Assignment 3 - Our own data analysis

Akaran Sivakumar (AS), Lelia Rønnow (LR), Shiraz Ben-Shoshan (SBS)

01-12-2024

Setup

Loading packages and data

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.3     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.3     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflic
ts to become errors
```

```
#install.packages("countrycode")
library(countrycode)
```

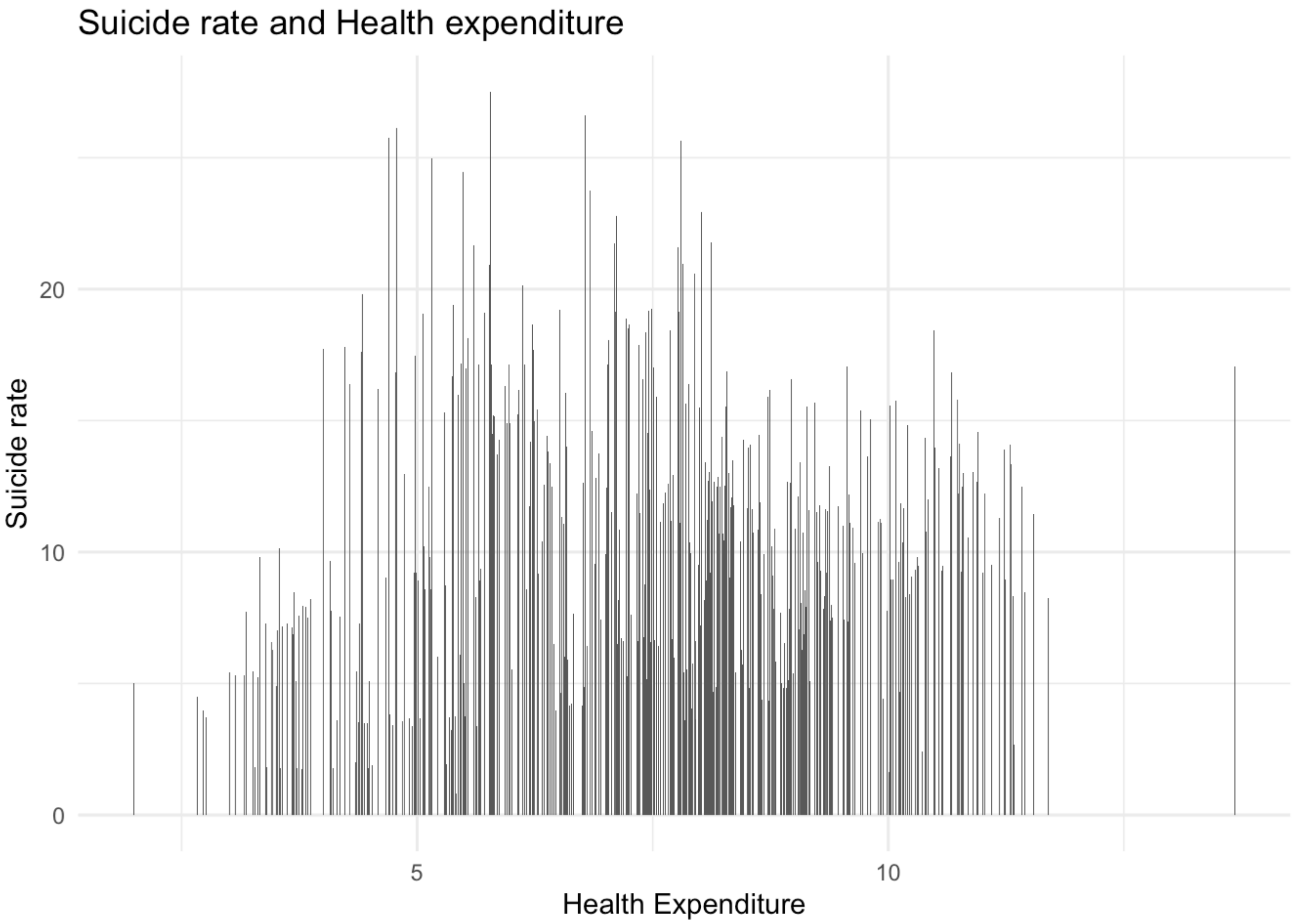
```
# Setting working directory
#setwd("~/Desktop/AarhusUni/Semester5/IntroCultDataSci")

# Loading Data
WHO_MHExp_and_Deaths <- read_csv("archive/WHO_MHExp_and_Deaths.csv")
```

```
## Rows: 531 Columns: 9
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (1): Country_Name
## dbl (8): Year, Population, Deaths_All_Types, Deaths_Suicides, HExp_Pctage_Y,...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

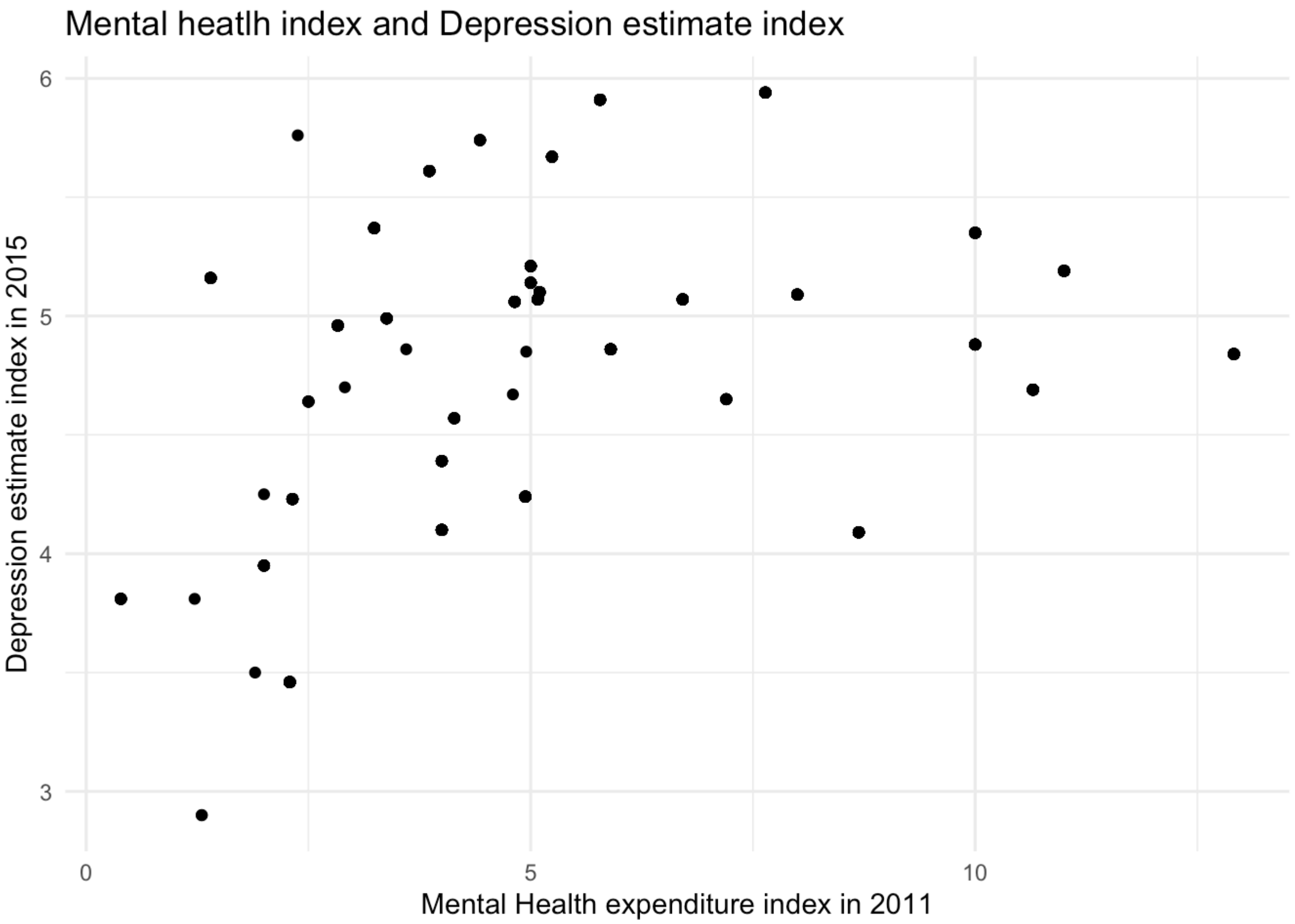
Preliminary data visualisations (AS)

```
WHO_MHExp_and_Deaths %>%
ggplot(aes(x = HExp_Pctage_Y,
y = Suicide_p100)) +
geom_bar(stat = "summary", fun = "mean") +
theme_minimal() +
labs(x = "Health Expenditure", y = "Suicide rate") +
ggtitle("Suicide rate and Health expenditure")
```



The plot above does not show any clear patterns, so we will now conduct further exploratory plots to discover hidden patterns.

```
WHO_MHExp_and_Deaths %>%
ggplot(aes(x = MHExp_Pctage_2011,
y = Dep_Num_2015)) +
geom_point()+
theme_minimal() +
labs(x = "Mental Health expenditure index in 2011", y = "Depression estimate index in
2015") +
ggtitle("Mental heatlh index and Depression estimate index")
```



## Data mutation and further visualisations (LR)

```
# Load the package
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

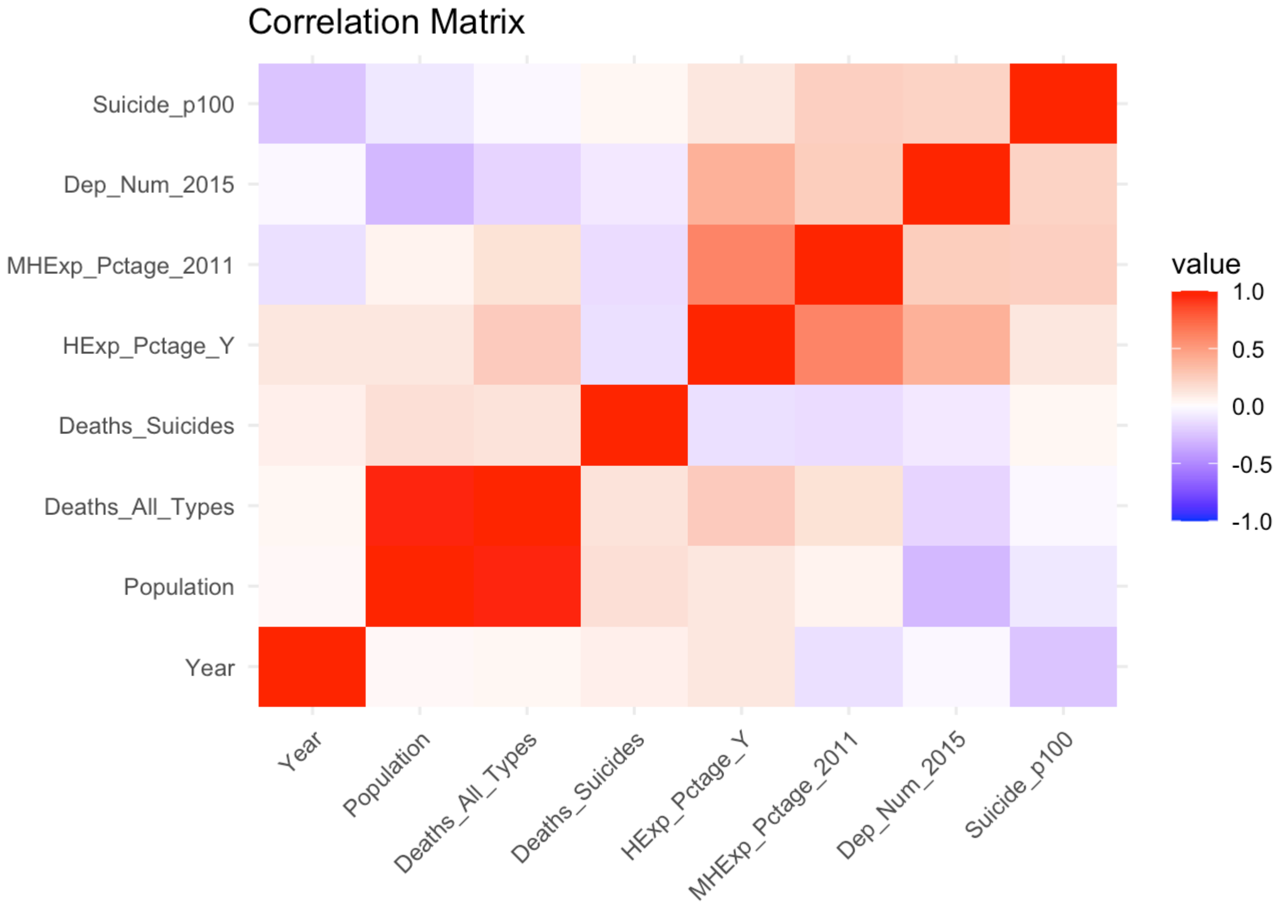
```
# Converting nominal deaths to a ratio of per 100k
WHO_MHExp_and_Deaths$Deaths_per_100k <- (WHO_MHExp_and_Deaths$Deaths_All_Types / WHO_MHExp_and_Deaths$Population) * 100000

# Convert country names to continent names
WHO_MHExp_and_Deaths$continents <- countrycode(WHO_MHExp_and_Deaths$Country_Name, origin = "country.name", destination = "continent")

# Remove the first and last columns
df_filtered <- WHO_MHExp_and_Deaths[, -c(1, ncol(WHO_MHExp_and_Deaths)-1)]

numeric_df <- df_filtered[, sapply(df_filtered, is.numeric)]
# Calculate the correlation matrix and convert it to a long format
cor_matrix <- cor(numeric_df, use = "complete.obs")
cor_long <- melt(cor_matrix)

# Correlation matrix plot
ggplot(cor_long, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(-1, 1)) +
  theme_minimal() +
  labs(title = "Correlation Matrix", x = "", y = "") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
```



Given the results in the correlation matrix, we see a semi-strong correlation between the mental health expenditure index and the health expenditure percentage. We have, therefore, decided not to include mental health expenditure index in any models, to avoid multicolinearity.

Modelling (AS)

```
# Creating the first and most simple model
simple_model <- lm(Suicide_p100 ~ HExp_Pctage_Y, WHO_MHExp_and_Deaths)

summary(simple_model)
```

```
##
## Call:
## lm(formula = Suicide_p100 ~ HExp_Pctage_Y, data = WHO_MHExp_and_Deaths)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8939 -4.2753 -0.6807  3.4656 19.4103
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     8.6435     0.8846   9.771   <2e-16 ***
## HExp_Pctage_Y   0.2898     0.1128   2.569   0.0105 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.463 on 529 degrees of freedom
## Multiple R-squared:  0.01232,    Adjusted R-squared:  0.01045
## F-statistic: 6.599 on 1 and 529 DF,  p-value: 0.01048
```

Health expenditure percentage seems to have an effect on the suicide rate, but there is still a lot of unexplained variance.

Adding Random Effects (SBS)

```
#install.packages("lmerTest")
#install.packages("MuMIn")
library(MuMIn)
```

```
## Warning: package 'MuMIn' was built under R version 4.3.3
```

```
library(lmerTest)
```

```
## Loading required package: lme4
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
##
## Attaching package: 'lmerTest'
```

```
## The following object is masked from 'package:lme4':
##
##     lmer
```

```
## The following object is masked from 'package:stats':
##
##     step
```

```
# Adding country as a random intercept
mixed_model <- lmer(Suicide_p100 ~ HExp_Pctage_Y+(1|Country_Name), WHO_MHExp_and_Deat
hs)
summary(mixed_model)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Suicide_p100 ~ HExp_Pctage_Y + (1 | Country_Name)
##    Data: WHO_MHExp_and_Deaths
##
## REML criterion at convergence: 2402.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.4558 -0.3558 -0.0301  0.3961  4.7487
##
## Random effects:
##  Groups       Name        Variance Std.Dev.
##  Country_Name (Intercept) 28.962   5.382
##  Residual                  3.885   1.971
## Number of obs: 531, groups:  Country_Name, 42
##
## Fixed effects:
##               Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)    12.5161     1.1110 108.2106  11.266  < 2e-16 ***
## HExp_Pctage_Y  -0.3750     0.1012 528.0301  -3.705 0.000234 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr)
## HExp_Pctg_Y -0.650
```

```
# Calculating the marginal and conditional R-sqaured
r_squared <- r.squaredGLMM(mixed_model)
print(r_squared)
```

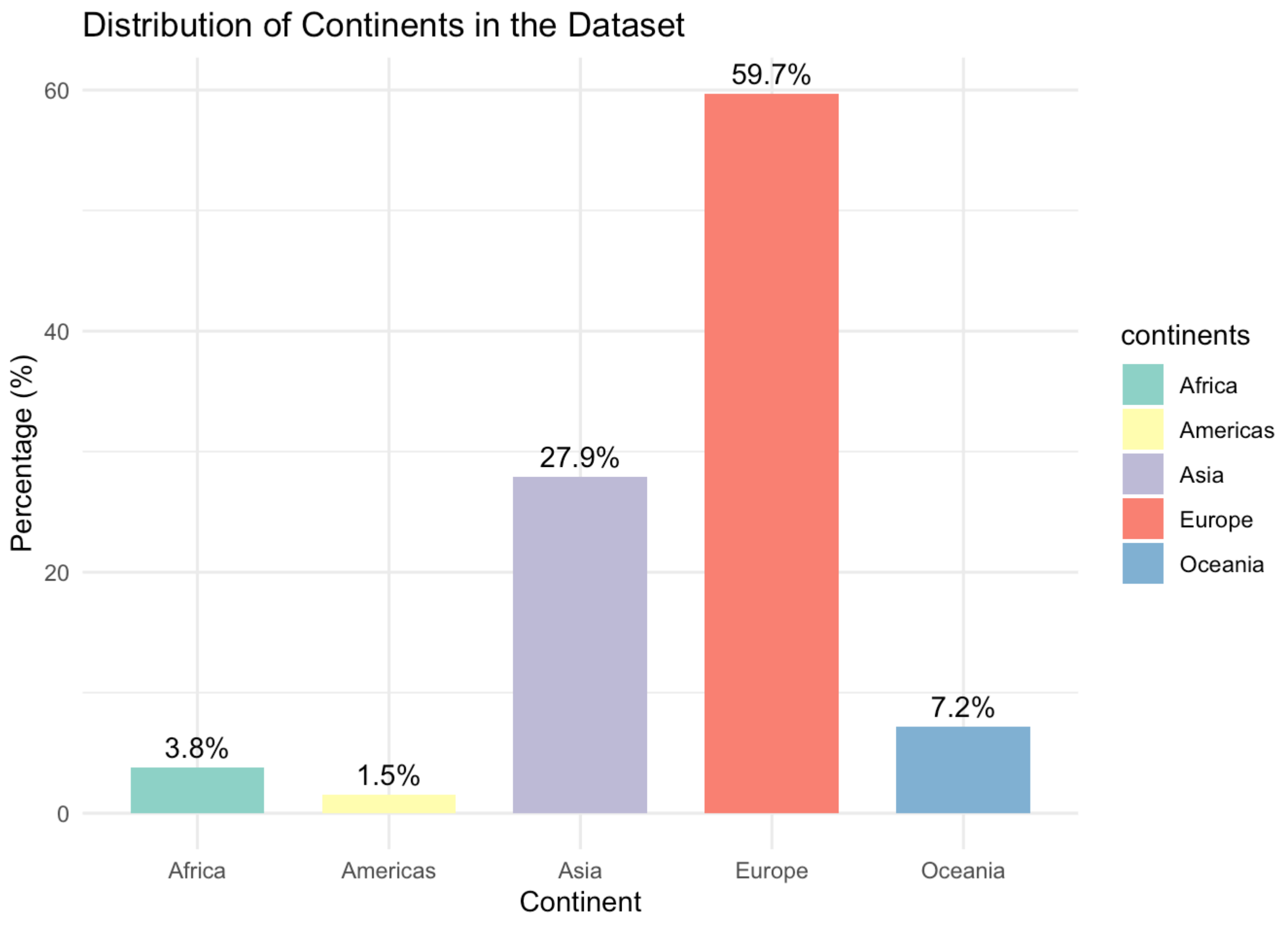
```
##             R2m      R2c
## [1,] 0.01859253 0.883929
```

Adding a random intercept for country has made the effect of health expenditure percentage more clear/significant, thus explaining 88% of the data.

```
# Adding continent as a column
data <- WHO_MHExp_and_Deaths %>%
  count(continents)

# Counting the occurrences of each continent and calculating the percentage
data <- WHO_MHExp_and_Deaths %>%
  count(continents) %>%
  mutate(percentage = n / sum(n) * 100)

# Creating a bar plot to visualise the imbalance in the data
ggplot(data, aes(x = continents, y = percentage, fill = continents)) +
  geom_bar(stat = "identity", width = 0.7) +
  labs(title = "Distribution of Continents in the Dataset",
       x = "Continent",
       y = "Percentage (%)") +
  scale_fill_brewer(palette = "Set3") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            vjust = -0.5) +
  theme_minimal()
```

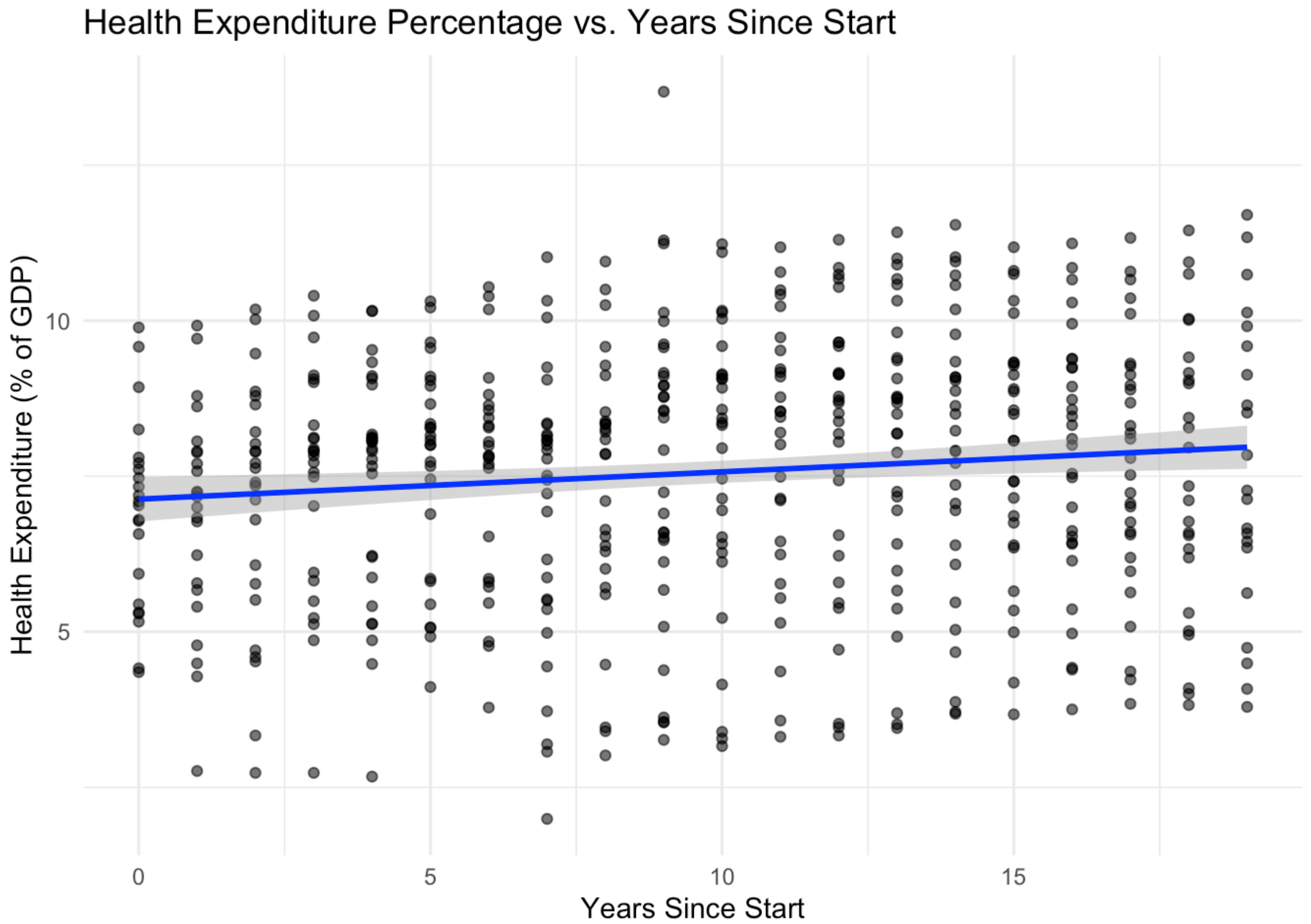


The bar plot shows that Europe is very prominent in the data, compared to the other countries, thus telling us that we should take the results with a grain of salt.

```
# recalculating the "Year" column to use in models
WHO_MHExp_and_Deaths$Year_since_start <- WHO_MHExp_and_Deaths$Year-2000

# Creating a plot with the Year as a predictor
ggplot(WHO_MHExp_and_Deaths, aes(x = Year_since_start, y = HExp_Pctage_Y)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", color = "blue", se = TRUE) +
  labs(
    title = "Health Expenditure Percentage vs. Years Since Start",
    x = "Years Since Start",
    y = "Health Expenditure (% of GDP)"
  ) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Visual inspection shows a slightly positive slope between the year and the health expenditure percentage.

Adding more Predictors (LR)

```
# Adding continent as an interaction term
continent_simple_model <- lm(Suicide_p100 ~ HExp_Pctage_Y+HExp_Pctage_Y:continents ,W
HO_MHExp_and_Deaths)
summary(continent_simple_model)
```

```
##
## Call:
## lm(formula = Suicide_p100 ~ HExp_Pctage_Y + HExp_Pctage_Y:continents,
##     data = WHO_MHExp_and_Deaths)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7478 -3.7308 -0.4028  3.1358 17.7829
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       11.6190     0.9188  12.645  < 2e-16 ***
## HExp_Pctage_Y                     -1.6077     0.2971  -5.411 9.53e-08 ***
## HExp_Pctage_Y:continentsAmericas   1.3934     0.3231   4.313 1.93e-05 ***
## HExp_Pctage_Y:continentsAsia       1.2286     0.2477   4.961 9.49e-07 ***
## HExp_Pctage_Y:continentsEurope     1.6497     0.2479   6.654 7.17e-11 ***
## HExp_Pctage_Y:continentsOceania    1.5840     0.2648   5.981 4.11e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.172 on 525 degrees of freedom
## Multiple R-squared:  0.1212, Adjusted R-squared:  0.1128
## F-statistic: 14.48 on 5 and 525 DF,  p-value: 2.672e-13
```

```
# Calculating the marginal and conditional R-sqaured
r_squared <- r.squaredGLMM(continent_simple_model)
print(r_squared)
```

```
##            R2m       R2c
## [1,] 0.1201577 0.1201577
```

```
# Adding continent as both an interaction term and as a fixed predictor
continent_model <- lm(Suicide_p100 ~ HExp_Pctage_Y*continents ,WHO_MHExp_and_Deaths)
summary(continent_model)
```

```
##
## Call:
## lm(formula = Suicide_p100 ~ HExp_Pctage_Y * continents, data = WHO_MHExp_and_Death
s)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4643 -3.2468  0.1032  2.5011 13.6578
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       3.600356  11.861307   0.304    0.762
## HExp_Pctage_Y                     0.001242   2.389326   0.001    1.000
## continentsAmericas                4.840862  25.063565   0.193    0.847
## continentsAsia                    0.262793  11.911741   0.022    0.982
## continentsEurope                 19.273319  11.936700   1.615    0.107
## continentsOceania                13.071291  12.655417   1.033    0.302
## HExp_Pctage_Y:continentsAmericas  0.147090   3.477538   0.042    0.966
## HExp_Pctage_Y:continentsAsia      0.723151   2.395111   0.302    0.763
## HExp_Pctage_Y:continentsEurope   -1.269896   2.394614  -0.530    0.596
## HExp_Pctage_Y:continentsOceania  -0.619527   2.446758  -0.253    0.800
##
## Residual standard error: 4.672 on 521 degrees of freedom
## Multiple R-squared:  0.2885, Adjusted R-squared:  0.2762
## F-statistic: 23.48 on 9 and 521 DF,  p-value: < 2.2e-16
```

```
# Calculating the marginal and conditional R-sqaured
r_squared <- r.squaredGLMM(continent_model)
print(r_squared)
```

```
##            R2m       R2c
## [1,] 0.2850245 0.2850245
```

The explainability of both models do not seem as strong as the models with countries as random intercepts.

Model Selection (SBS)

```
#using AIC to select the best and simplest model
AIC(simple_model, continent_model,continent_simple_model, mixed_model)
```

```
##                        df      AIC
## simple_model            3 3314.114
## continent_model        11 3155.935
## continent_simple_model  7 3260.115
## mixed_model             4 2410.244
```

We choose the “mixed_model” as it has best explainability while having the lowest AIC score.