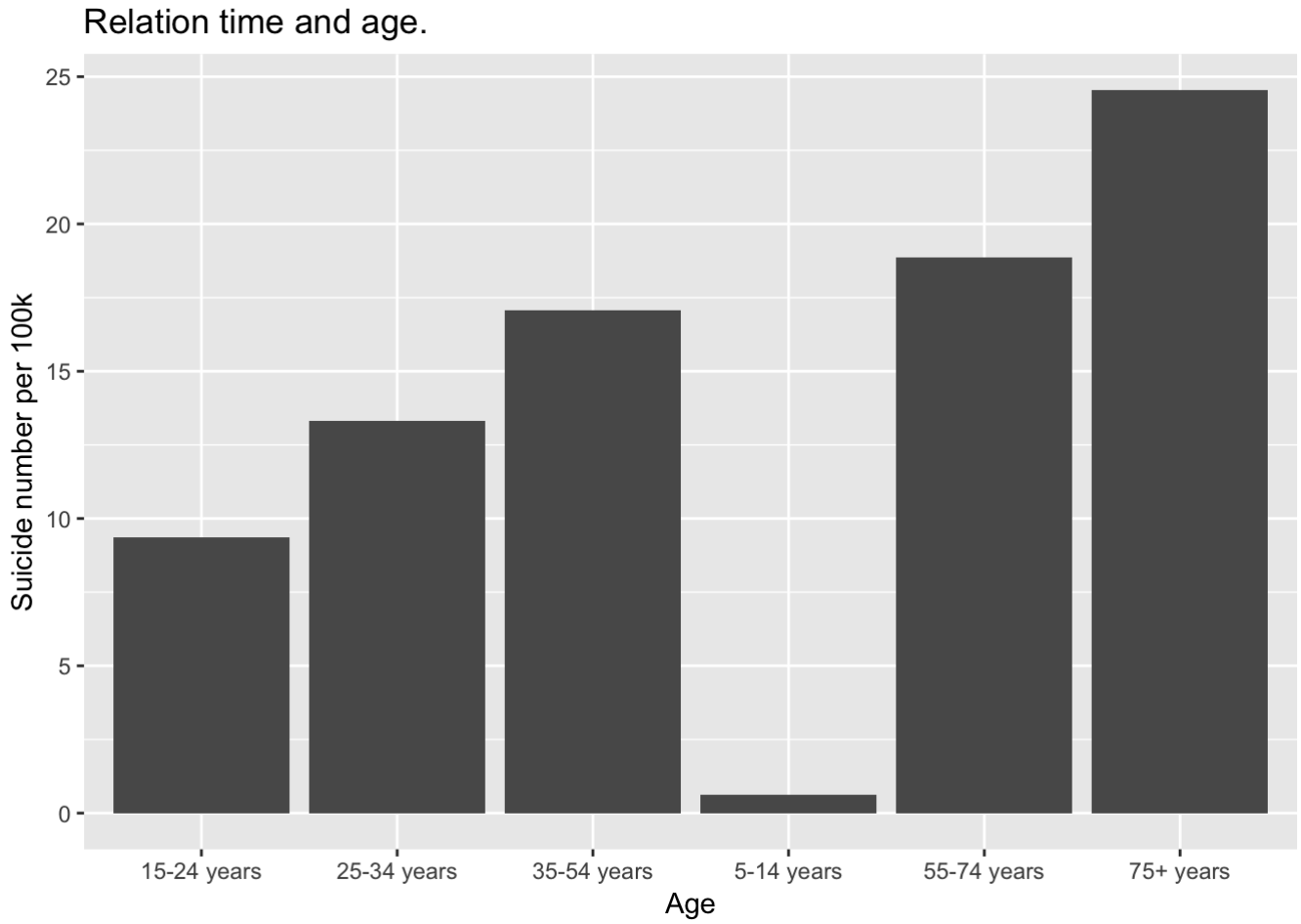
```
suicide_per_df <- suicide %>% group_by(year) %>% summarize(pop = sum(population), number
= sum(suicides_no), suicide_per_100 = (number/pop) * 100000)
suicide_per_df %>% ggplot(mapping=aes(x= year, y = suicide_per_100)) +  geom_point()+  l
abs(
       title = "Relation year and suicide.",
       x = "year",
       y = "Suicide per 100k")
```



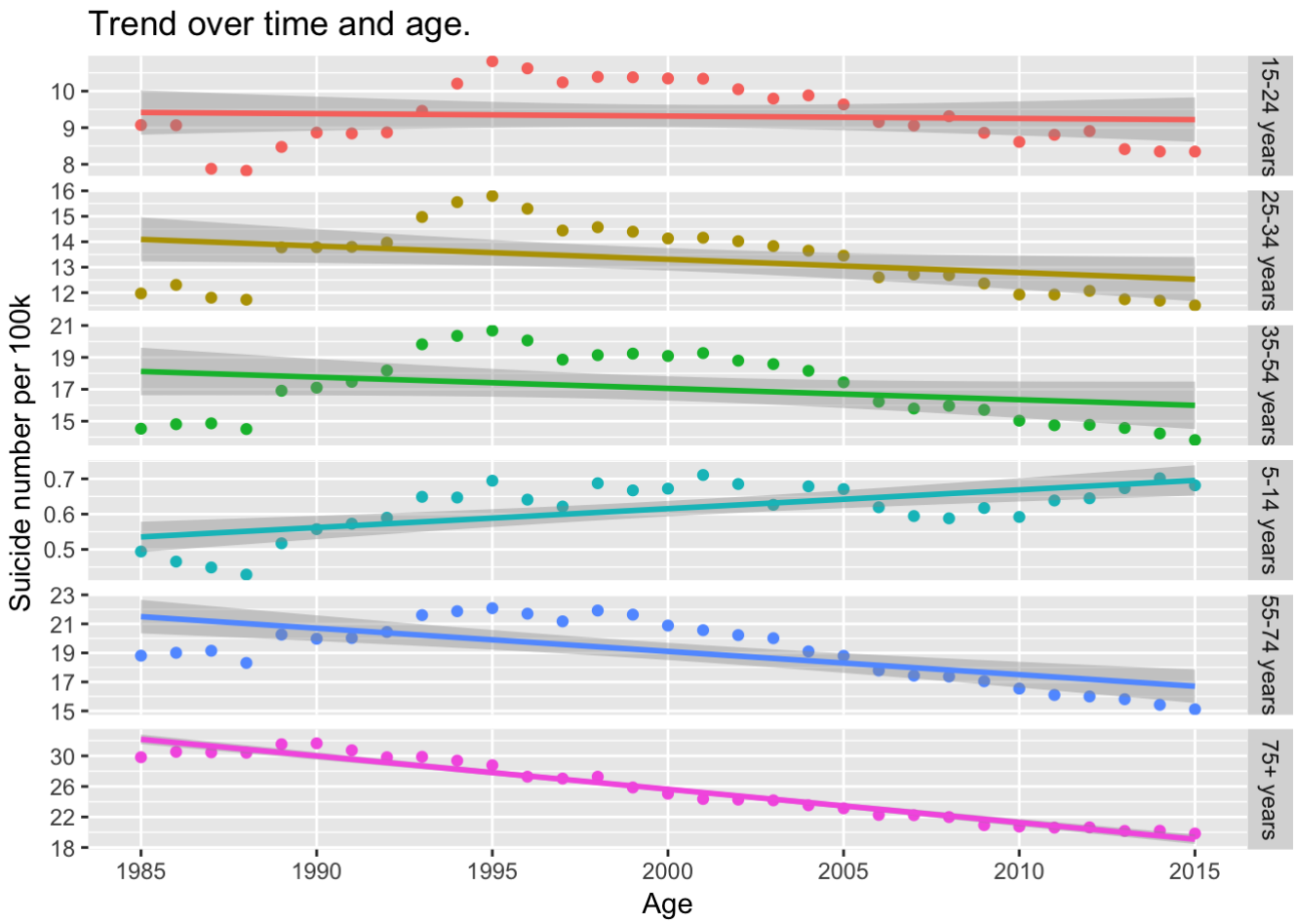
a general trend of decreasing of suicide

maybe that is related to age because as people's life is prolong nowadays, people may not willing to death

```
suicide_per_df <- suicide %>% group_by(age) %>% summarize(pop = sum(population), number
= sum(suicides_no), suicide_per_100 = (number/pop) * 100000)
suicide_per_df %>% ggplot(mapping=aes(x= age, y = suicide_per_100)) +  geom_bar(stat="id
entity") + labs(
       title = "Relation time and age.",
       x = "Age",
       y = "Suicide number per 100k")
```



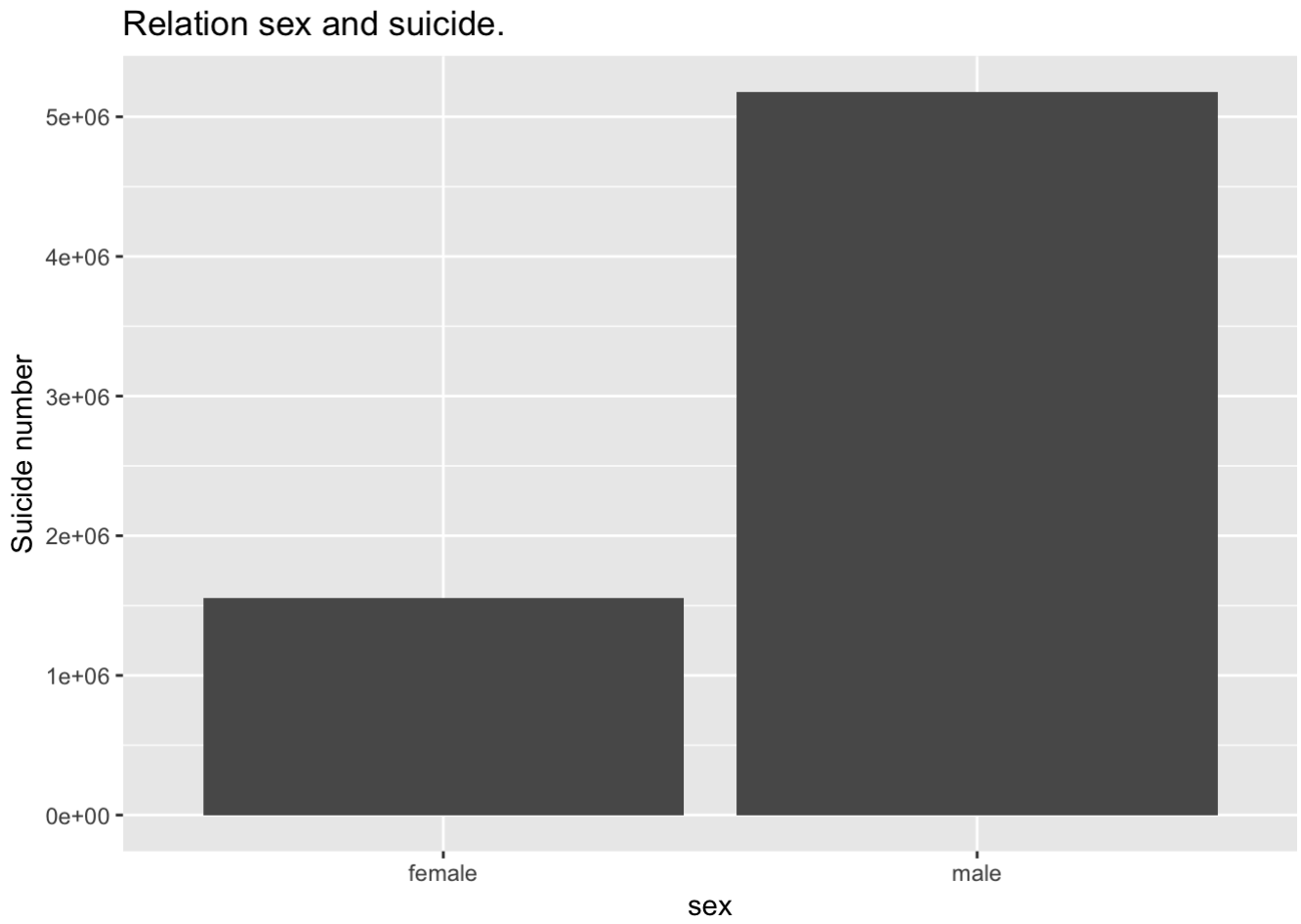

```
suicide_per_df <- suicide %>% group_by(age,year) %>% summarize(pop = sum(population), nu
mber = sum(suicides_no), suicide_per_100 = (number/pop) * 100000)
suicide_per_df  %>% ggplot(mapping=aes(x= year, y = suicide_per_100,col = age)) +  facet
_grid(age~., scales = "free_y") + labs(
       title = "Trend over time and age.",
       x = "Age",
       y = "Suicide number per 100k") +  geom_point() +geom_smooth(method = lm) +
  theme(legend.position = "none") +
  scale_x_continuous(breaks = seq(1985, 2015, 5), minor_breaks = F)
```



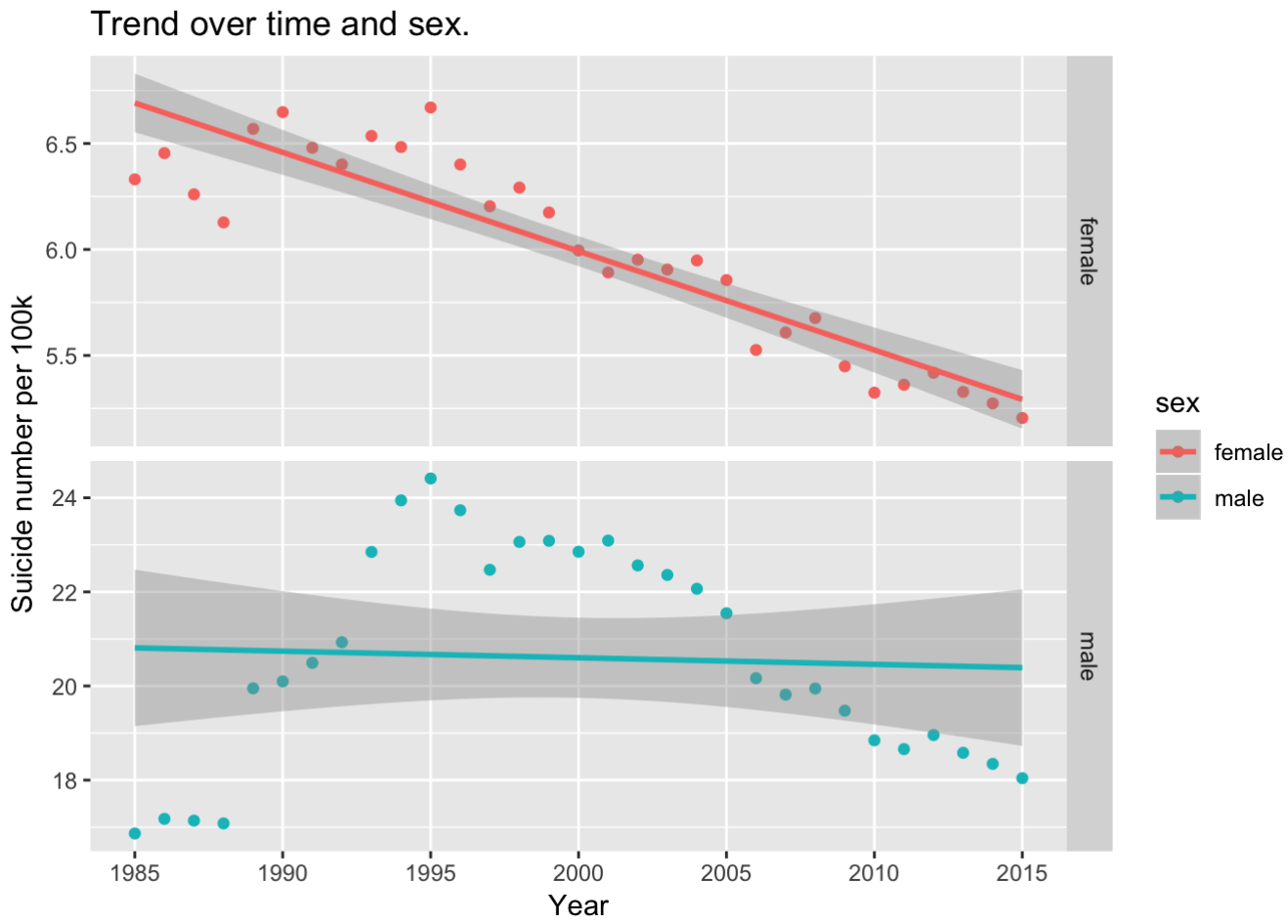
Well, just opposite to my hypothesis, 75+ years old is most prone to suicide,but the trend for suicide is decreasing, expect for 5-14 year old, it is increasing maybe because of the increasing peer pressure

Then how does it related to sex?

```
suicide %>% group_by(sex) %>% summarize(number = sum(suicides_no))  %>% ggplot(mapping=a
es(x= sex, y = number)) +
  labs(
       title = "Relation sex and suicide.",
       x = "sex",
       y = "Suicide number") + geom_bar(stat="identity")
```




```
suicide_per_df <- suicide %>% group_by(sex,year) %>% summarize(pop = sum(population), nu
mber = sum(suicides_no), suicide_per_100 = (number/pop) * 100000)
suicide_per_df %>% ggplot(mapping=aes(x= year, y = suicide_per_100, col =sex)) +  facet_
grid(sex~., scales = "free_y") + labs(
       title = "Trend over time and sex.",
       x = "Year",
       y = "Suicide number per 100k") +  geom_point() +geom_smooth(method = lm)+
  scale_x_continuous(breaks = seq(1985, 2015, 5), minor_breaks = F)
```



male have a higher suicide number,female keep decreasing, and male first increase then decrease,almost keep the same

Does sex and years old all correlated with suicide?

```
temp <- suicide %>% group_by(year,sex,age) %>%
summarize(suicide_per_100k = (sum(suicides_no) / sum(population)) * 100000)
fit <- aov(formula = suicide_per_100k~age+sex, data = temp)
summary(fit)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## age           5  31703    6341   183.9 <2e-16 ***
## sex           1  26858   26858   779.0 <2e-16 ***
## Residuals   365  12584      34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

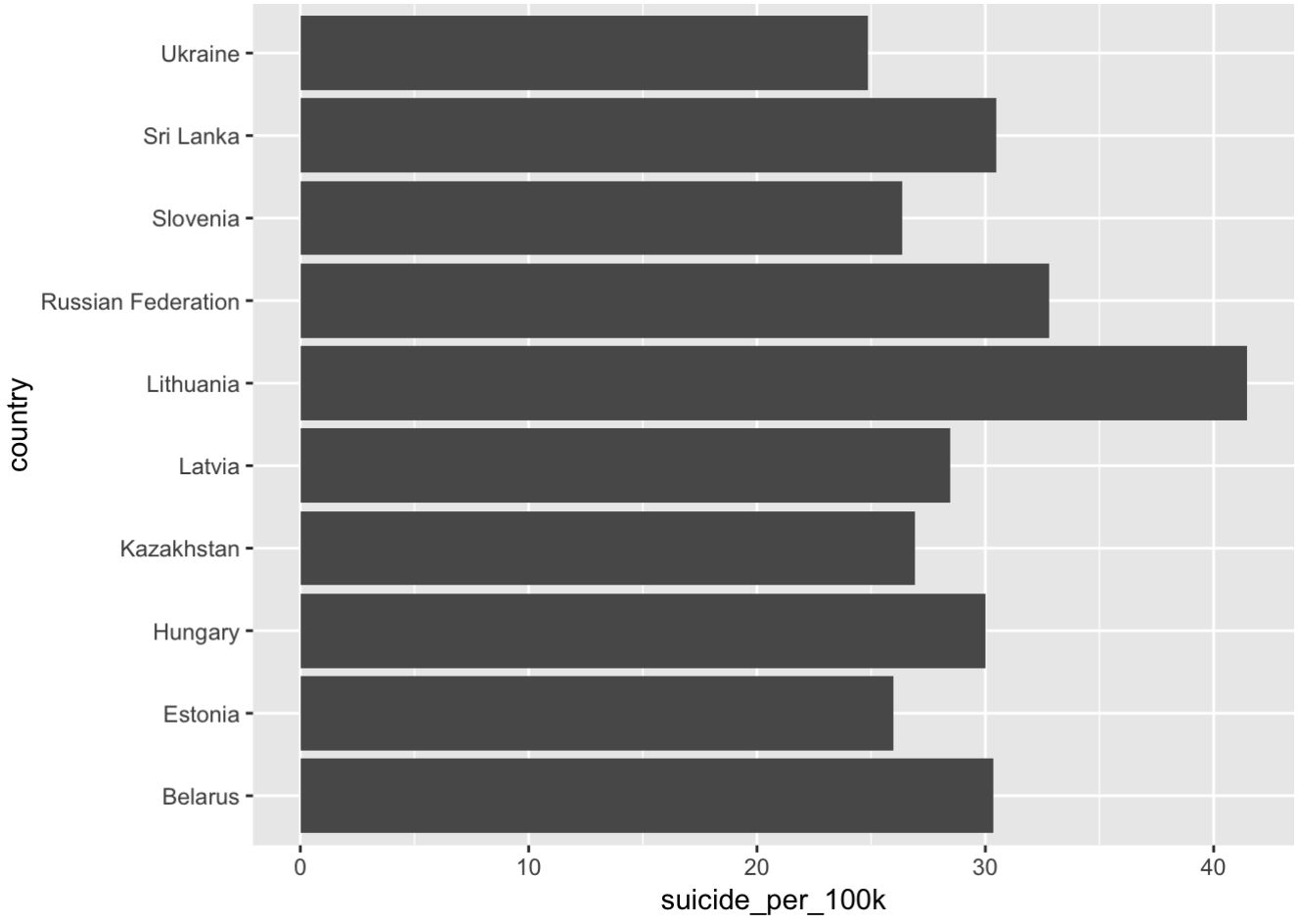
We find that both p value is less than 0.05, so we can't reject there is no relationship of sex and age with suicide, and because the F value of sex is higher, sex impact more than age

we want to find if the suicide rate have some relation with country,we select top 10 highet suicide rate country

```
country <- suicide %>%
  group_by(country) %>%
  summarize(suicide_per_100k = (sum(as.numeric(suicides_no)) / sum(as.numeric(populatio
n))) * 100000) %>%
  arrange(desc(suicide_per_100k)) %>% slice(c(1:10))
country
```

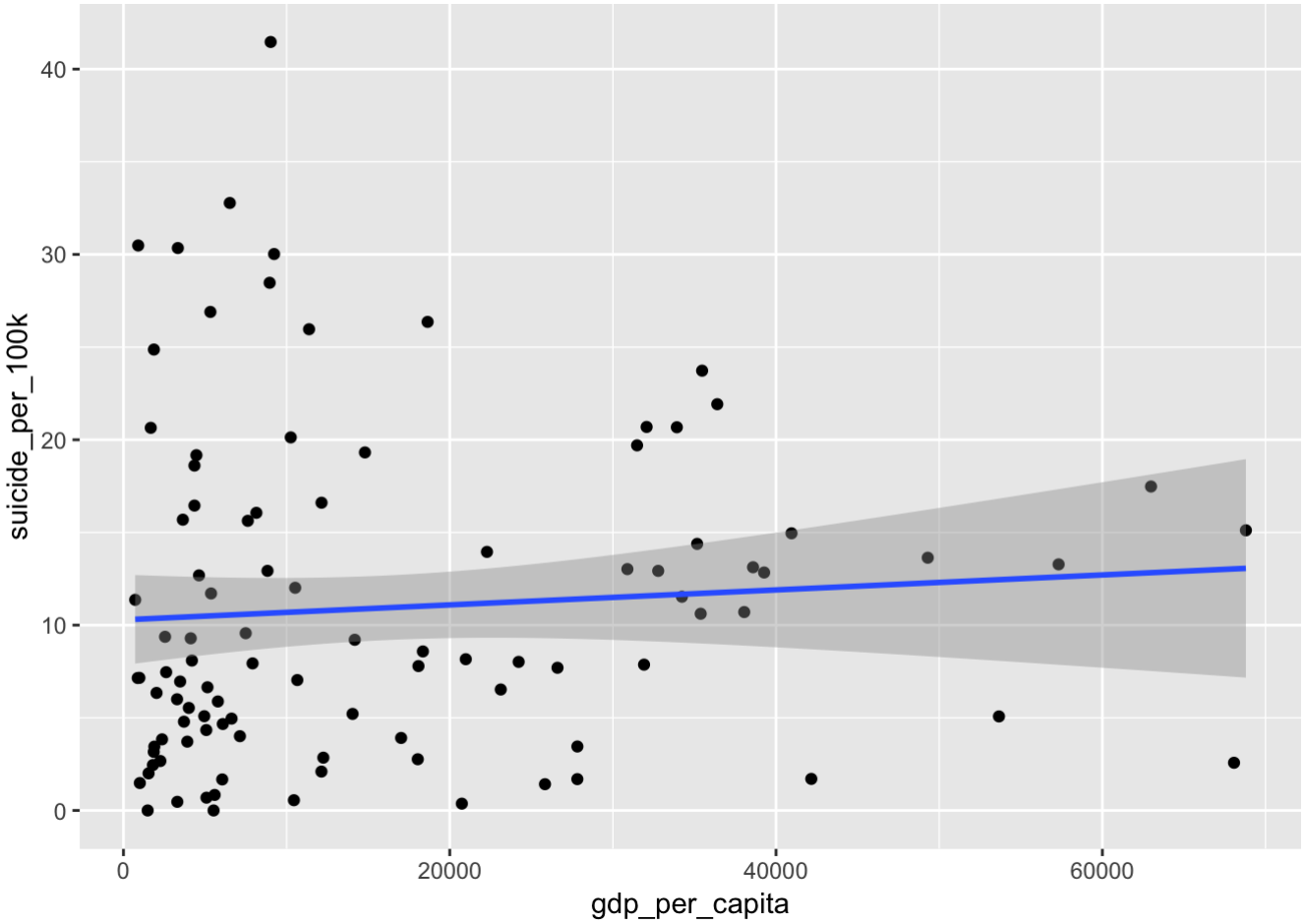
```
## # A tibble: 10 x 2
##    country            suicide_per_100k
##    <chr>                         <dbl>
##  1 Lithuania                      41.5
##  2 Russian Federation             32.8
##  3 Sri Lanka                      30.5
##  4 Belarus                        30.3
##  5 Hungary                        30.0
##  6 Latvia                         28.5
##  7 Kazakhstan                     26.9
##  8 Slovenia                       26.4
##  9 Estonia                        26.0
## 10 Ukraine                        24.9
```

```
ggplot(country, aes(x = country, y = suicide_per_100k)) +
  geom_bar(stat = "identity") +
coord_flip()+ theme(legend.position = "bottom")
```



seems like the top 10 highest country is relatively poor, we go to explore the GDP per capita vs suicide rate.

```
country_mean_gdp <- suicide %>%
  group_by(country) %>%
  summarize(suicide_per_100k = (sum(suicides_no) / sum(population)) * 100000,
            gdp_per_capita = mean(gdp_per_capita))

ggplot(country_mean_gdp, aes(x = gdp_per_capita, y = suicide_per_100k)) +
  geom_point()  + geom_smooth(method = lm)
```




```
  labs(
    title = " GDP per capita vs Suicides per 100k",
       x = "GDP per capita",
       y = "Suicides per 100k")
```

```
## $x
## [1] "GDP per capita"
##
## $y
## [1] "Suicides per 100k"
##
## $title
## [1] " GDP per capita vs Suicides per 100k"
##
## attr(,"class")
## [1] "labels"
```

seems like have slight linear relation, let's generate a linear fit for that

```
Auto_fit <- lm(suicide_per_100k ~gdp_per_capita, data = country_mean_gdp)
Auto_fit
```

```
##
## Call:
## lm(formula = suicide_per_100k ~ gdp_per_capita, data = country_mean_gdp)
##
## Coefficients:
##    (Intercept)  gdp_per_capita
##      1.028e+01       4.036e-05
```

```
summary(Auto_fit)
```

```
##
## Call:
## lm(formula = suicide_per_100k ~ gdp_per_capita, data = country_mean_gdp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.755  -6.606  -2.556   4.748  30.815
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.028e+01  1.229e+00   8.370  4.1e-13 ***
## gdp_per_capita 4.036e-05  5.374e-05   0.751    0.454
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.769 on 98 degrees of freedom
## Multiple R-squared:  0.005723,   Adjusted R-squared:  -0.004423
## F-statistic: 0.5641 on 1 and 98 DF,  p-value: 0.4544
```

pvalue is 0.4544 > 0.05, so we don't reject the assumption that there is no relation between gdp per capita and suicide rate.

To summarize, suicide rate is highly related with sex and years old, but no highly related with gdp per capita

Here is some reference more about suicides: I find data from https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016 (https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016),

and some other reference will be United Nations Development Program. (2018). Human development index (HDI). Retrieved from http://hdr.undp.org/en/indicators/137506 (http://hdr.undp.org/en/indicators/137506)

World Bank. (2018). World development indicators: GDP (current US$) by country:1985 to 2016. Retrieved from http://databank.worldbank.org/data/source/world-development-indicators# (http://databank.worldbank.org/data/source/world-development-indicators#)

[Szamil]. (2017). Suicide in the Twenty-First Century [dataset]. Retrieved from https://www.kaggle.com/szamil/suicide-in-the-twenty-first-century/notebook (https://www.kaggle.com/szamil/suicide-in-the-twenty-first-century/notebook)

World Health Organization. (2018). Suicide prevention. Retrieved from http://www.who.int/mental_health/suicide-prevention/en/ (http://www.who.int/mental_health/suicide-prevention/en/)