suicide

Nguyen Diem Huong

12/15/2019

```
load(file = "suicide.RData")

library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------- tidyverse 1.3.0 --
```

```
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## v purrr   0.3.3
```

```
## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(nortest)
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.6.2
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

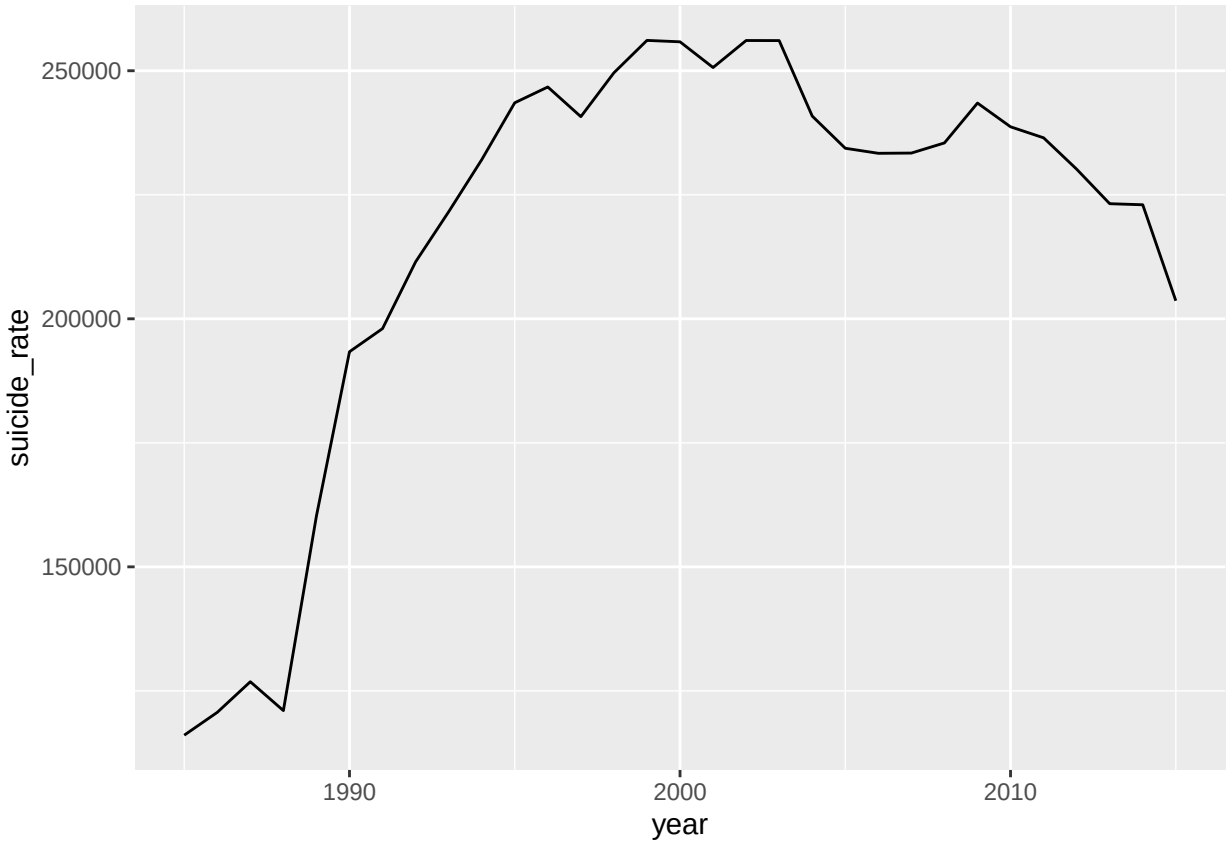
Exploracni analyza

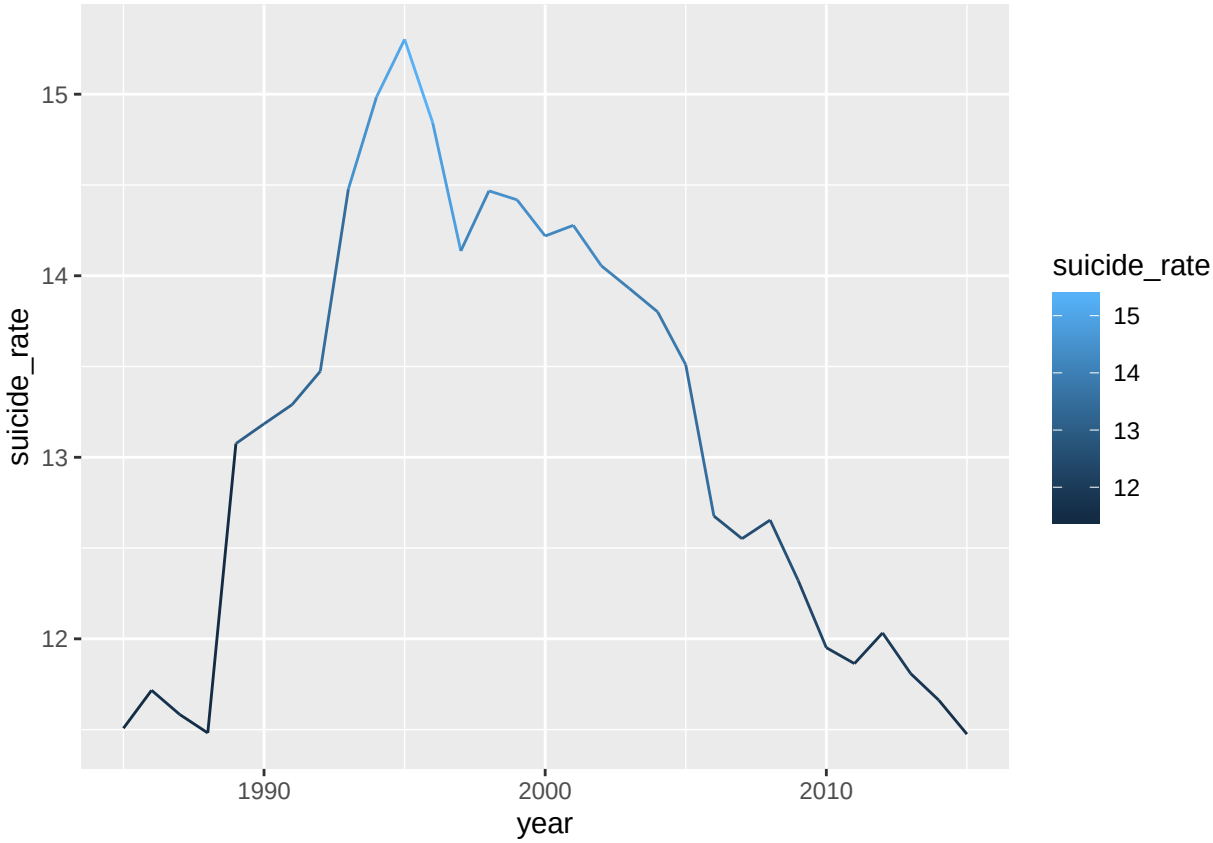
V explorační analýze byly vykresleny grafy proměnných vzhledem k časové ose. Cílem bylo zobrazit trend míry sebevražd (počet sebevražd na 100 000 obyvatel) dle různých atributů. Informace vydedukované z grafů byly pak použity ke zformování hypotéz.

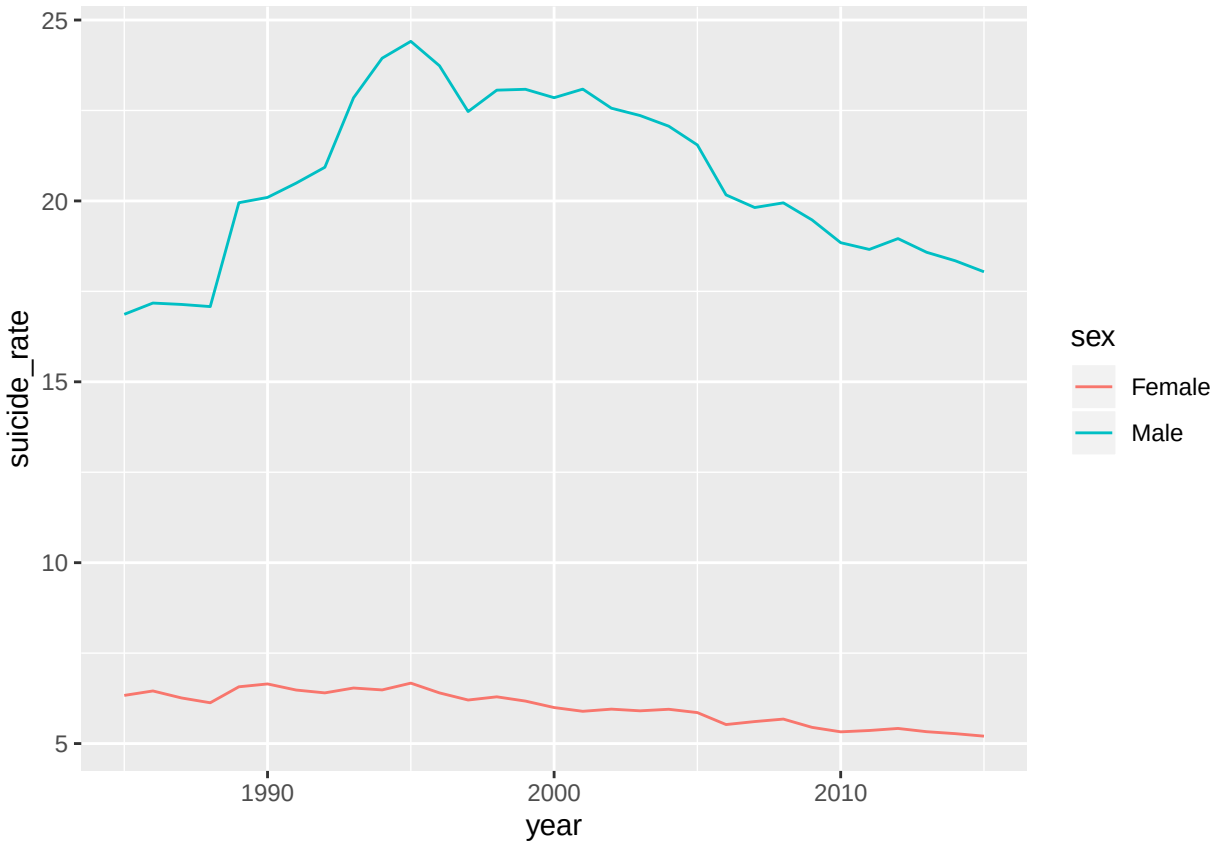
```
##        row col
## [1,] 20857   5
```

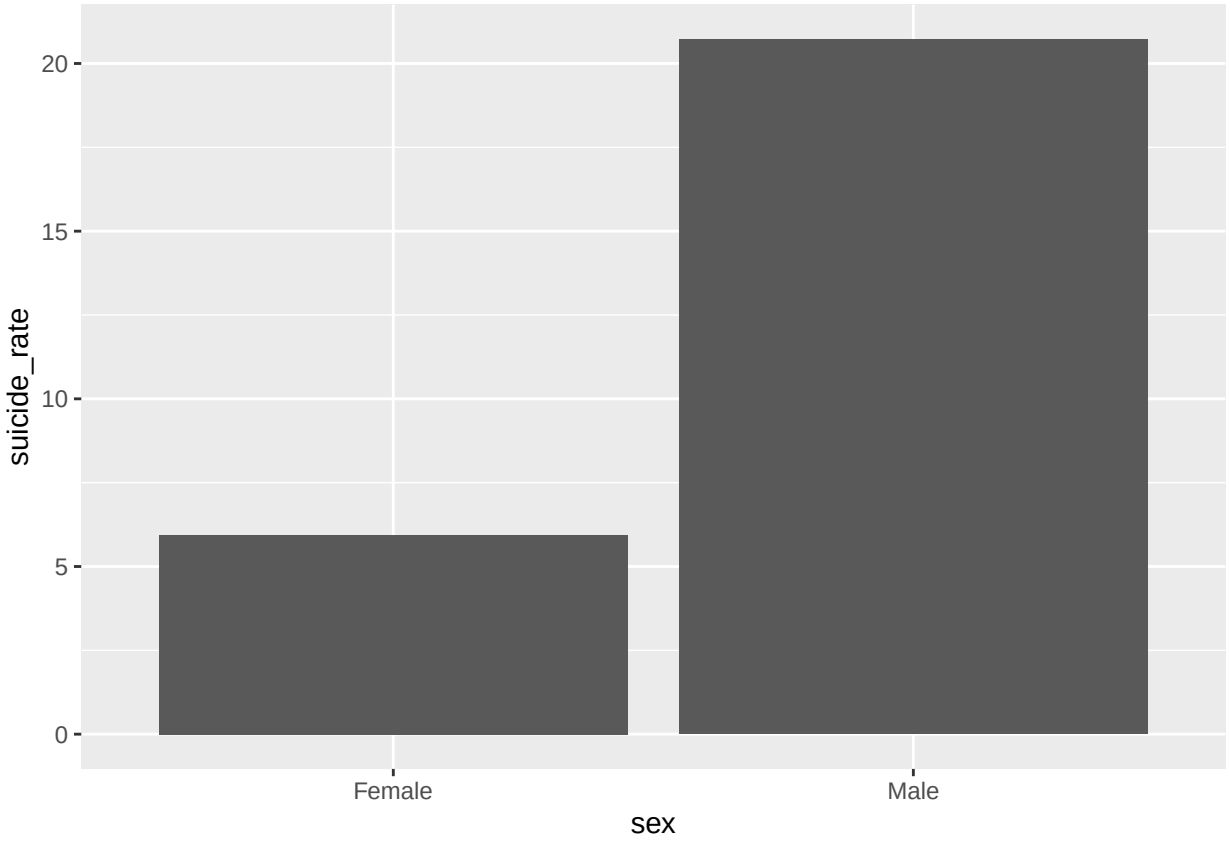
```
## # A tibble: 31 x 2
##     year num_suicide
##    <dbl>       <dbl>
##  1  1999      256119
##  2  2002      256095
##  3  2003      256079
##  4  2000      255832
##  5  2001      250652
##  6  1998      249591
##  7  1996      246725
##  8  1995      243544
##  9  2009      243487
```

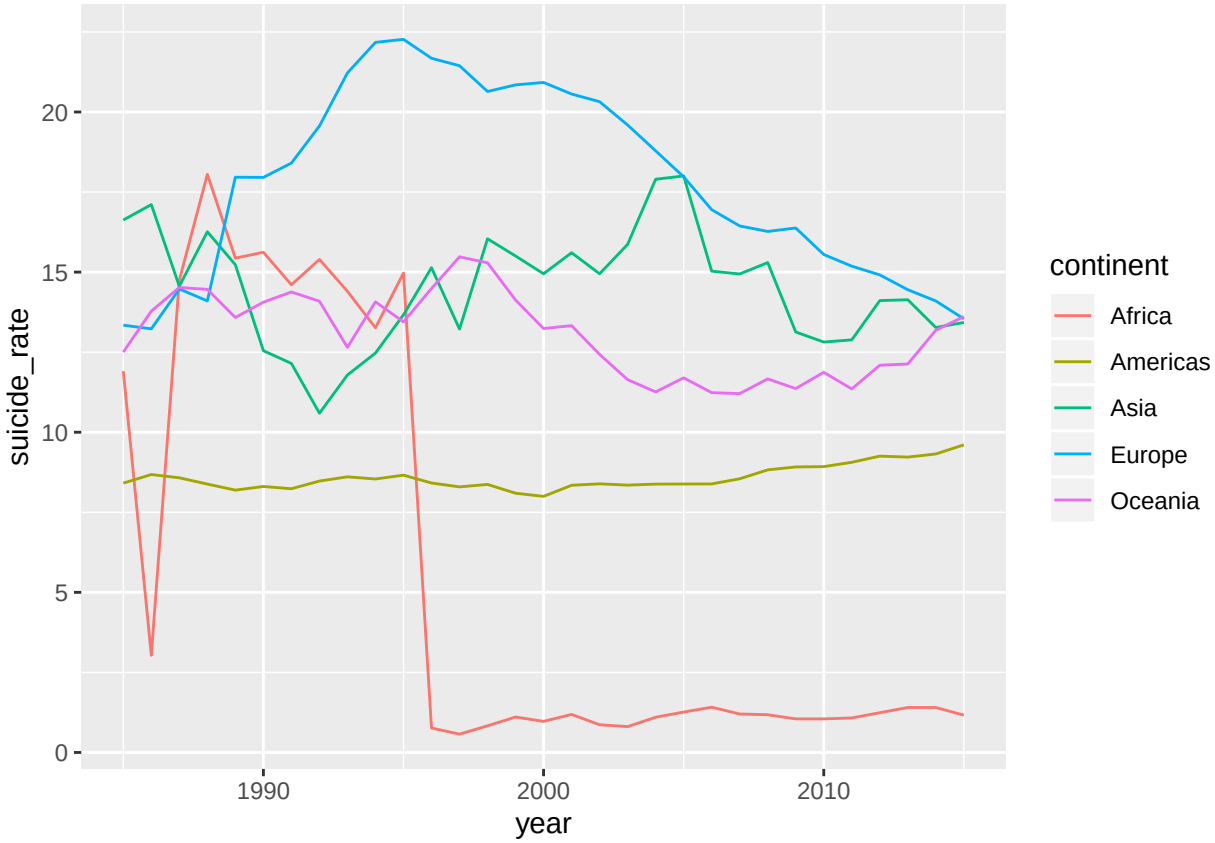
```
## 10  2004      240861
## # ... with 21 more rows
```

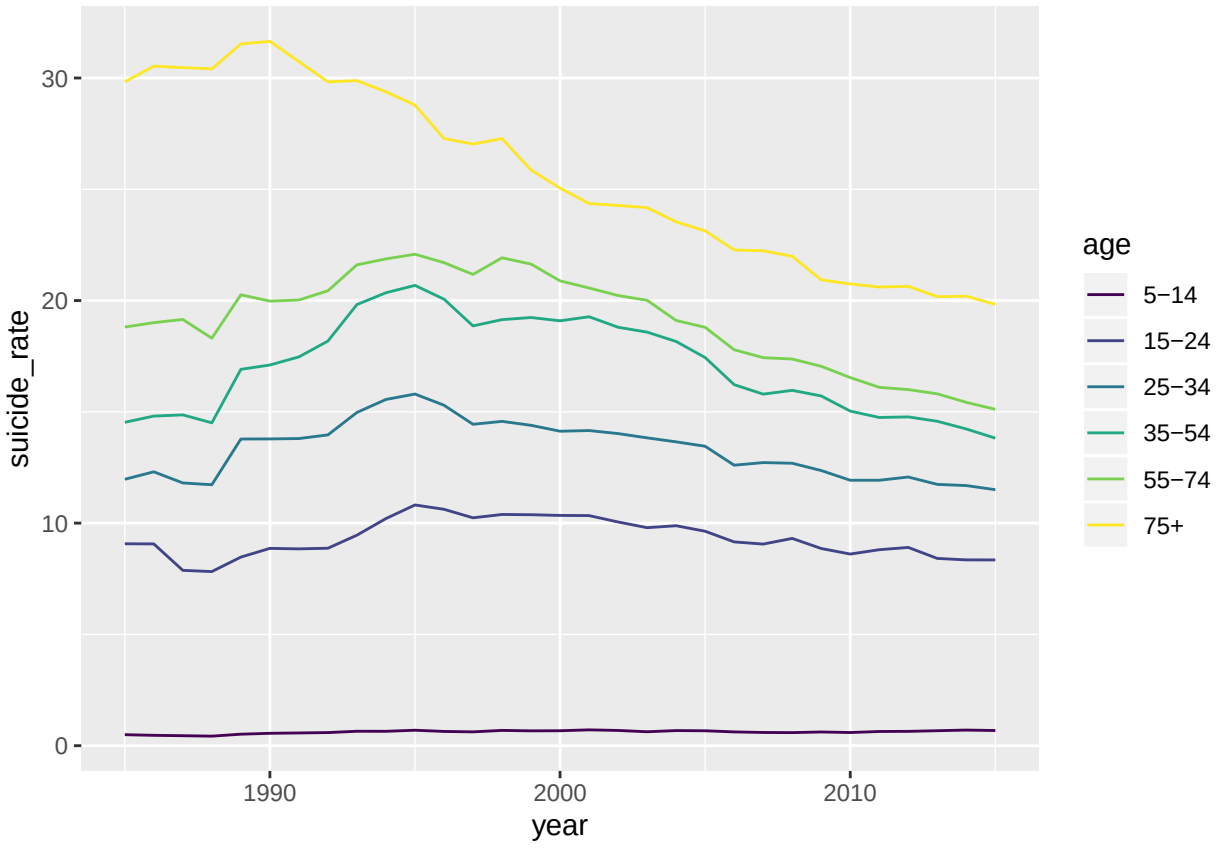


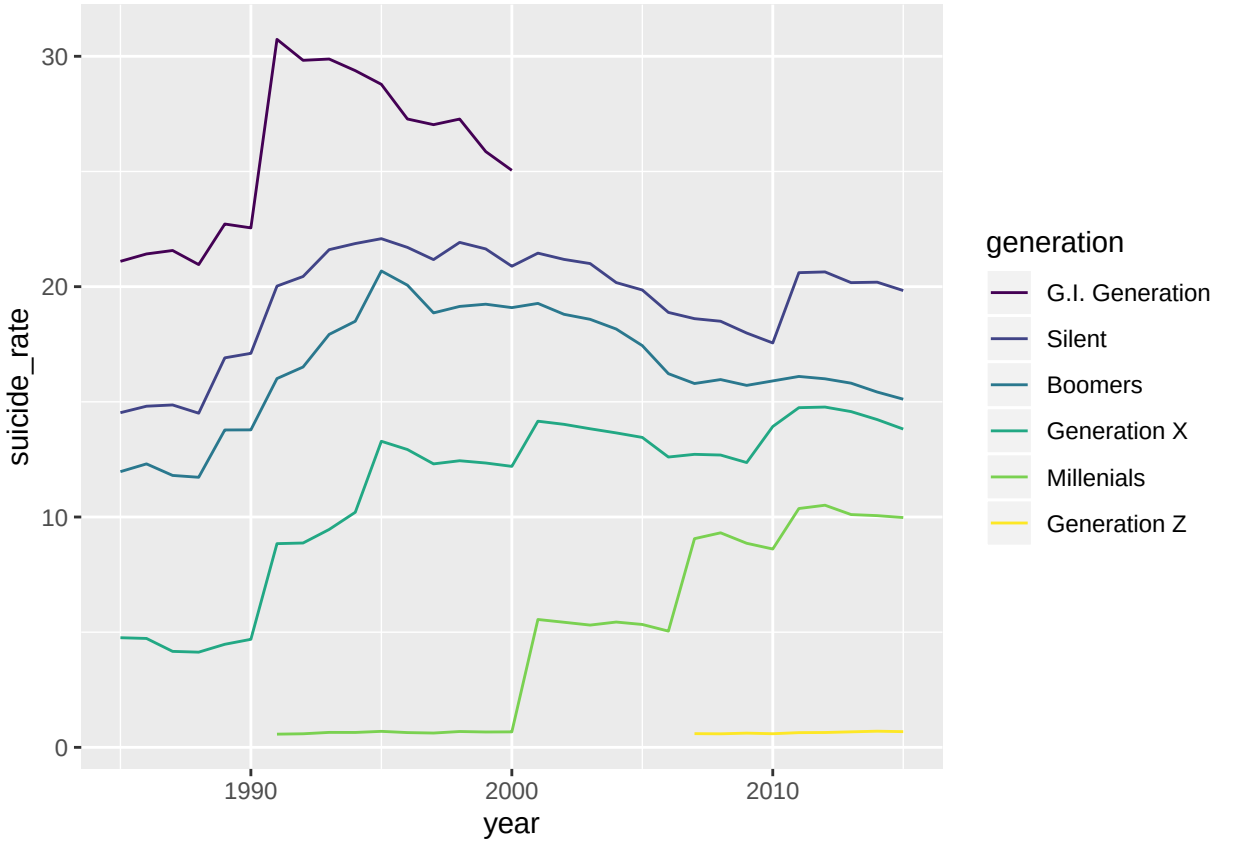

15
14
13
12
suicide_rate
1990
2000
2010
year
suicide_rate
15
14
13
12


25
20
15
10
5
suicide_rate
1990
2000
2010
year
sex
Female
Male


20
15
10
5
0
suicide_rate
Female
Male
sex


suicide_rate
20
15
10
5
0
1990
2000
2010
year
continent
Africa
Americas
Asia
Europe
Oceania


suicide_rate
30
20
10
0
1990
2000
2010
year
age
5−14
15−24
25−34
35−54
55−74
75+




```
## # A tibble: 10 x 2
##    country            suicide_rate
##    <fct>                     <dbl>
##  1 Finland                    32.3
##  2 Russian Federation         28.6
##  3 Austria                    25.2
##  4 Ukraine                    22.2
##  5 Belarus                    22.2
##  6 France                     21.5
##  7 Kazakhstan                 21.5
##  8 Czech Republic             20.6
##  9 Belgium                    20.2
## 10 Luxembourg                 19.0
```

Nejvyšší počet sebevražd měli celkově Rusové v roce 1994, ve věkovém rozmězí 35-54 let. Rusko mělo v roce 1994 také nejvyšší míru sebevražd ze všech států.Z roků měl nejvyšší počet sebevražd rok 1999 s celkovým 256119 počtem úmrtí. Z pohlaví vycházejí muži 3x - 4x hůře než ženy. Z kontinentů vychází nejhůře Evropa. Trendová přímka Afriky naznačuje, že mohlo pravděpodobně dojít ke změně metodiky sběru dat (podezřele radikální změna v míře sebevražd). Asie, která je známá svou vysokou mírou sebevražd je také poměrně vysoko, i když ne tolik, jak bych očekávala. Při vykreslení histogramu rozložení míry sebevražd Asiatů jsem vykoukala, že asijské ženy se zřejmě zabíjejí více (viz histogram v hypotéze) a podle toho jsem se inspirovala ke zformování hypotéz. Trend mezi věkovými skupinami je velmi podobný, přestože je značný rozdíl mezi tím, kde přímky začínají - vyšší věkové skupiny mají vyšší míru sebevražd. Podobně jsou na tom generace - starší generace jsou náchylnější k vyšší míře sebevražd.

funfact: Česko bylo v roce 1990 na 8. místě.

Hypotezy

H1: Míra sebevražd je vyšší v zemích, které mají vyšší HDP.

První hypotéza vznikla na základě myšlenky, že čím bohatší jsou lidé, tím více budou jejich problémy psychického rázu. GDP je společné pro všechny skupiny, v jednom státě a roce.

```
ad.test(data$suicides.100k.pop)
```

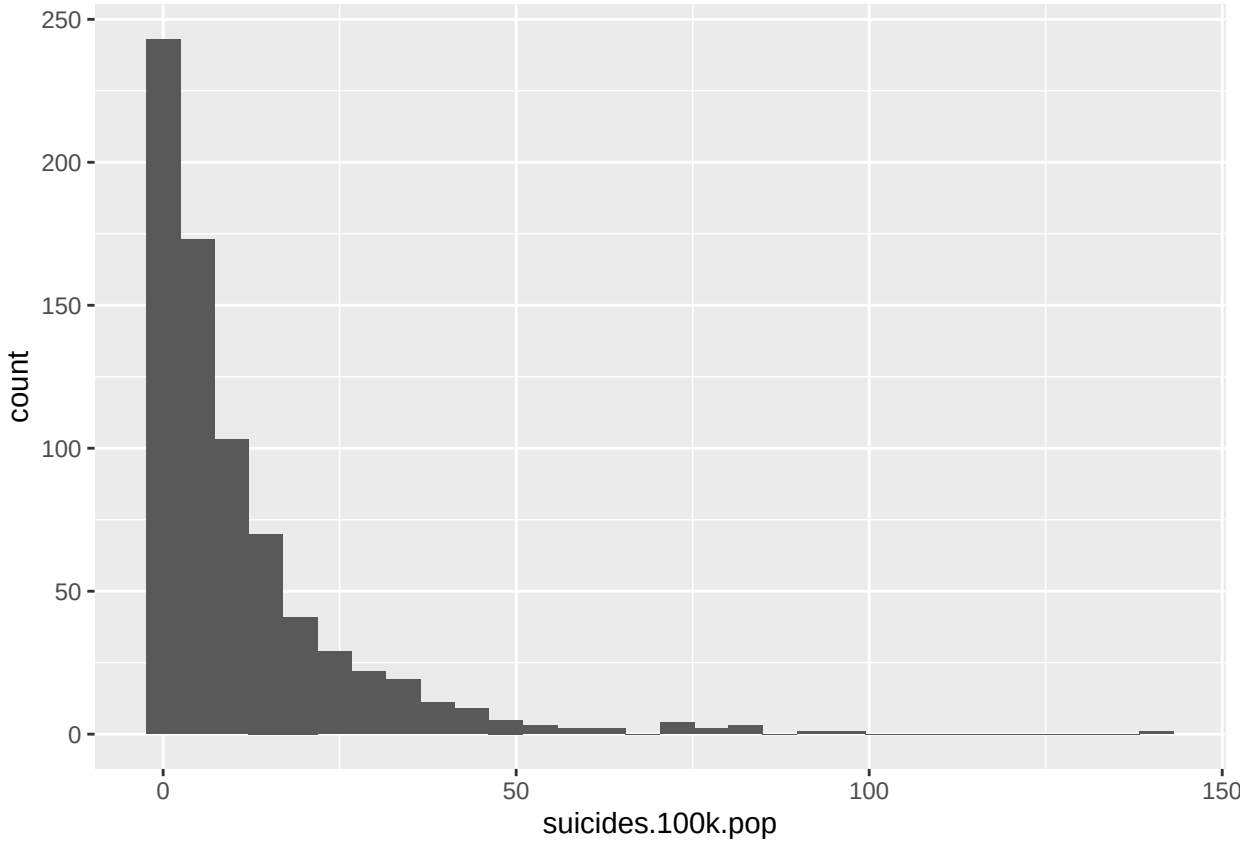
```
##
##  Anderson-Darling normality test
##
## data:  data$suicides.100k.pop
## A = 2569.3, p-value < 2.2e-16
```

```
ad.test(data$gdp_per_capita)
```

```
##
##  Anderson-Darling normality test
##
## data:  data$gdp_per_capita
## A = 1706.9, p-value < 2.2e-16
```

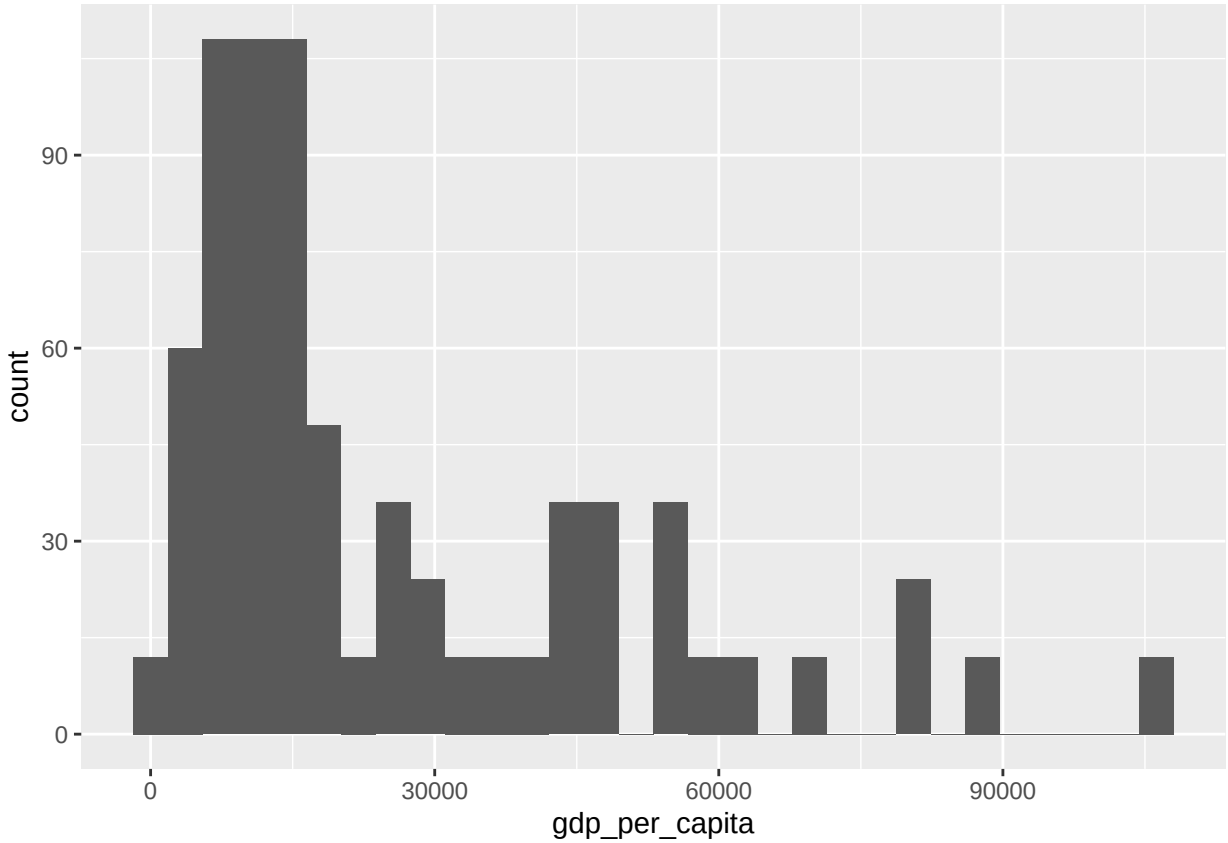
```
# Pouze za rok 2015, aby vysledek nebyl ovlivnen vyvojem HDP v letech
data2 <- data %>% filter(year==2015)

# Vykresleni grafu, abych se podivala, jestli jsou data normalniho rozdeleni
data2 %>% ggplot(aes(x=suicides.100k.pop)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
data2 %>% ggplot(aes(x=gdp_per_capita)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# Data jsou velmi skewed a opravdu nemuzu rict, ze jsou aspon trochu normalni

# Data nepochazi z normalniho rozdeleni -> neparametricky koeficient -> spearman
cor.test(data2$suicides.100k.pop, data2$gdp_per_capita, method = "spearman")
```

```
## Warning in cor.test.default(data2$suicides.100k.pop,
## data2$gdp_per_capita, : Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  data2$suicides.100k.pop and data2$gdp_per_capita
## S = 59655706, p-value = 0.0003449
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.130869
```

```
# Velmi nizka hladina korelace -> neexistujici vztah mezi GDP per capita a mirou sebevrazd
```

Data nepochází z normálního rozdělení a kvůli tomu je použit Spearmanův test. Vychází velmi nízká hladina korelace, což naznačuje, že asi neexistuje vztah mezi GDP per capita a mírou sebevražd (p-value naznačuje, že je to statisticky významné - musíme ovšem počítat s tím, že neparametrické testy jsou slabší).

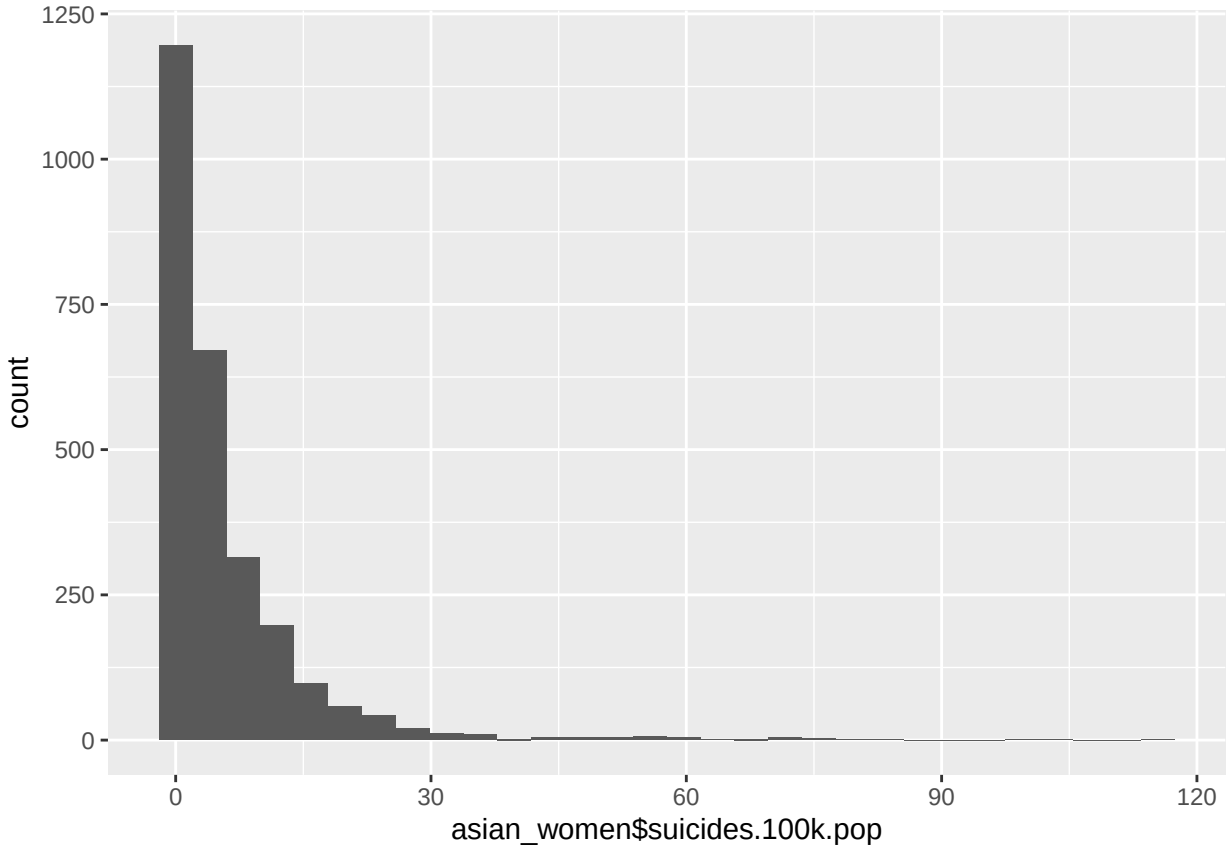
Hypotéza byla počítána za konkrétní rok. Důvodem je odproštění se od vývoje státu během let - dochází ke změně velké řady ukazatelů, které se ani nemusí nacházet v datasetu, ale výrazně ovlivňují parametry jako HDP, populace, atd.

H2: Ženy v Asii mají větší sklon k sebevraždám než ženy v jiných oblastech

Z histogramů distribuce je vidět, že asijské ženy mají míru sebevražd více rozloženou doprava, než ženy ze zbytku světa.

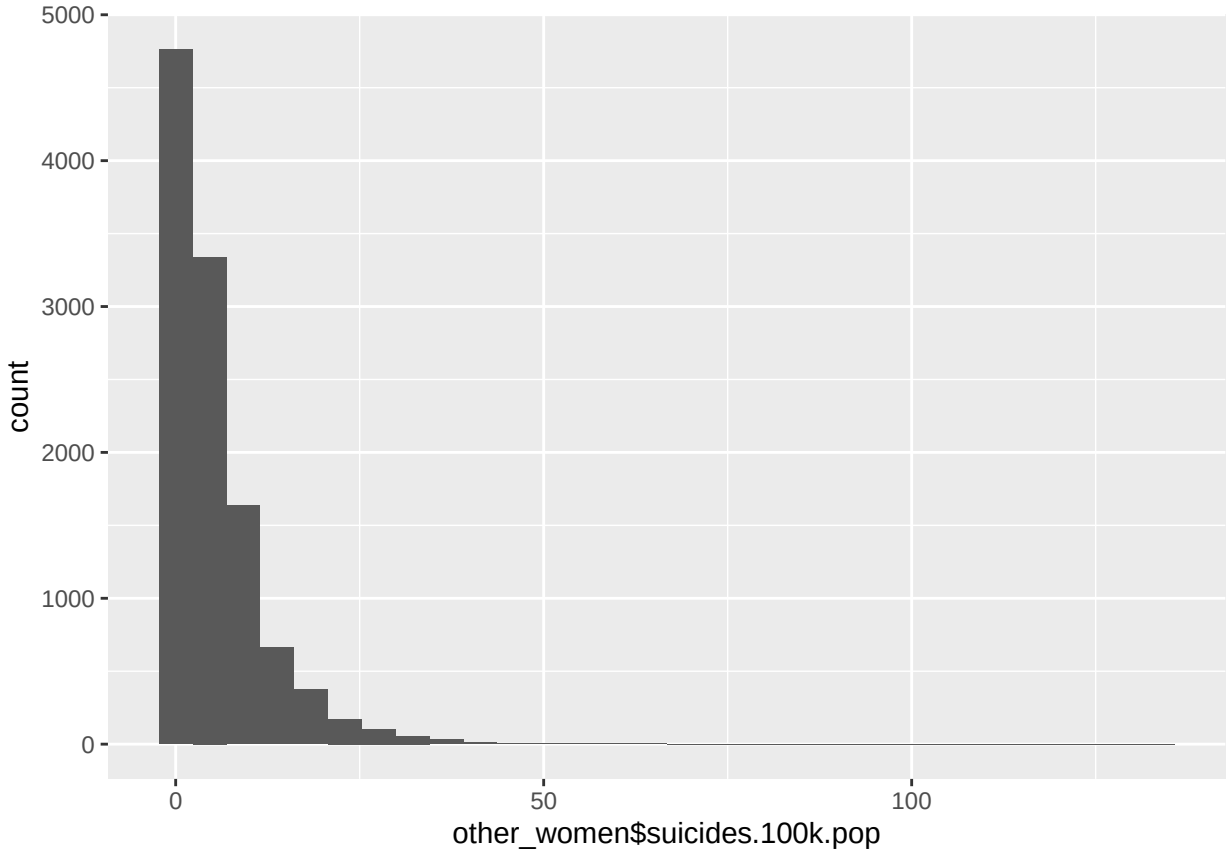
```
# 2 samples - zeny z Asie a zeny z jinych oblasti

asian_women <- data %>% filter(sex=="Female", continent=="Asia")
other_women <- data %>% filter(sex=="Female", continent!="Asia")

# Histogramy - rozlozeni
asian_women %>% ggplot(aes(x=asian_women$suicides.100k.pop)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
other_women %>% ggplot(aes(x=other_women$suicides.100k.pop)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# Shapiro test
shapiro.test(asian_women$suicides.100k.pop) # p-value je mensi - neni z normalniho rozdeleni
```

```
##
##  Shapiro-Wilk normality test
##
## data:  asian_women$suicides.100k.pop
## W = 0.58087, p-value < 2.2e-16
```

```
#shapiro.test(other_women$suicides.100k.pop) # tenhle sample je moc velky

# t-test je parametricky test, mel by byt pouzivan na normalne rozdelena data
# a tohle nejsou normalne rozdelena data
# ale pak ty means jsou normalne rozdeleny
t.test(asian_women$suicides.100k.pop, other_women$suicides.100k.pop)
```

```
##
##  Welch Two Sample t-test
##
## data:  asian_women$suicides.100k.pop and other_women$suicides.100k.pop
## t = 2.5718, df = 3311.4, p-value = 0.01016
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1199029 0.8892617
## sample estimates:
## mean of x mean of y
##  5.803829  5.299247
```

```r
# neparametrickym protejskem je Wilcoxuv test, ktery je slabsi, ale
# lze porusit predpoklady normality
# neparametricky test na means of idependent samples, not paired
wilcox.test(asian_women$suicides.100k.pop, other_women$suicides.100k.pop)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  asian_women$suicides.100k.pop and other_women$suicides.100k.pop
## W = 14586542, p-value = 0.1568
## alternative hypothesis: true location shift is not equal to 0
```

Data nemají normální rozdělení, a proto byly provedeny dva testy, t-test a jeho neparametrická alternativa, Wilcoxonův test. Wilcoxonův test tu je pro srovnání, jelikož t-test předpokládá normalní rozdělení dat. t-test nicméně stejně použijeme, protože střední hodnoty by nakonec měly být normálně rozdělené. V t-testu vyšlo na hladině významnosti 0.05, že zamítáme, že střední hodnoty skupin jsou stejné, tedy nezamítame hypotézu, že asijské ženy mají vyšší sklon k sebevraždám.

Wilcoxonův test vyšel s p-value vyšší než hladina významnosti, a tedy nezamítáme, že střední hodnoty skupin jsou stejné. Testy sice říkají přesný opak, ale střední hodnoty se liší pouze o kousek (ale asijské ženy mají stále vyšší střední hodnotu). Proto je možné, že Wilcoxonův test považuje takový rozdíl za nevýznamný. Neparametrické testy jsou taky slabší než parametrické.

H3: Míra sebevraždy ve věkové skupině 15-24 je vyšší v Asii než jinde

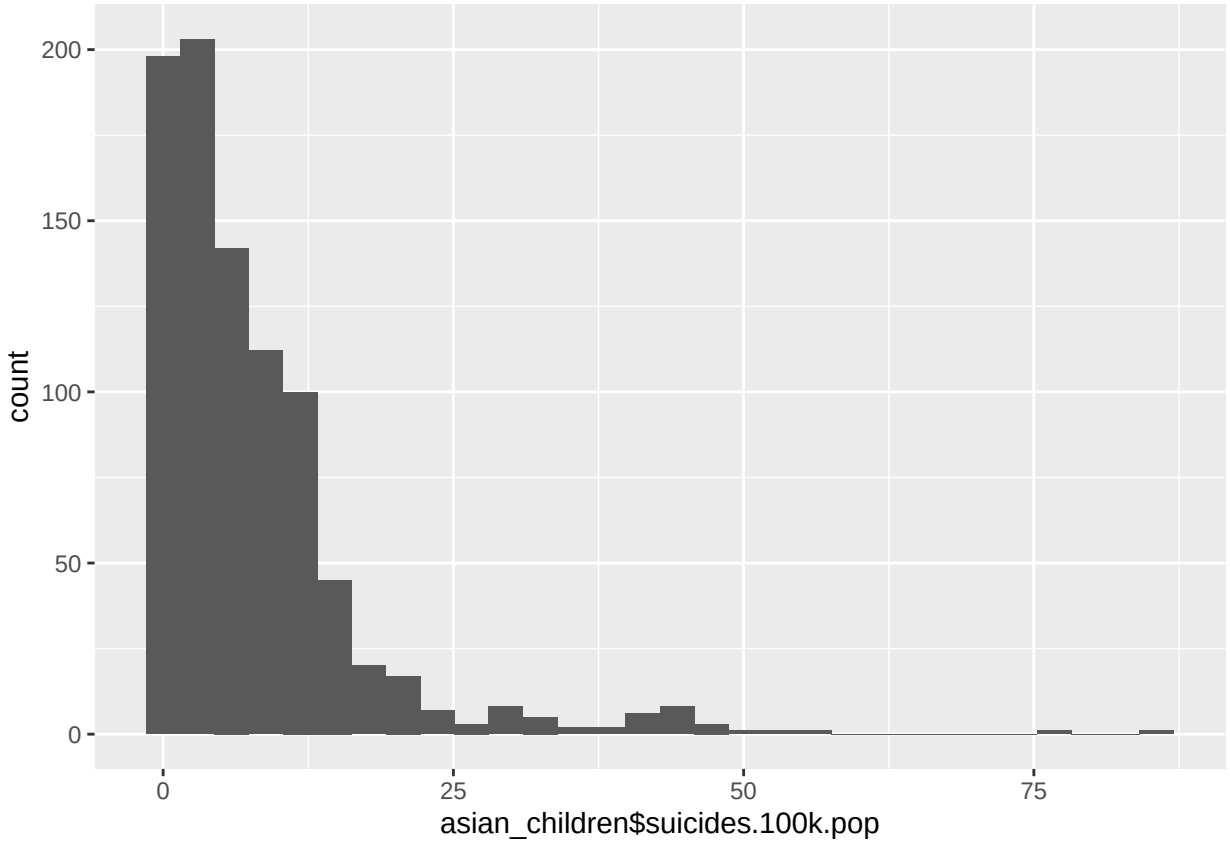
Provádí se ty samé testy jako pro H2. Opět vidíme, že histogram pro asijské děti je více rozložený.

```r
asian_children <- data %>% filter(age=="15-24", continent=="Asia")
other_children <- data %>% filter(sex=="Female", continent!="Asia")

shapiro.test(asian_children$suicides.100k.pop) # p-value je mensi - neni z normalniho rozdeleni
```

```
##
##  Shapiro-Wilk normality test
##
## data:  asian_children$suicides.100k.pop
## W = 0.70888, p-value < 2.2e-16
```

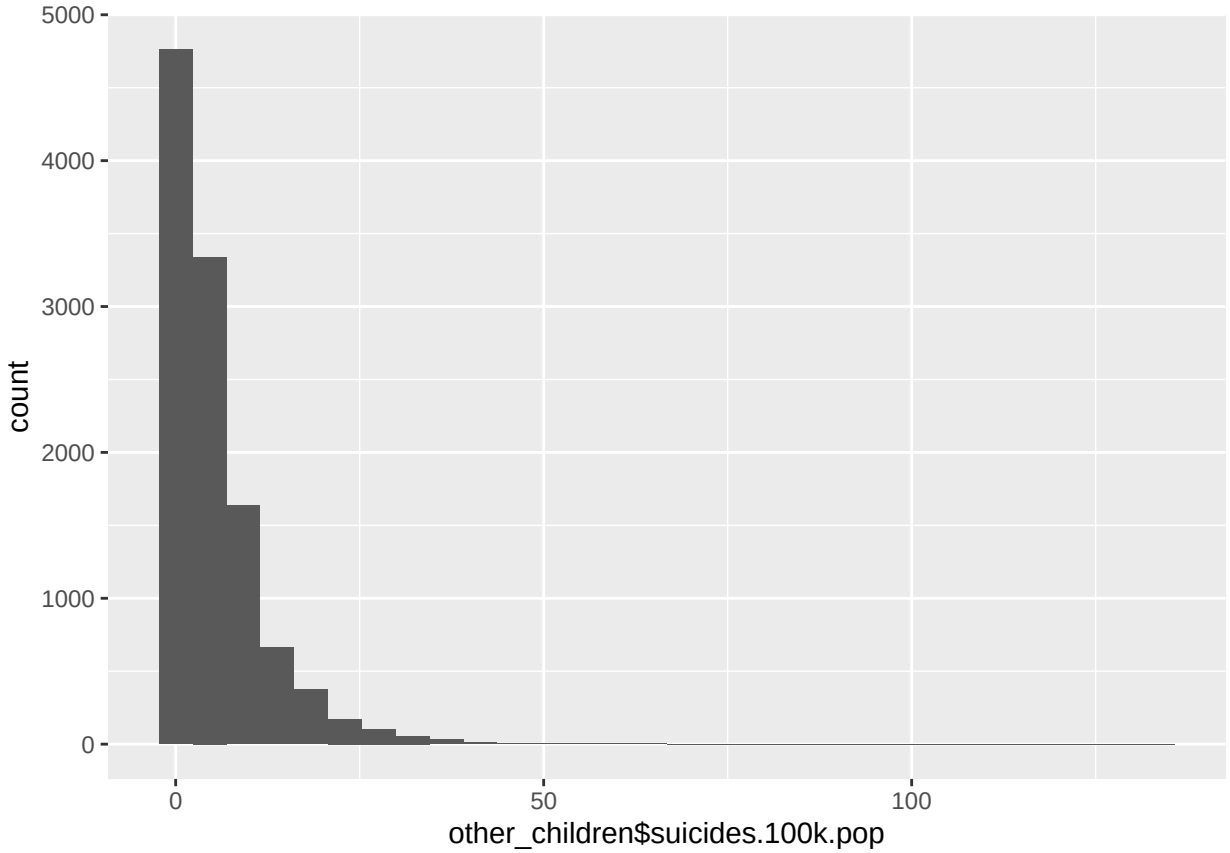
```r
# Histogram
asian_children %>% ggplot(aes(x=asian_children$suicides.100k.pop)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
other_children %>% ggplot(aes(x=other_children$suicides.100k.pop)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
t.test(asian_children$suicides.100k.pop, other_children$suicides.100k.pop)
```

```
##
##  Welch Two Sample t-test
##
## data:  asian_children$suicides.100k.pop and other_children$suicides.100k.pop
## t = 7.6498, df = 959.5, p-value = 4.886e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.821614 3.078716
## sample estimates:
## mean of x mean of y
##  7.749412  5.299247
```

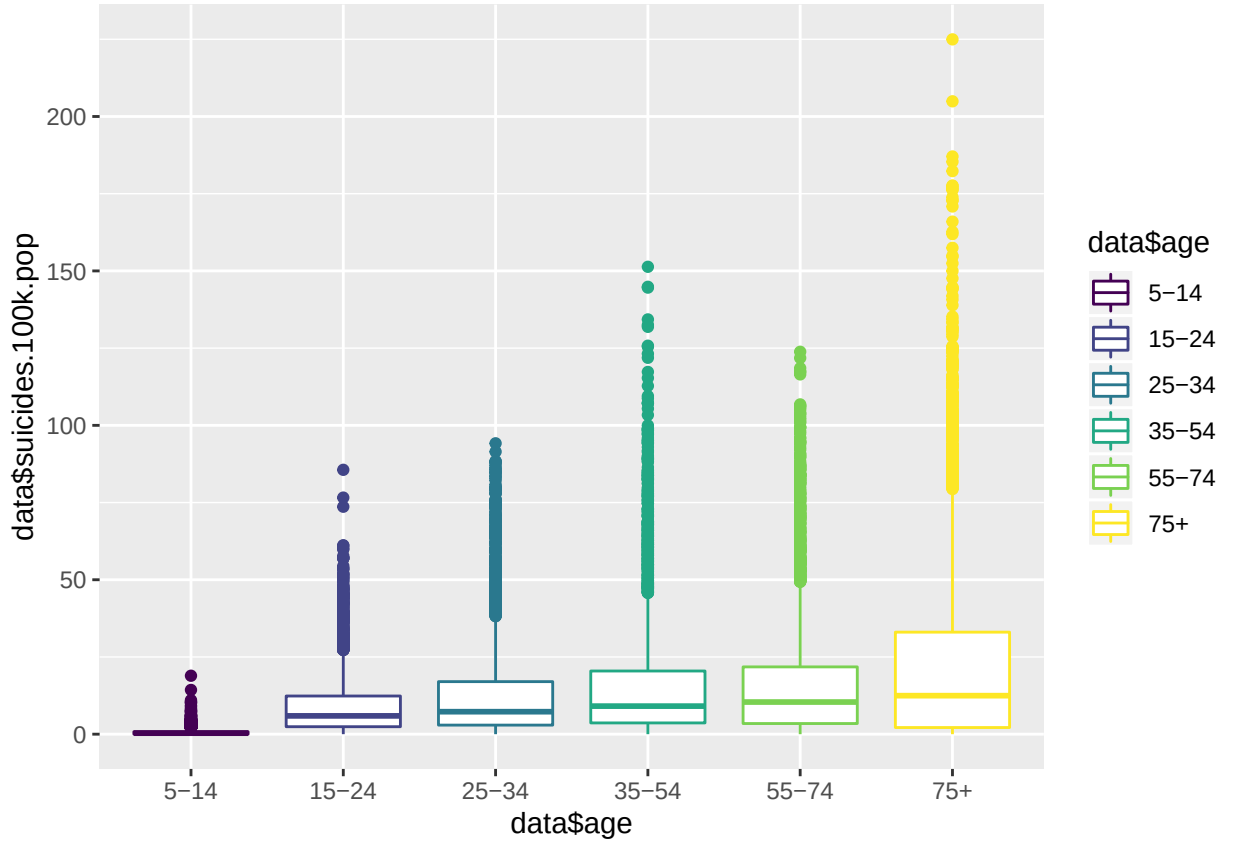
```
# neparametricky test na means of idependent samples, not paired
wilcox.test(asian_children$suicides.100k.pop, other_children$suicides.100k.pop)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  asian_children$suicides.100k.pop and other_children$suicides.100k.pop
## W = 5969197, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Už z grafu je znát větší rozdíl než v předešlé hypotéze. Zde lze vyčíst, že se střední hodnoty liší o dost více než u asijských žen. Zde už se oba testy shodly a zamítají, že by střední hodnoty mezi skupinami byly stejné a nezamítáme, že se v Asii tato věková skupina zabíjí více než jinde.

H4: Střední hodnota se nemění mezi věkovými skupinami

```
data_by_age <- split(data, data$age)

# cute boxplot, abychom videli, ze mame asi pravdu
ggplot(data, aes(x=data$age, y=data$suicides.100k.pop, color=data$age)) + geom_boxplot()
```




```
# ty means se moc nehybou, takze mame asi pravdu

# udelej kruskalluv test, protoze anova assumptions were not met
# ale neni tak stronk :(
kruskal.test(lapply(split(data, data$age), function(x) { x$suicides.100k.pop }))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  lapply(split(data, data$age), function(x) {    x$suicides.100k.pop})
## Kruskal-Wallis chi-squared = 6997, df = 5, p-value < 2.2e-16
```

```
# p-value je mensi nez hladina vyznamnosti
```

Byl použit Kruskal-Wallis test, který je neparametrickou alternativou k anově (protože nebyly splněny předpoklady normality). p-value je menší než hladina významnosti 0.05, tedy zamítáme, že střední hodnoty skupin jsou stejné.

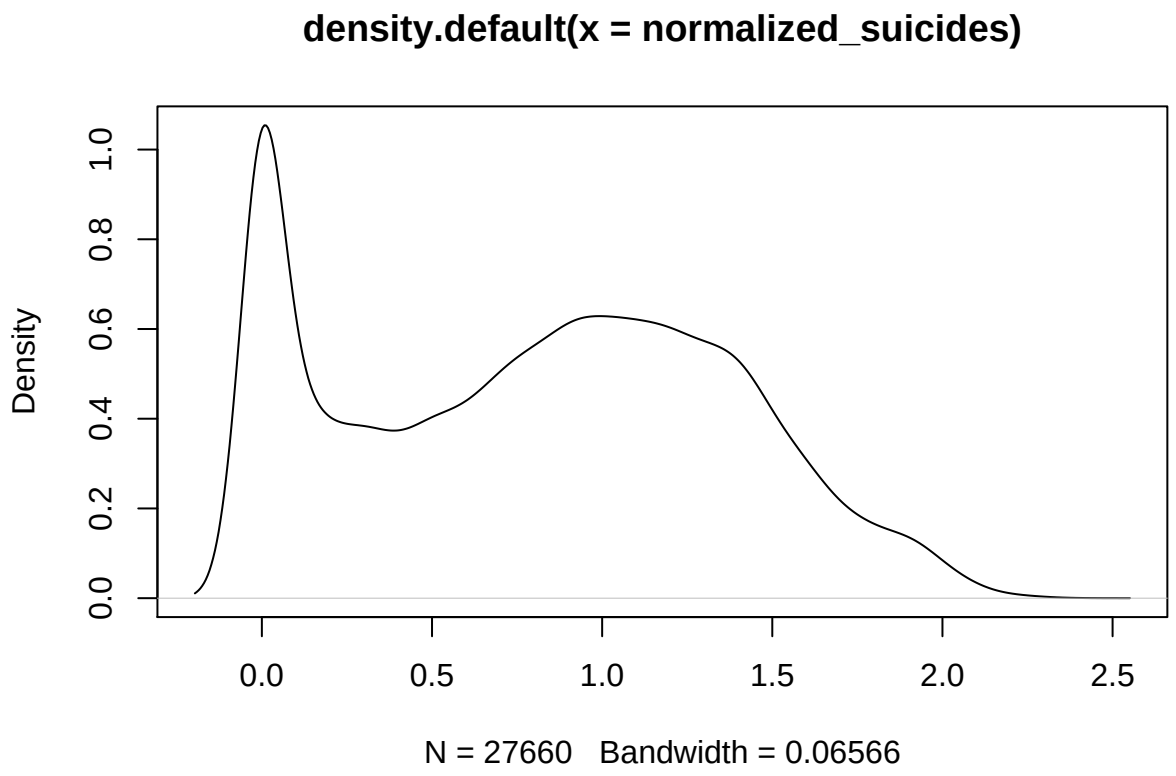
Pre-process

```
#####
# Selekce atributu k pouziti v modelu + shuffle + stratifikovany sampling + factor reduction
#####

## Prevod na klasifikacni ulohu pro nektere alg ##
# Prevod miry sebevrazd na intervaly, vytvoreno 5 binu (kvantily 0 - 100, krok po 20 %)
data$bin <- data$suicides.100k.pop %>%
  cut(include.lowest=TRUE, breaks=quantile(data$suicides.100k.pop, probs = seq(0,1,0.2)))
# data$bin <- discretize(data$suicides.100k.pop, method="frequency", breaks=5)

# Pocatecni manualni vyber faktoru + classes
train_data <- data %>% select(country, age, year, sex, generation, gdp_per_capita, bin, suicides.100k.p

# Remove rows with missing data
train_data <- na.omit(train_data)

# Stratifikovany sampling
train_data <- train_data[sample(nrow(train_data)),] %>% group_by(bin) %>% filter(row_number() <= 200) %
train_classes <- train_data$bin
train_classes_cont <- train_data$suicides.100k.pop

# Odstraneni classes z trenovacich dat
train_data <- train_data %>% select(-c(bin, suicides.100k.pop))

# Odstraneni zero a near-zero variance promennych
nzv <- nearZeroVar(train_data)
if (length(nzv) != 0) {
  train_data <- train_data[, -nzv]
}

# Odstraneni vysoce korelovanych promennych
hcor <- findCorrelation(cor(mutate_if(train_data, is.factor, as.numeric)), cutoff=.75)
if (length(hcor) != 0) {
  train_data <- train_data[, -hcor]
}

# Odstraneni linearne zavislych promennych (linearnich kombinaci)
lin_comb <- findLinearCombos(mutate_if(train_data, is.factor, as.numeric))$remove
if (length(lin_comb) != 0) {
  train_data <- train_data[, -lin_comb]
}

# Prevod z tibble na data.frame, tibble obcas dela problemy caret algoritmum
train_data <- as.data.frame(train_data)

# Prevedeni vsech atributu na numericke pro algoritmy, ktere to vyzaduji
train_data_numeric <- train_data %>% mutate_if(is.factor, as.numeric) %>% as.data.frame()


# VOLITELNE: centrovani a normalizace, lze i PCA
# !!!!!!!!!
# !!! POZOR - pri pouziti skalovani pote neni mozno predikovat neznamy neskalovany zaznam !!!
```

```r
# !!!!!!!!!
# train_data <- mutate_if(train_data, is.numeric, scale)  # pouze normalizace, ale jednodussi usage
preproc <- preProcess(train_data, method=c("center", "scale"))  # pro PCA staci pridat "pca", ale ztrat
train_data <- predict(preproc, train_data)

# Transformace dat
library(e1071)
skewness(data$suicides.100k.pop) # skewness je 2.9 (highly-skewed)
```

```
## [1] 2.964852
```

```r
normalized_suicides <- log10(data$suicides.100k.pop + 1) # +1, protoze log(0) je undefined
plot(density(normalized_suicides))
```




```r
# po plotovani density jsme zjistili, ze transformace vubec nepomohla >:(
```

Trenovani a testovani modelu

Byly natrénovány jak klasifikační, tak regresní modely. Vyskytují se tam ty, které se probíraly v rámci kurzu (qda, lda, gam, lm, lasso), tak i složitější jako neuronové sítě, random forest atd. Klasifikační algoritmy jsou mezi sebou porovnávány v rámci accuracy, regresní v rámci RMSE. K otestování se používá 10-fold cross-validation. Kromě toho, jak se předzpracovaly data v předešlé části ještě některé z algoritmů dělají

vnitřně vlastní feature selection. Vzhledem k velikosti datasetu bylo potřeba kvůli některým z algoritmů udělat sampling, protože by mi bouchnul počítač (což ale ovlivňuje výkon algoritmů).

Klasifikacni modely

```
models <- c("gam", "lda2", "ctree", "nnet", "xgbTree", "rf", "naive_bayes", "qda")
needs_numeric <- c("gam", "lda2", "xgbTree", "rf", "naive_bayes", "qda")
results <- c()

for (i in 1:length(models)) {

  if (models[i] %in% needs_numeric) {
    model_data <- train_data_numeric
  } else {
    model_data <- train_data
  }
  if (models[i]=="nnet") {
    model <- train(model_data, train_classes, method=models[i], trControl = trainControl(method="cv", nu
  } else {
    model <- train(model_data, train_classes, method=models[i], trControl = trainControl(method="cv", nu
  }
  # ggplot(varImp(model)) # nejvetsi faktory: sex & age, obcas country
  results <- rbind(results, c(models[i], mean(model$resample$Accuracy)))
}
```

```
## Loading required package: mgcv
```

```
## Loading required package: nlme
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
## This is mgcv 1.8-28. For overview type 'help("mgcv-package")'.
```

```
colnames(results) <- c("Model", "Accuracy")
results[order(results[,2], decreasing = TRUE),] # serazeno od nejvyssi acc
```

```
##      Model         Accuracy
## [1,] "nnet"        "0.633"
## [2,] "rf"          "0.58"
## [3,] "xgbTree"     "0.578"
## [4,] "ctree"       "0.577"
## [5,] "qda"         "0.452"
## [6,] "naive_bayes" "0.44"
## [7,] "lda2"        "0.411"
## [8,] "gam"         "0.228"
```

```
# # nejlepsi klasifikacni vysledky
# class_res <- readRDS("classification_results")
```

pozn: Do neuronky a do ctree se dávaly jako jediné kategorické proměnné, do zbytku se dávaly převedené na čísla.

Nejhůř dopadly modely gam, lda což se dalo očekávat vzhledem k jednoduchosti algoritmů. Dále na tom nebyl tak dobře naivní bayes, který má silný naivní předpoklad, že jsou features nezávislé, což ho dost často omezuje. Nejlépe na tom byla neuronová síť, poté random forest a pak extreme gradient boosting trees.
Regresni modely

```
models <- c("lm", "gam", "nnet", "xgbTree", "rf", "lasso", "svmPoly")
results_reg <- c()

for (i in 1:length(models)) {
  if (models[i] == "nnet") {
    model <- train(train_data_numeric, train_classes_cont, method=models[i], trControl = trainControl(m
  } else {
    model <- train(train_data_numeric, train_classes_cont, method=models[i], trControl = trainControl(m
  }
  results_reg <- rbind(results_reg, c(models[i], mean(model$resample$RMSE)))  # pripadne Rsquared nebo
}
```

```
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info =
## trainInfo, : There were missing values in resampled performance measures.
```

```
colnames(results_reg) <- c("Model", "RMSE")
results_reg[order(results_reg[,2]),] # serazeno od nejnizsiho rmse
```

```
##      Model     RMSE
## [1,] "xgbTree" "12.1135666879483"
## [2,] "rf"      "12.3930700884383"
## [3,] "svmPoly" "14.858751864138"
## [4,] "gam"     "15.3175498035179"
## [5,] "lm"      "15.3449132809182"
## [6,] "lasso"   "15.4120238305522"
## [7,] "nnet"    "21.5156223895499"
```

```
# reg_res <- readRDS("regression_results")
# nejlepsi regresni vysledky

# dostupne modely: names(getModelInfo())
# modely lze ulozit a nahrat pomoci funkci saveRDS a readRDS
```

Je zajímavé, že neuronka měla nejvyšší accuracy u klasifikace, ale největší RMSE u regrese. Měli bychom ovšem počítat s tím, že převedením úlohy rozdělením do binů se možná snížila přesnost a accuracy nemusí být úplně vypovídající. Použití kategorické/numerické proměnné také mohlo ovlivnit výkon. Dle očekávání jednodušší algoritmy neměly příliš dobré výsledky. Algoritmy random trees a xgb trees měly nejlepší výsledky, což není překvapením, jelikož se z toho důvodu velmi často v praxi používají. Při odendání některých proměnných se RMSE i accuracy snížila - v regresní části si často vedl nejlépe SVM (zřejmě kvůli tomu, že dobře generalizuje).

Závěrečná diskuze

Tento dataset byl zajímavý z hlediska zjištění nějakých krizových skupin, zemí, ovlivnění krize na míru sebevražd atd. V rešerši jsem ovšem ještě nenarazila na výzkum, který by tento jev dokázal spolehlivě predikovat, i když se jednalo přímo o predikci sebevraždy podle množiny nějakých emocí jednoho člověka.

Mnoho článků (spíše psychologického rázu) souhlasí s problémem vysoké míry sebevražd žen i dětí-adolescentů v Asii. To je způsobeno přetrváváním velmi silného patriarchismu a náporu na vzdělání/jiné skillsety u asijských

dětí. Podle WHO tabulky z roku 2016-2000 (na 100k obyvatel) je suicide female rate v jihovýchodní Asii nejvyšší ze všech zmíněných regionů (http://apps.who.int/gho/data/view.main.MHSUICIDEREGv?lang=en). Podle WHO se 79 % sebevražd dějí v low-middle-income zemí, což vyvrátilo myšlenku, dle které jsem dělala H1 (https://www.who.int/news-room/fact-sheets/detail/suicide).

Z tohoto datasetu nejde zjistit nic ohledně příčiny vysoké míry sebevražd. Také by bylo zajímavé mít nějakou úroveň vzdělání v dané zemi, kriminalitu, státní uspořádání, obecně nějaké indexy, např. životní úrovně.