Task3 – Happiness vs Mood of Songs

2024-11-18

```
df <- read.csv("happiness_w_mood_R/happiness_w_mood_R.csv")
df$developed_country = as.logical(df$developed_country)
df[is.na(df)] <- 0
```

Data Exploration

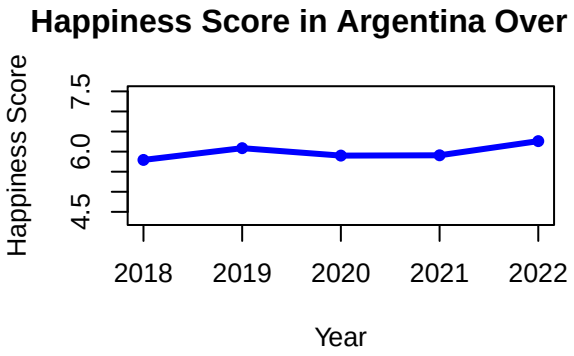
Variables: - year - country - positive: number of positive songs - neutral: number of neutral songs - negative: number of negative songs - developed_country: indicator, true if the country is considered as developed country

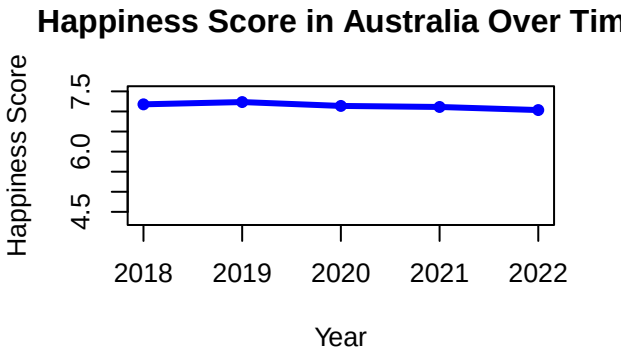
```
summary(df)
```

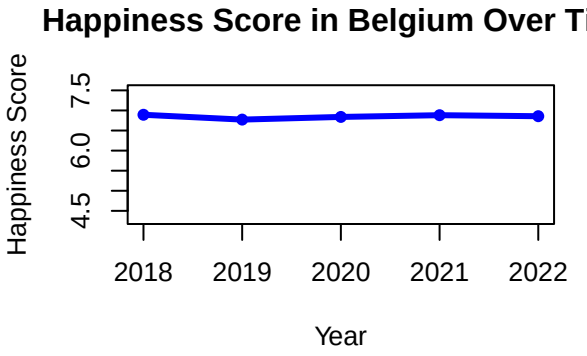
```
##       year        country             positive        neutral
##  Min.   :2018   Length:90          Min.   :17.00   Min.   : 0.00
##  1st Qu.:2019   Class :character   1st Qu.:35.25   1st Qu.: 5.00
##  Median :2020   Mode  :character   Median :41.50   Median : 8.00
##  Mean   :2020                      Mean   :41.67   Mean   : 9.90
##  3rd Qu.:2021                      3rd Qu.:47.00   3rd Qu.:12.75
##  Max.   :2022                      Max.   :78.00   Max.   :41.00
##     negative       happiness     developed_country
##  Min.   :27.00   Min.   :4.367   Mode :logical
##  1st Qu.:52.00   1st Qu.:5.889   FALSE:45
##  Median :56.00   Median :6.116   TRUE :45
##  Mean   :58.43   Mean   :6.241
##  3rd Qu.:66.25   3rd Qu.:6.882
##  Max.   :89.00   Max.   :7.444
```

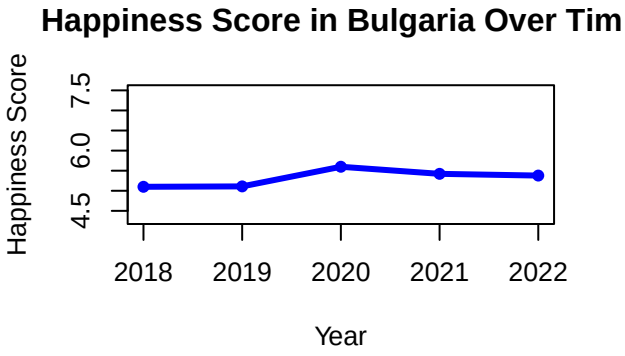
```
par(mfrow = c(2, 2))
plot_happiness <- function(data, country_name) {
  country_data <- data[data$country == country_name, ]
  # output as individual plots
  #png(paste("plots/Happiness Score in", country_name, "Over Time.png"), width = 600, height = 400)
  plot(
    country_data$year,
    country_data$happiness,
    type = "l", # Line plot
    main = paste("Happiness Score in", country_name, "Over Time"),
    xlab = "Year",
    ylab = "Happiness Score",
    col = "blue",
    lwd = 3,
    ylim = c(4.3, 7.5)
  )

  points(country_data$year, country_data$happiness, col = "blue", pch = 16)
```

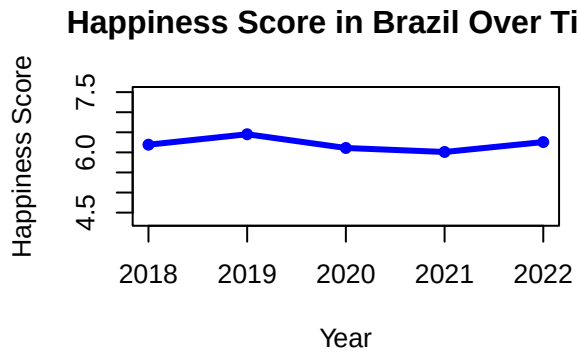
```
  #dev.off()
}

for (c in unique(df$country)){
  plot_happiness(df, c)
}
```

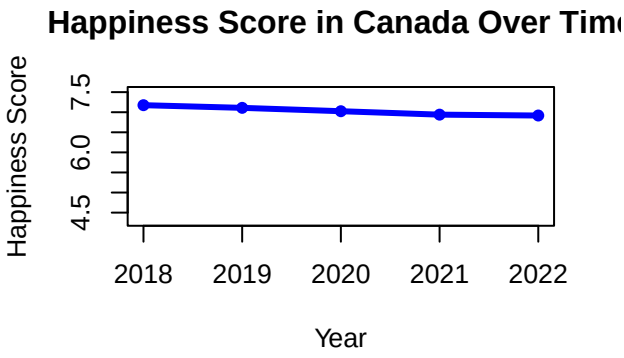


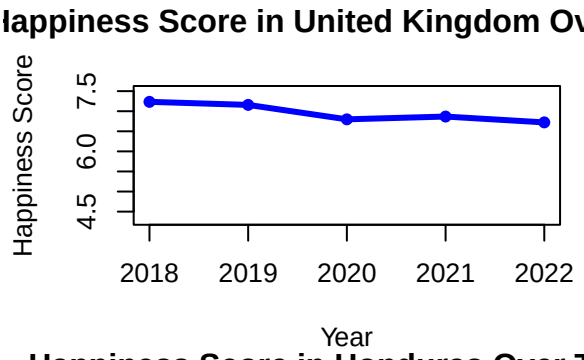


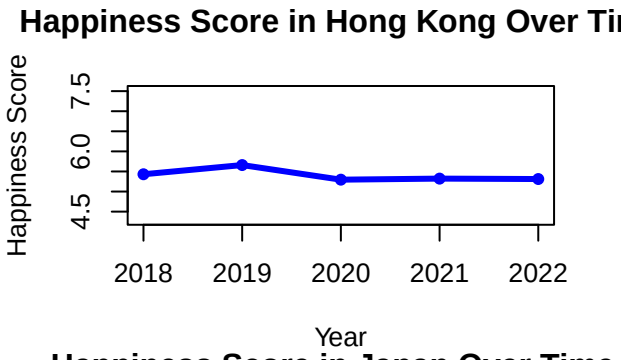


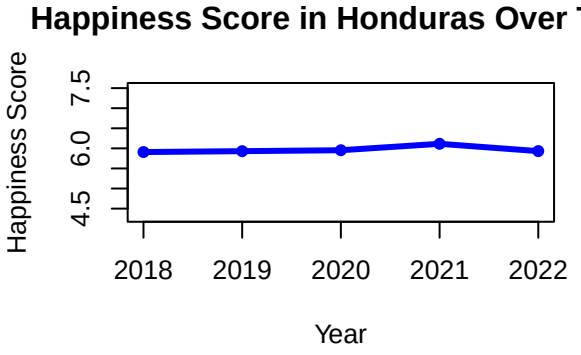


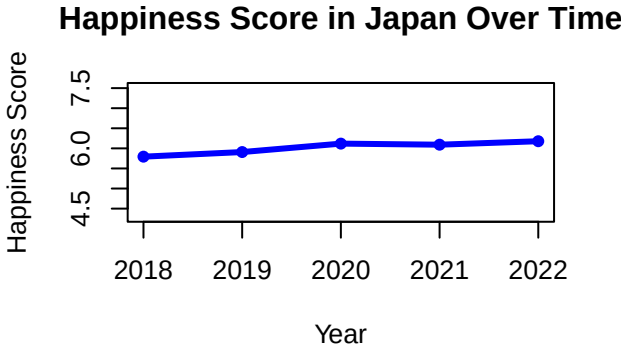

Happiness Score in Brazil Over Time
Happiness Score
4.5
6.0
7.5
2018
2019
2020
2021
2022
Year

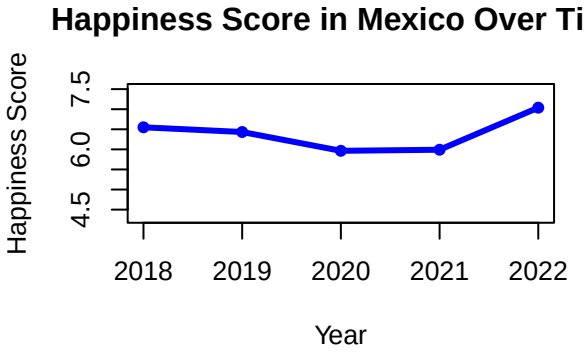

Happiness Score in Canada Over Tim
Happiness Score
4.5
6.0
7.5
2018
2019
2020
2021
2022
Year

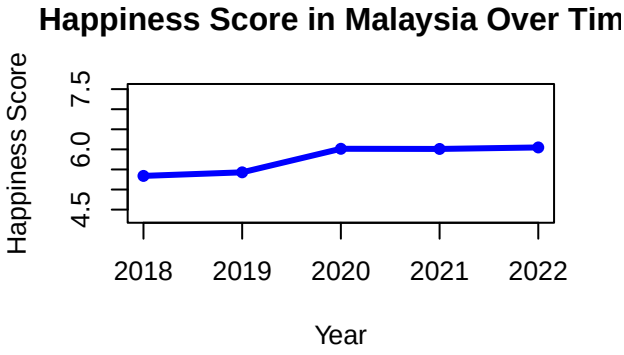

lappiness Score in United Kingdom Over
Happiness Score
4.5
6.0
7.5
2018
2019
2020
2021
2022
Year

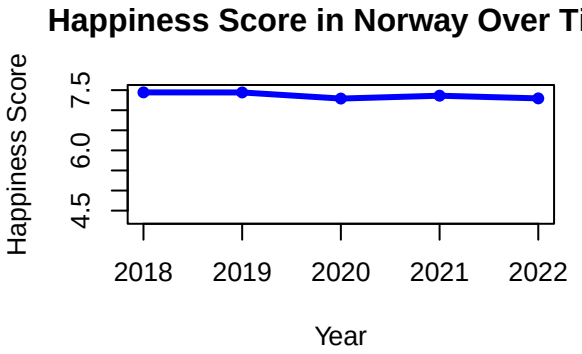

Happiness Score in Hong Kong Over Ti
Happiness Score
4.5
6.0
7.5
2018
2019
2020
2021
2022
Year

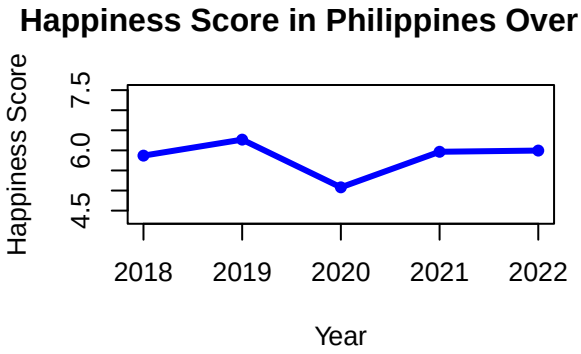

Happiness Score in Honduras Over Tin
Happiness Score
4.5
6.0
7.5
2018
2019
2020
2021
2022
Year

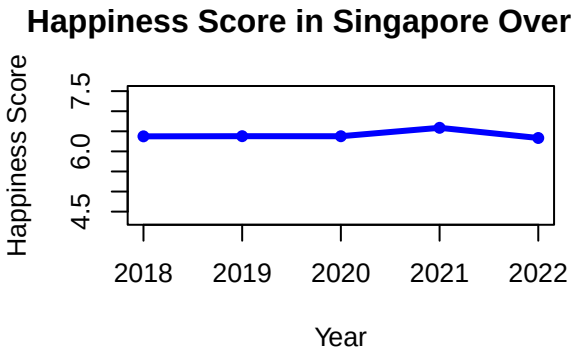

Happiness Score in Japan Over Time
Happiness Score
4.5
6.0
7.5
2018
2019
2020
2021
2022
Year

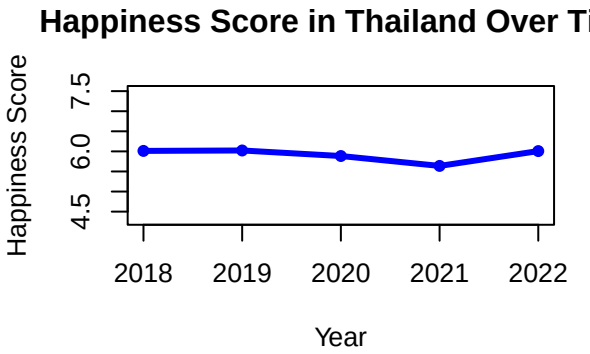

Happiness Score in Mexico Over Time
Happiness Score
4.5
6.0
7.5
2018
2019
2020
2021
2022
Year


Happiness Score in Malaysia Over Tim
Happiness Score
4.5
6.0
7.5
2018
2019
2020
2021
2022
Year

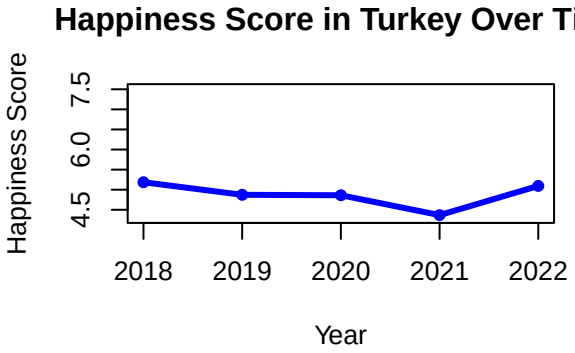


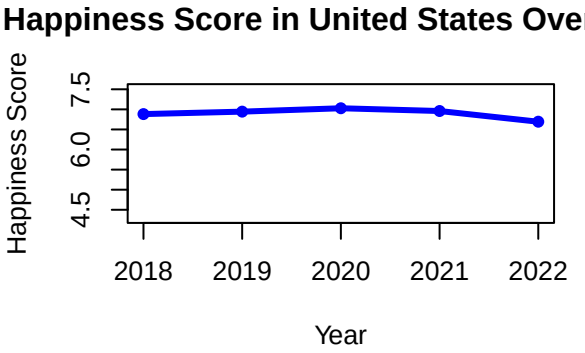







```r
# Rescale
# multiply by 10 to match the scale of the happiness score
df$positive <- df$positive / 110 * 10
df$neutral <- df$neutral / 110 * 10
df$negative <- df$negative / 110 * 10
```






```r
# Correlation matrix
cor_data <- df[, c("happiness", "positive", "neutral", "negative")]
cor_matrix <- cor(cor_data)
print(cor_matrix)
```

```
##           happiness   positive    neutral   negative
## happiness  1.0000000 -0.1867838 -0.1486041  0.2511243
## positive  -0.1867838  1.0000000 -0.1087886 -0.8062066
## neutral   -0.1486041 -0.1087886  1.0000000 -0.5004165
## negative   0.2511243 -0.8062066 -0.5004165  1.0000000
```

- We see that `positive` and `negative` are strongly negatively correlated, which can cause multicollinearity issues in the regression models. Hence, we combine `postive` and `negative` into a new feature to

avoid such issues.

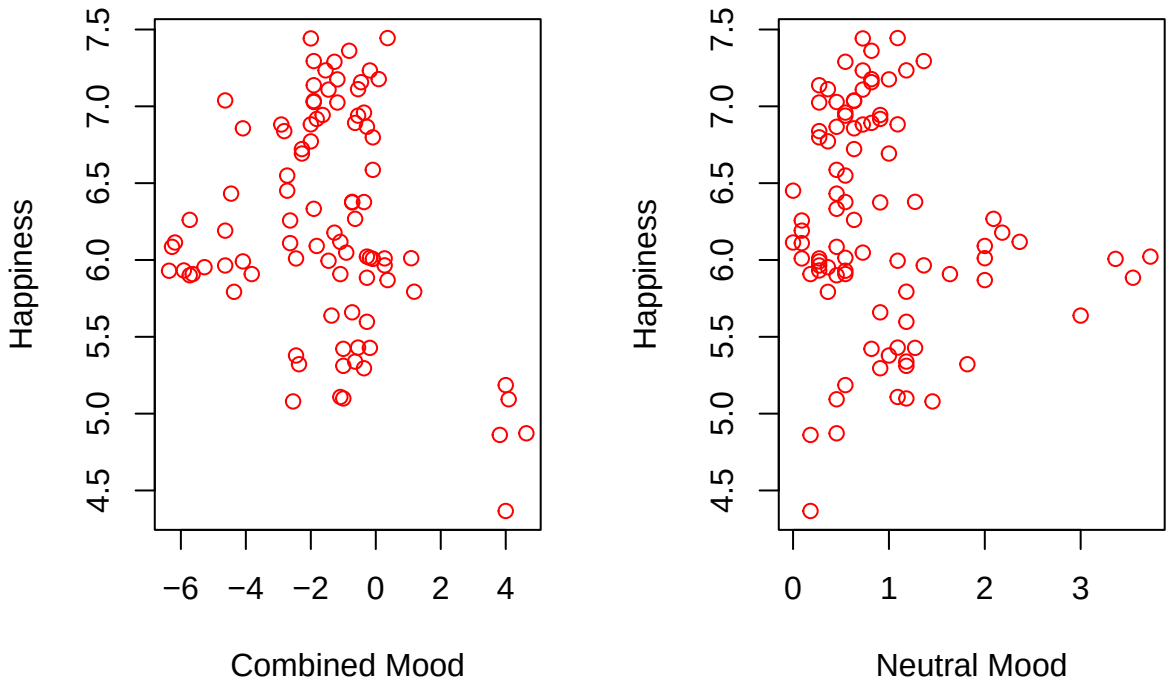
- Also, notice that happiness **is not correlated** with the count of positive/neutral/negative song in general.

```
# higher value of comb means more positive songs were listened
df$comb <- df$positive - df$negative

# re-examine correlation
cor_data <- df[, c("happiness", "neutral", "comb")]
cor_matrix <- cor(cor_data)
print(cor_matrix)
```

```
##             happiness    neutral       comb
## happiness  1.0000000 -0.1486041 -0.2326765
## neutral   -0.1486041  1.0000000  0.2281007
## comb      -0.2326765  0.2281007  1.0000000
```

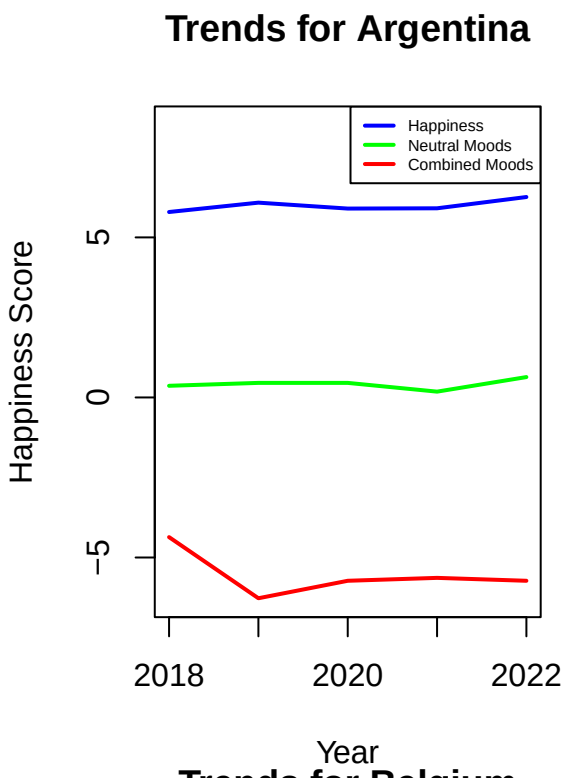
No highly correlated variables!

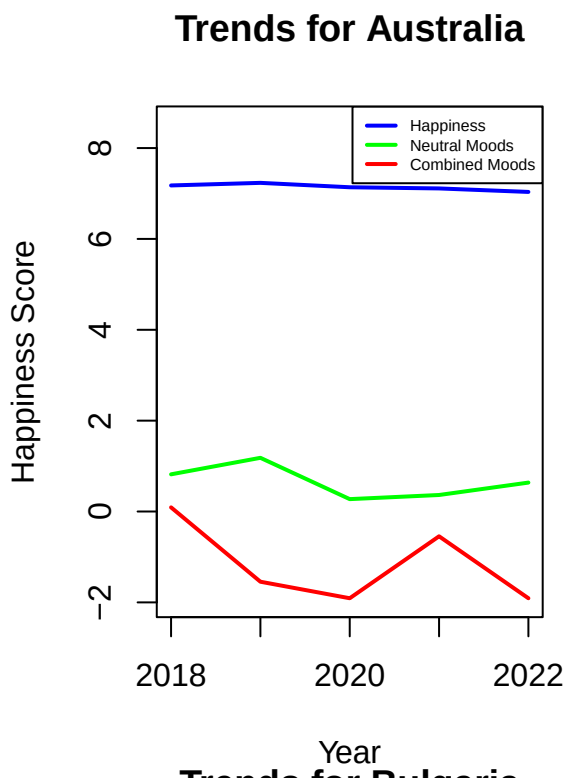
```
par(mfrow = c(1, 2))

plot(df$comb, df$happiness,
     xlab = "Combined Mood",
     ylab = "Happiness",
     col = "red")

plot(df$neutral, df$happiness,
     xlab = "Neutral Mood",
     ylab = "Happiness",
     col = "red")
```

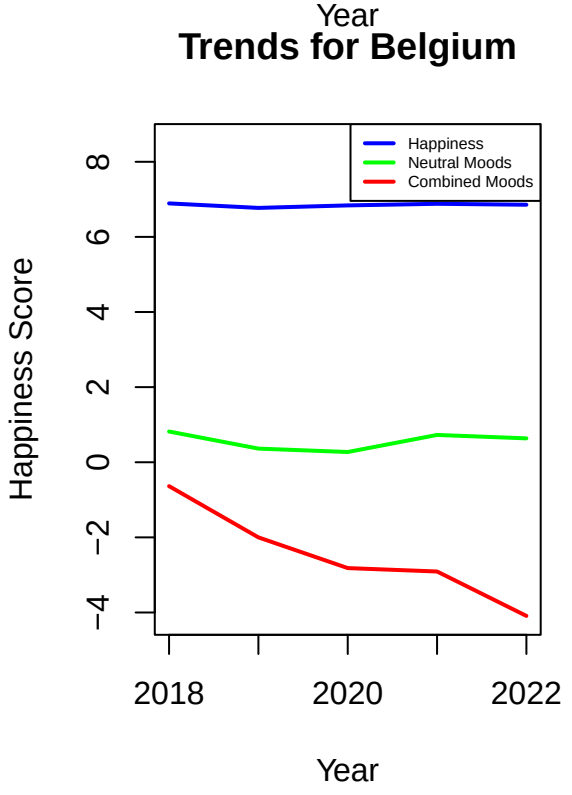


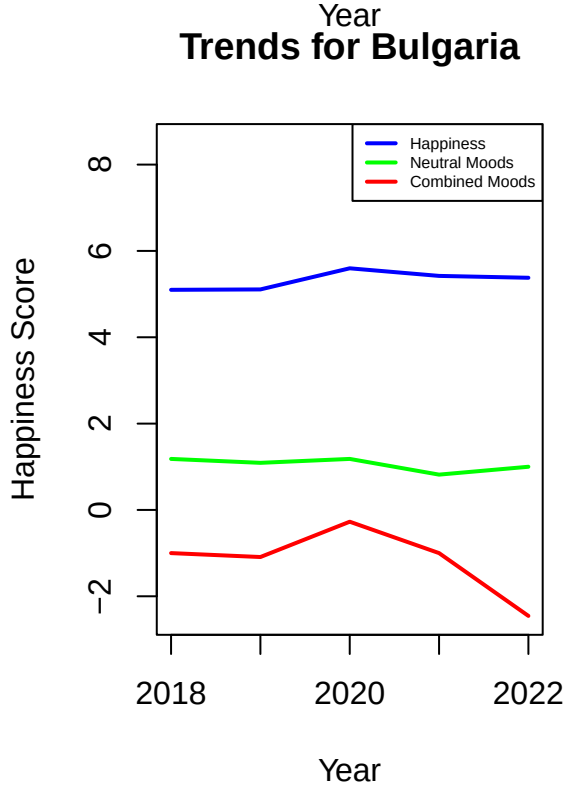
Trend over Time by Country

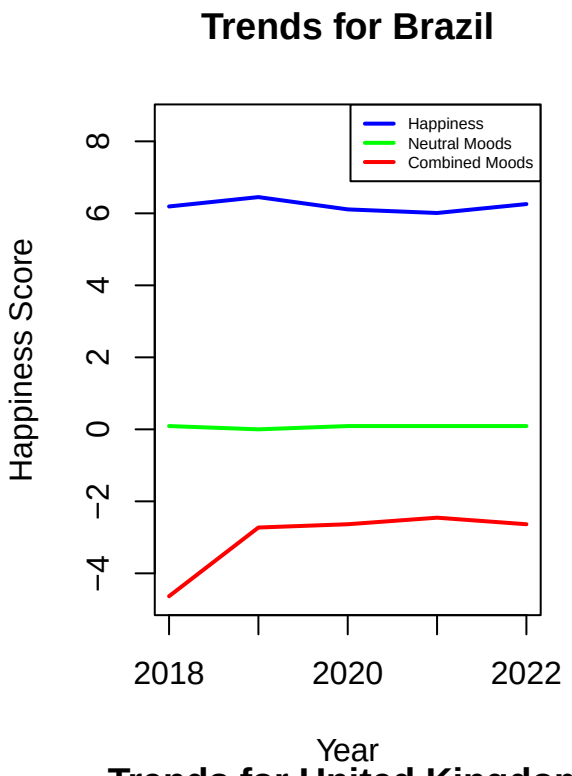
```r
par(mfrow = c(1,2))

plot_country_trends <- function(data) {
  countries <- unique(data$country)

  for (country in countries) {
    country_data <- data[data$country == country, ]

    plot(
      country_data$year,
      country_data$happiness,
      type = "l",
      main = paste("Trends for", country),
      xlab = "Year",
      ylab = "Happiness Score",
      col = "blue",
      ylim = c(min(c(country_data$comb, country_data$neutral))
               , 8.5),
      lwd = 2
    )

    lines(country_data$year, country_data$neutral, col = "green", lwd = 2)
    lines(country_data$year, country_data$comb, col = "red", lwd = 2)

    legend(
      "topright",
      legend = c("Happiness", "Neutral Moods", "Combined Moods"),
      col = c("blue", "green", "red"),
      lty = 1,
      lwd = 2,
      cex = 0.5
    )
  }
}

plot_country_trends(df)
```

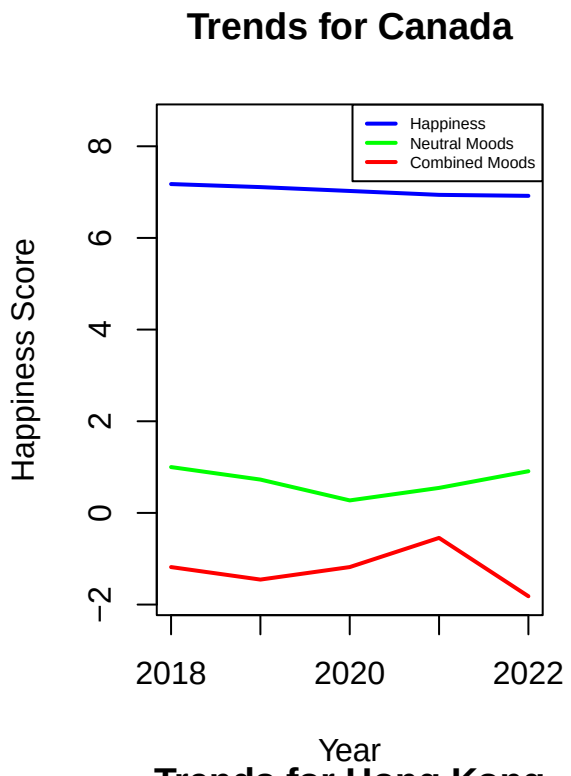

Trends for Argentina
Happiness
Neutral Moods
Combined Moods
Happiness Score
5
0
−5
2018
2020
2022
Year

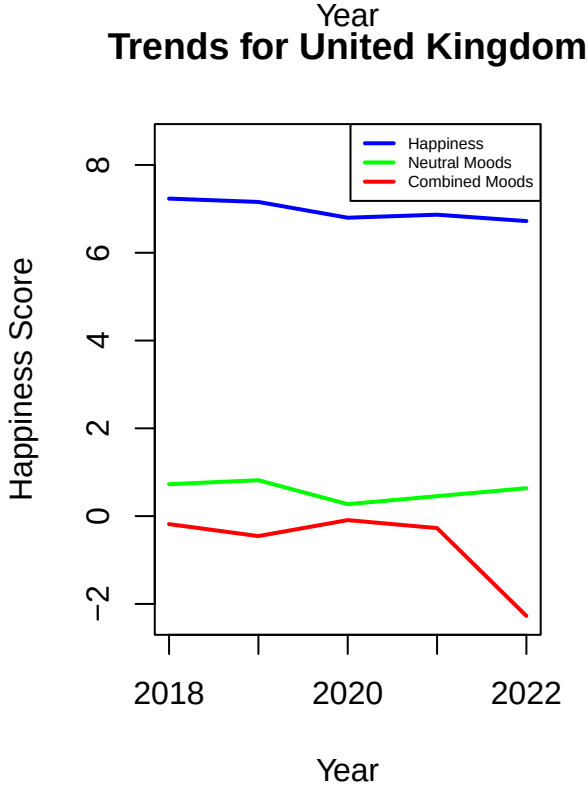

Trends for Australia
Happiness
Neutral Moods
Combined Moods
Happiness Score
8
6
4
2
0
−2
2018
2020
2022
Year

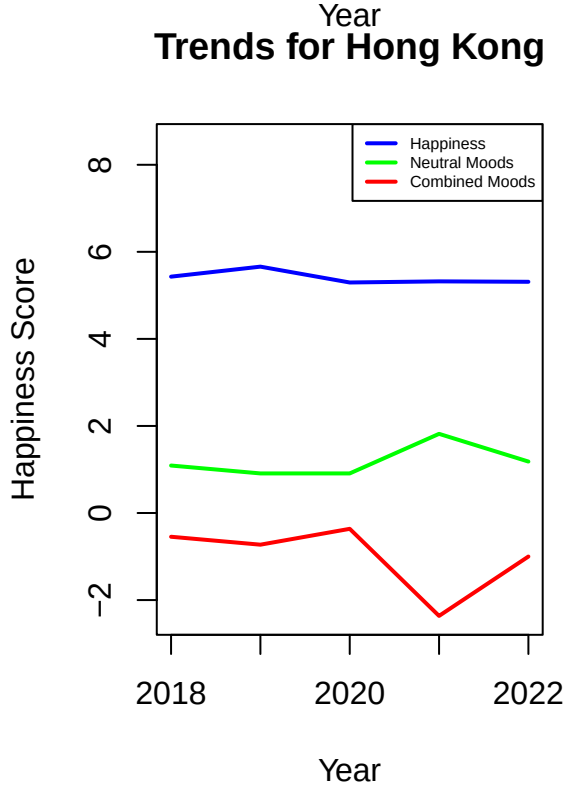

Trends for Belgium
Happiness
Neutral Moods
Combined Moods
Happiness Score
8
6
4
2
0
−2
−4
2018
2020
2022
Year

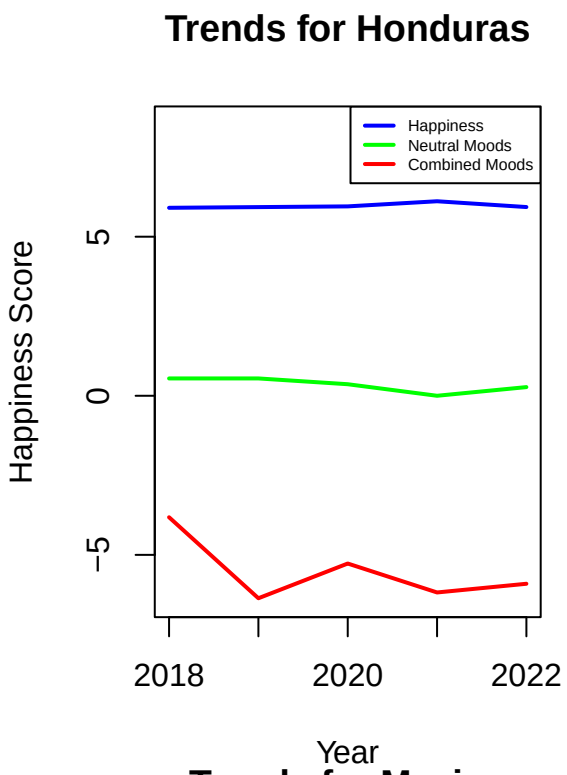

Trends for Bulgaria
Happiness
Neutral Moods
Combined Moods
Happiness Score
8
6
4
2
0
−2
2018
2020
2022
Year


Trends for Brazil
Happiness
Neutral Moods
Combined Moods
Happiness Score
8
6
4
2
0
−2
−4
2018
2020
2022
Year

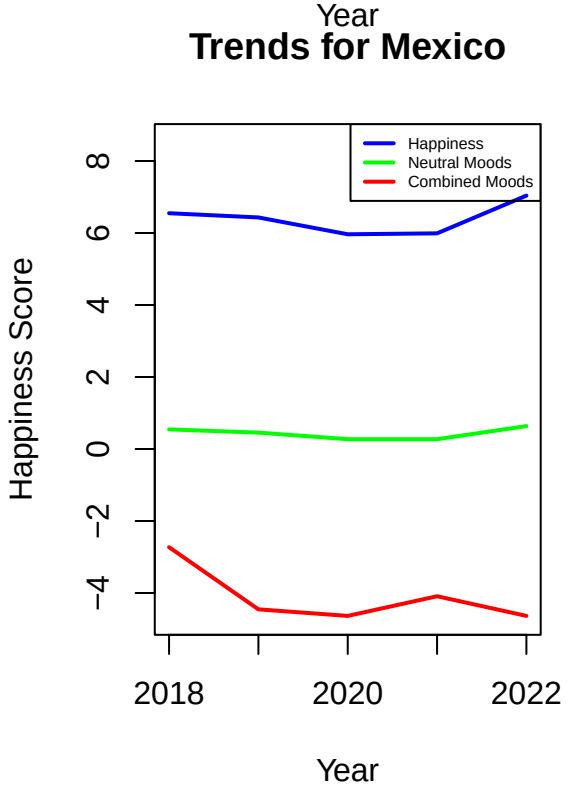

Trends for Canada
Happiness
Neutral Moods
Combined Moods
Happiness Score
8
6
4
2
0
−2
2018
2020
2022
Year

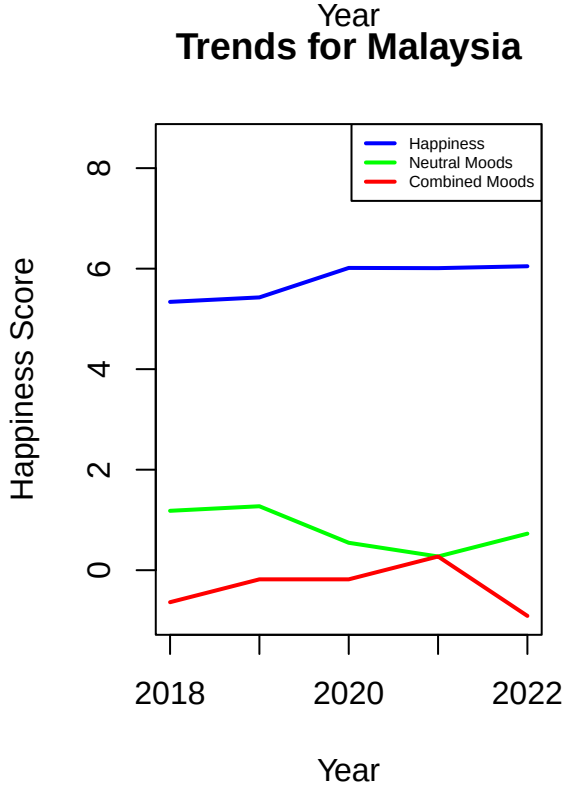

Trends for United Kingdom
Happiness
Neutral Moods
Combined Moods
Happiness Score
8
6
4
2
0
−2
2018
2020
2022
Year

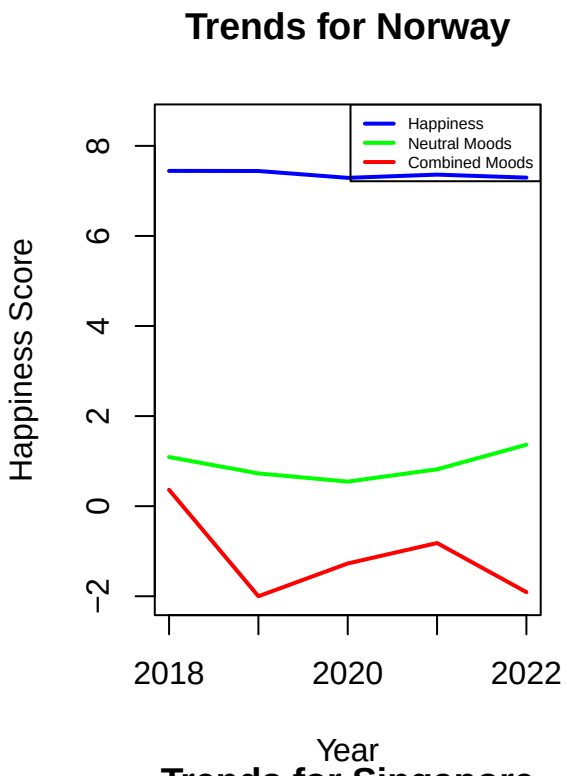

Trends for Hong Kong
Happiness
Neutral Moods
Combined Moods
Happiness Score
8
6
4
2
0
−2
2018
2020
2022
Year

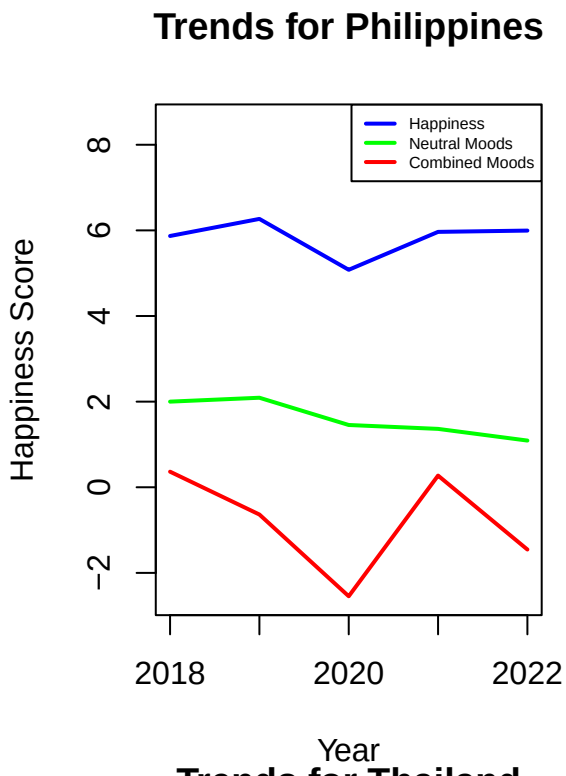

Trends for Honduras
Happiness
Neutral Moods
Combined Moods
Happiness Score
5
0
−5
2018
2020
2022
Year

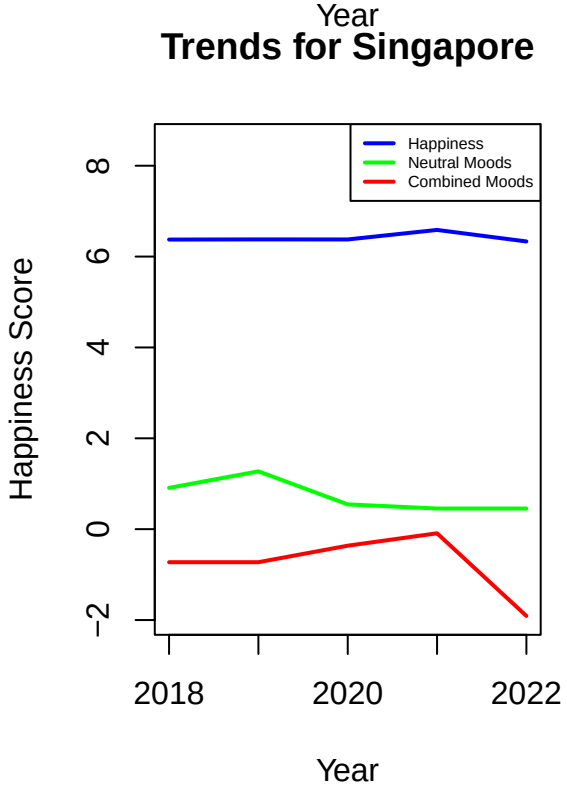

Trends for Japan
Happiness
Neutral Moods
Combined Moods
Happiness Score
8
6
4
2
0
−2
2018
2020
2022
Year

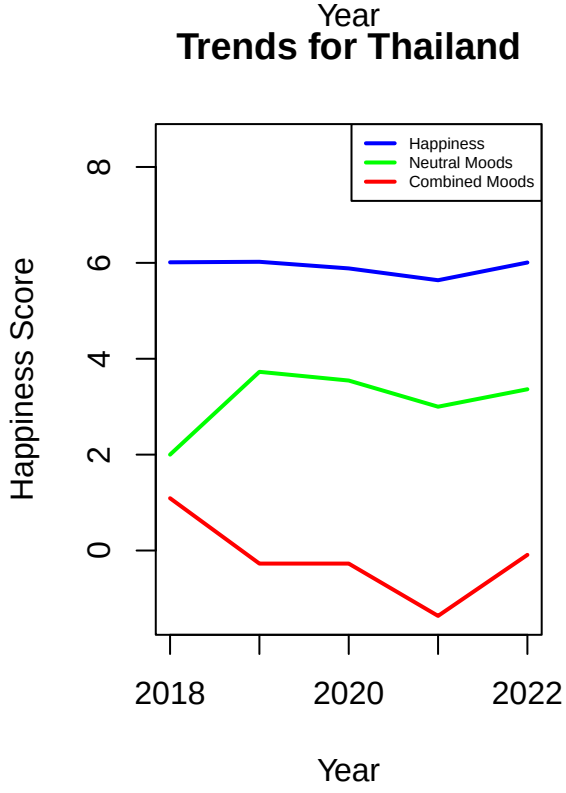

Trends for Mexico
Happiness
Neutral Moods
Combined Moods
Happiness Score
8
6
4
2
0
−2
−4
2018
2020
2022
Year

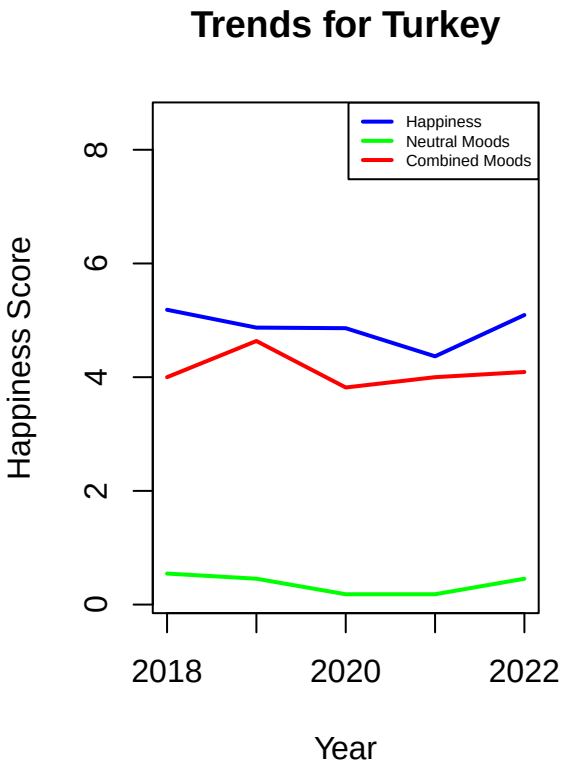

Trends for Malaysia
Happiness
Neutral Moods
Combined Moods
Happiness Score
8
6
4
2
0
2018
2020
2022
Year

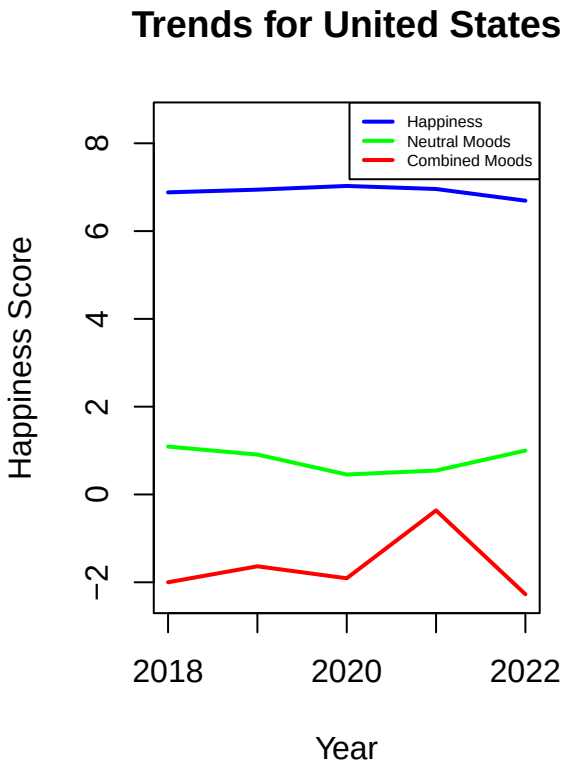

Trends for Norway
Happiness
Neutral Moods
Combined Moods
Happiness Score
−2
0
2
4
6
8
2018
2020
2022
Year


Trends for Philippines
Happiness
Neutral Moods
Combined Moods
Happiness Score
−2
0
2
4
6
8
2018
2020
2022
Year


Trends for Singapore
Happiness
Neutral Moods
Combined Moods
Happiness Score
−2
0
2
4
6
8
2018
2020
2022
Year


Trends for Thailand
Happiness
Neutral Moods
Combined Moods
Happiness Score
0
2
4
6
8
2018
2020
2022
Year





Modelling

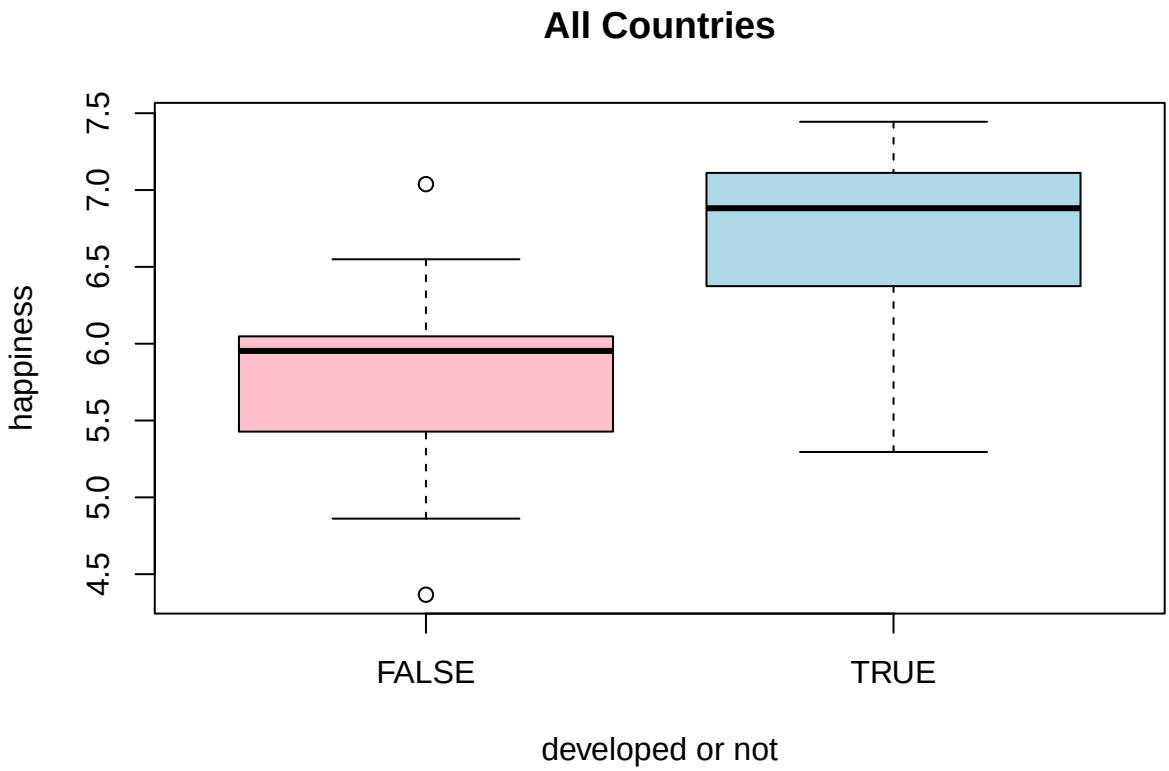
Baseline Model

```
m0 <- lm(happiness ~  comb + neutral, data = df)
summary(m0)
```

```
##
## Call:
## lm(formula = happiness ~ comb + neutral, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57999 -0.53897 -0.03452  0.61905  1.34475
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.22612    0.13420  46.393   <2e-16 ***
## comb        -0.06559    0.03333  -1.968   0.0522 .
## neutral     -0.09419    0.09959  -0.946   0.3469
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.6919 on 87 degrees of freedom
## Multiple R-squared:  0.06377,    Adjusted R-squared:  0.04224
## F-statistic: 2.963 on 2 and 87 DF,  p-value: 0.05692
```

Examine the effect of the development of a country

```
boxplot(happiness~developed_country, main = "All Countries", col = c("pink", "lightblue"),
        data = df, xlab = "developed or not")
```




```
fit0 <- lm(happiness ~ developed_country, data = df)
summary(fit0)
```

```
##
## Call:
## lm(formula = happiness ~ developed_country, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4410 -0.3316  0.1728  0.3523  1.2308
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.80762    0.08343  69.611  < 2e-16 ***
## developed_countryTRUE  0.86740    0.11799   7.352 9.57e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5597 on 88 degrees of freedom
## Multiple R-squared:  0.3805, Adjusted R-squared:  0.3734
## F-statistic: 54.05 on 1 and 88 DF,  p-value: 9.57e-11
```

- `developed_countryTRUE`: the difference in happiness between developed and developing countries on average is 0.867
- we see from the boxplot that the IQR and mean of happiness score for developed countries are much

higher than those of developing countries, which suggests the happiness score of developed countries are higher in general. Also, the `developed_countryTRUE` coefficient has a positive value of 0.867, which implies developed countries has higher happiness score. The null hypothesis $H_0 : \beta_1 = 0$ has a p-value of 9.57e-11 « 0.001, this means there is strong evidence against the null hypothesis, which aligns with our previous conclusion that developed countries has higher happiness score on average.

Model with New Variable

```
m1 <- lm(happiness ~ comb*developed_country +
           neutral*developed_country,
         data = df)
summary(m1)
```

```
##
## Call:
## lm(formula = happiness ~ comb * developed_country + neutral *
##     developed_country, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30516 -0.25677  0.00975  0.29010  1.00032
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     5.51949    0.11844  46.602  < 2e-16 ***
## comb                           -0.11237    0.02469  -4.551 1.79e-05 ***
## developed_countryTRUE           1.57145    0.20840   7.540 5.00e-11 ***
## neutral                         0.08931    0.07840   1.139 0.257889
## comb:developed_countryTRUE      0.03927    0.07726   0.508 0.612622
## developed_countryTRUE:neutral -0.65802    0.16435  -4.004 0.000134 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4768 on 84 degrees of freedom
## Multiple R-squared:  0.5708, Adjusted R-squared:  0.5453
## F-statistic: 22.35 on 5 and 84 DF,  p-value: 3.524e-14
```

```
anova(m0, m1)
```

```
## Analysis of Variance Table
##
## Model 1: happiness ~ comb + neutral
## Model 2: happiness ~ comb * developed_country + neutral * developed_country
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     87 41.655
## 2     84 19.095  3     22.56 33.081 3.246e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

m1 is a better model, adding `developed_country` is reasonble.

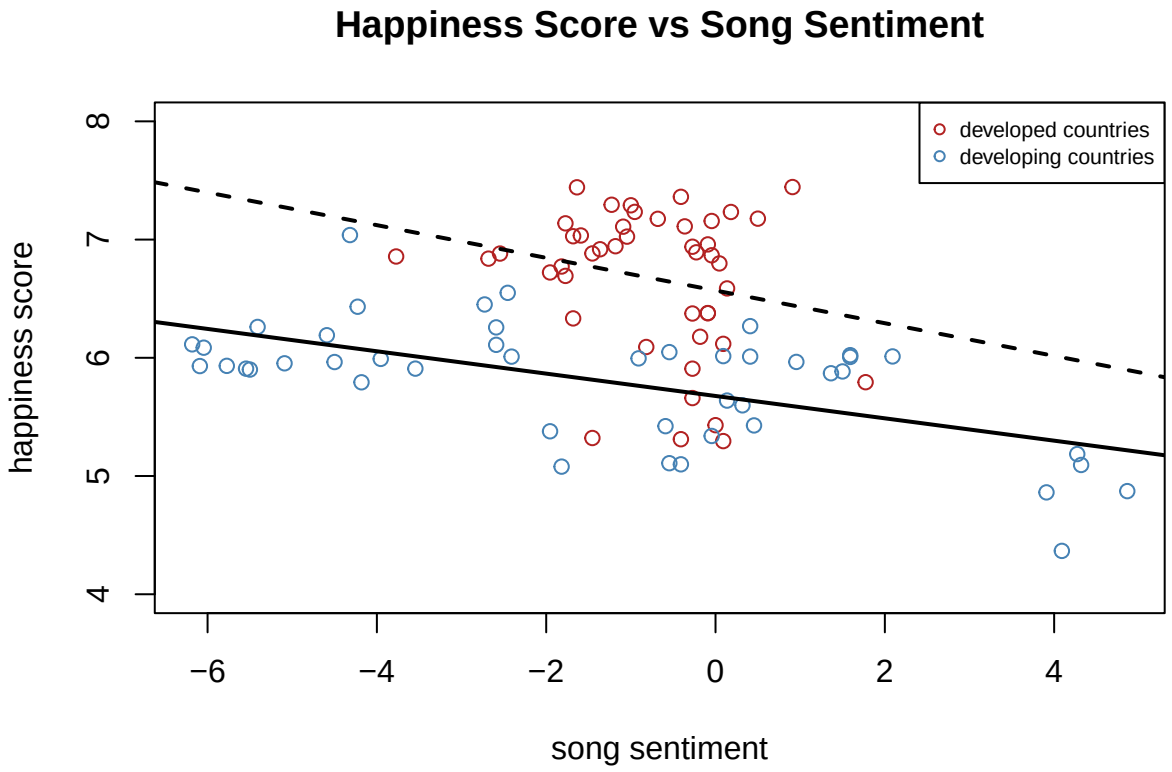
More Analysis on `devloped_country`

For better interpretability, we define a new variable `in_one` to combine all 3 moods

```r
# higher value of in_one means people tend to listen to more positive songs
# lower value of in_one means people tend to listen to more negative songs
# 0.5 is a random weight
df$in_one <- df$positive - df$negative + 0.5 * df$neutral

# the final model
m3 <- lm(happiness ~ in_one*developed_country, data = df)
summary(m3)
```

```
##
## Call:
## lm(formula = happiness ~ in_one * developed_country, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.44856 -0.29238  0.00525  0.35386  1.00089
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    5.67701    0.08359  67.915  < 2e-16 ***
## in_one                        -0.09452    0.02426  -3.897 0.000193 ***
## developed_countryTRUE          0.89203    0.12760   6.991 5.52e-10 ***
## in_one:developed_countryTRUE -0.04371    0.08015  -0.545 0.586881
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5137 on 86 degrees of freedom
## Multiple R-squared:   0.49,  Adjusted R-squared:  0.4722
## F-statistic: 27.54 on 3 and 86 DF,  p-value: 1.413e-12
```

```r
plot(happiness ~ in_one, main = "Happiness Score vs Song Sentiment",
     xlab = "song sentiment", ylab = "happiness score",
     data = df[df$developed_country == TRUE,], col = "firebrick",
     ylim = c(4, 8), xlim = c(min(df$in_one), max(df$in_one)))
points(happiness ~ in_one, data = df[df$developed_country == FALSE,],
       col = "steelblue")
legend("topright", c("developed countries", "developing countries"), cex = 0.7,
       col = c("firebrick", "steelblue"), pch = 1)
coefs_hat <- coefficients(m3)
abline(coefs_hat[1], coefs_hat[2], lty = 1, lwd = 2)
abline(coefs_hat[1]+coefs_hat[3], coefs_hat[2]+coefs_hat[4], lty = 2, lwd = 2)
```



Interpretation:

- We plot the linear fitted lines for developed and developing countries separately. Notice that there is a more obvious trend for developing countries while there is no obvious relationship between the sentiments of songs and the happiness score of one country for developed countries.
- Notice that the R-squared value for the model is low, which suggests happiness score is not well-explained by the variables we used.

Additionally, we examine the correlation between happiness and sentiment of songs separately for developed and developing countries.

```
# General correlation
cor_data <- df[, c("happiness", "in_one")]
cor_matrix <- cor(cor_data)
print(cor_matrix)
```

```
##             happiness     in_one
## happiness  1.0000000 -0.2450744
## in_one    -0.2450744  1.0000000
```

```
# for developed countries
developed_df <- df[df$developed_country == TRUE, ]
cor_data <- developed_df[, c("happiness", "in_one")]
cor_matrix <- cor(cor_data)
print(cor_matrix)
```

```
##             happiness     in_one
## happiness  1.0000000 -0.2315824
## in_one    -0.2315824  1.0000000
```

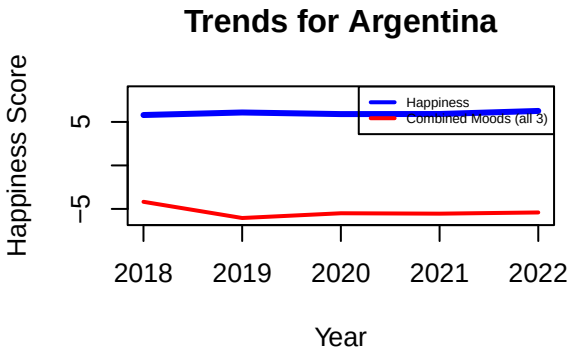
```
# for developing countries
developing_df <- df[df$developed_country == FALSE, ]
cor_data <- developing_df[, c("happiness", "in_one")]
```

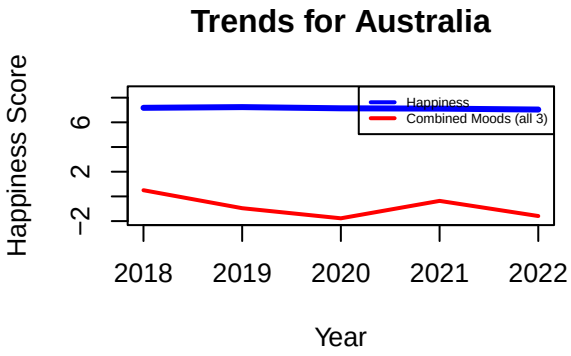
```r
cor_matrix <- cor(cor_data)
print(cor_matrix)
```

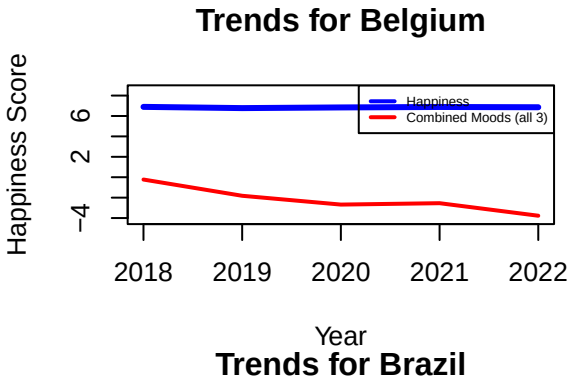
```
##            happiness     in_one
## happiness  1.0000000 -0.5915152
## in_one    -0.5915152  1.0000000
```

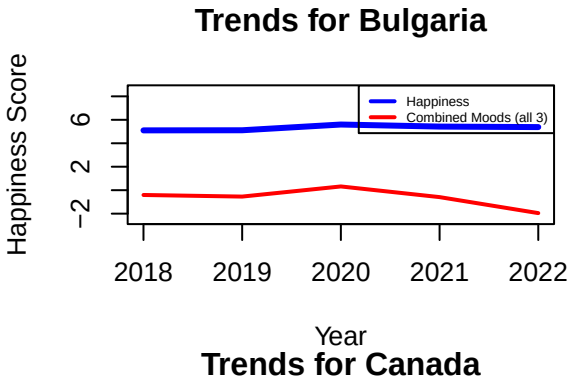
Based on the three correlation matrices, we can conclude that, overall, there is no significant relationship between a country's happiness score and the moods of songs its people prefer. A similar observation holds true for developed countries. However, in developing countries, a moderate negative correlation exists, indicating that individuals in happier countries tend to listen to songs with more negative moods.

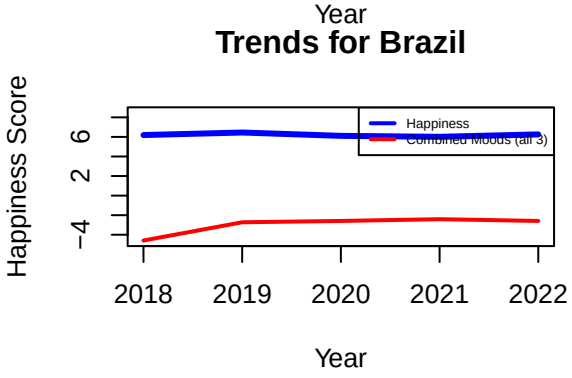
```r
par(mfrow = c(2,2))
plot_country_trends <- function(data) {
  countries <- unique(data$country)

  for (country in countries) {
    country_data <- data[data$country == country, ]
    # output the happiness and mood over time plots as individual files
    #png(paste("plots/Trends for", country, ".png"), width = 800, height = 600)
    plot(
      country_data$year,
      country_data$happiness,
      type = "l",
      main = paste("Trends for", country),
      xlab = "Year",
      ylab = "Happiness Score",
      col = "blue",
      ylim = c(min(c(country_data$comb, country_data$neutral))
               , 8.5),
      lwd = 3
    )
    lines(country_data$year, country_data$in_one, col = "red", lwd = 2)
    legend(
      "topright",
      legend = c("Happiness", "Combined Moods (all 3)"),
      col = c("blue", "red"),
      lty = 1,
      lwd = 2,
      cex = 0.5
    )
    #dev.off()
  }
}

plot_country_trends(df)
```

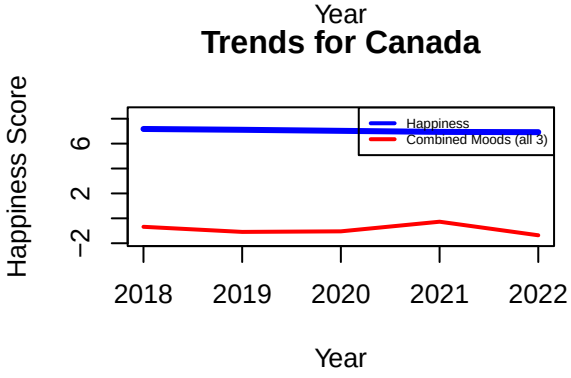

Trends for Argentina
Happiness Score
−5
5
2018
2019
2020
2021
2022
Year
Happiness
Combined Moods (all 3)

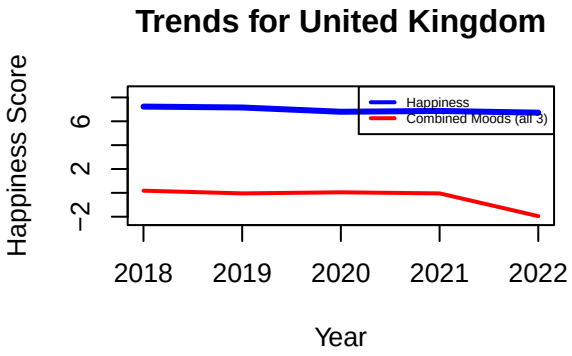

Trends for Australia
Happiness Score
−2
2
6
2018
2019
2020
2021
2022
Year
Happiness
Combined Moods (all 3)

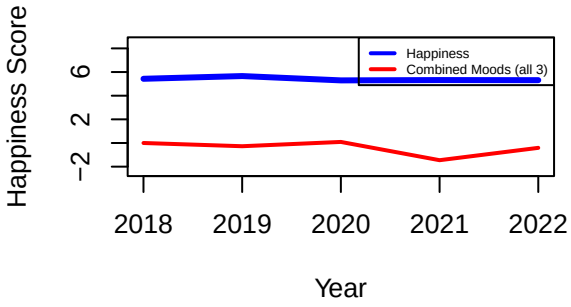

Trends for Belgium
Happiness Score
−4
2
6
2018
2019
2020
2021
2022
Year
Happiness
Combined Moods (all 3)

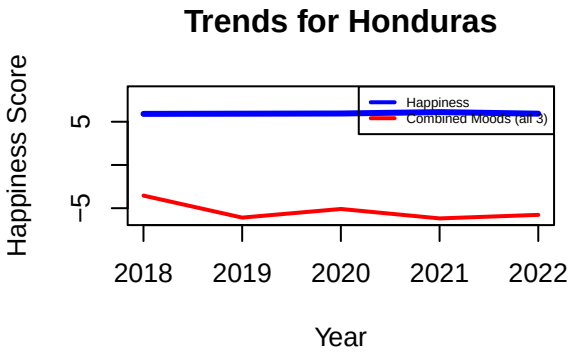

Trends for Bulgaria
Happiness Score
−2
2
6
2018
2019
2020
2021
2022
Year
Happiness
Combined Moods (all 3)

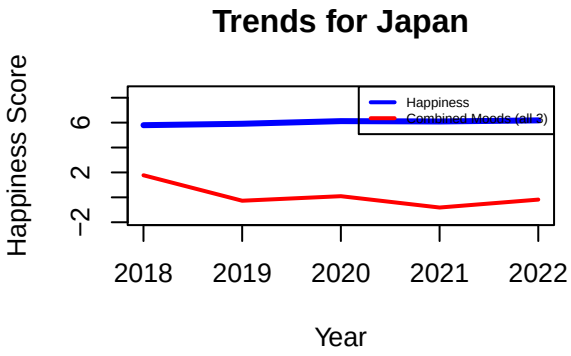

Trends for Brazil
Happiness Score
−4
2
6
2018
2019
2020
2021
2022
Year
Happiness
Combined Moods (all 3)

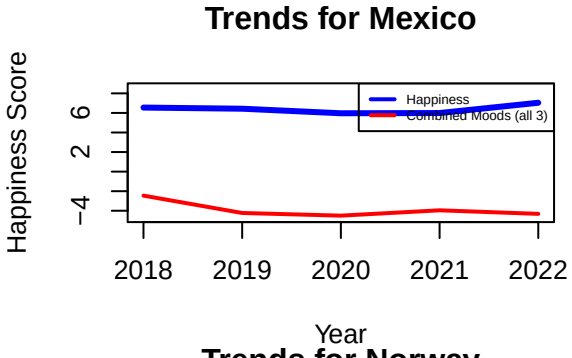

Trends for Canada
Happiness Score
−2
2
6
2018
2019
2020
2021
2022
Year
Happiness
Combined Moods (all 3)

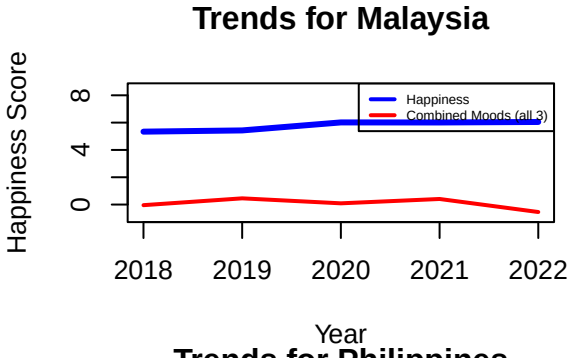

Trends for United Kingdom
Happiness Score
−2
2
6
2018
2019
2020
2021
2022
Year
Happiness
Combined Moods (all 3)

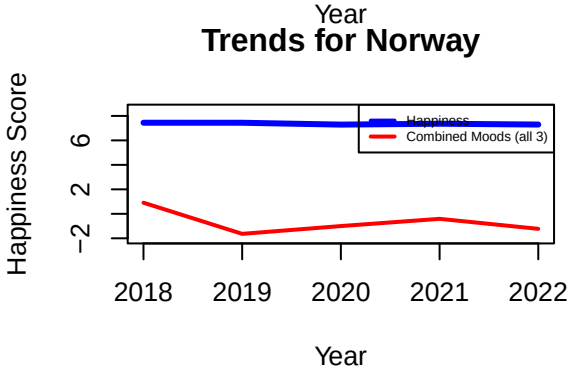

Trends for Hong Kong
Happiness Score
−2
2
6
2018
2019
2020
2021
2022
Year
Happiness
Combined Moods (all 3)

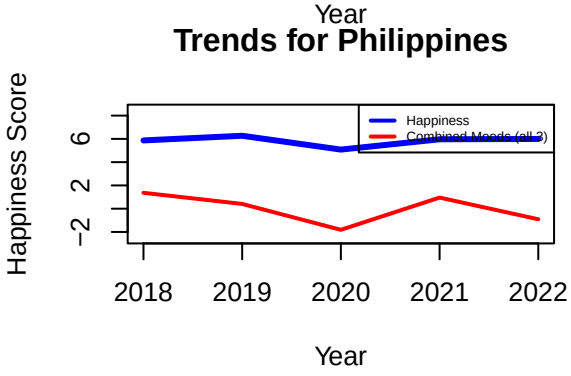

Trends for Honduras
Happiness Score
−5
5
2018
2019
2020
2021
2022
Year
Happiness
Combined Moods (all 3)

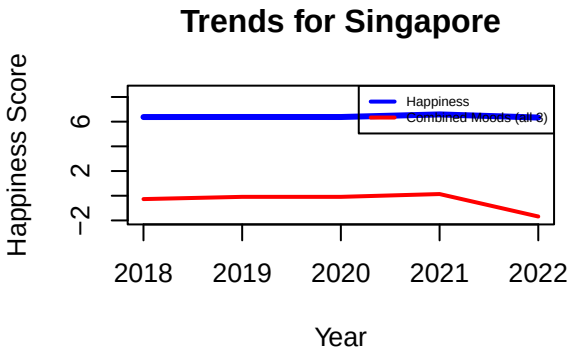

Trends for Japan
Happiness Score
−2
2
6
2018
2019
2020
2021
2022
Year
Happiness
Combined Moods (all 3)

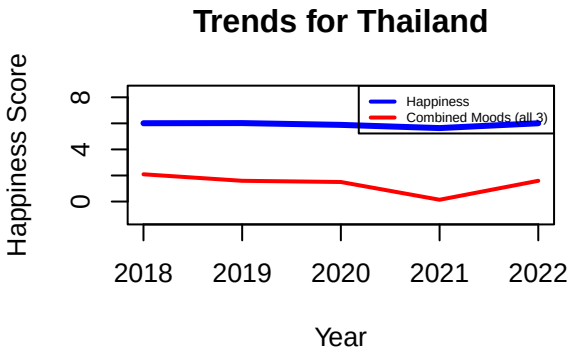

Trends for Mexico
Happiness Score
−4
2
6
2018
2019
2020
2021
2022
Year
Happiness
Combined Moods (all 3)

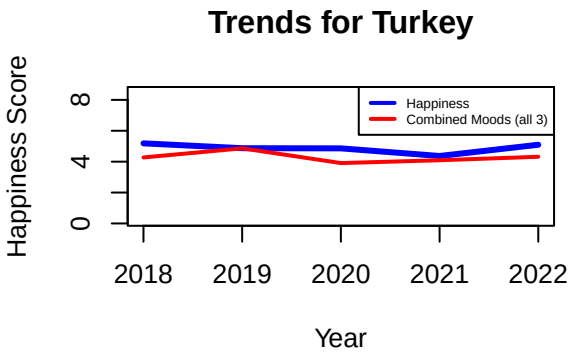

Trends for Malaysia
Happiness Score
0
4
8
2018
2019
2020
2021
2022
Year
Happiness
Combined Moods (all 3)

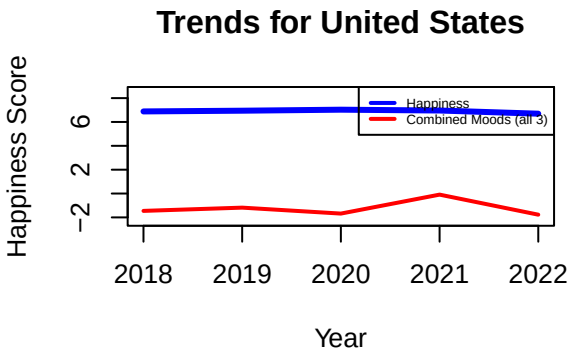

Trends for Norway
Happiness Score
−2
2
6
2018
2019
2020
2021
2022
Year
Happiness
Combined Moods (all 3)


Trends for Philippines
Happiness Score
−2
2
6
2018
2019
2020
2021
2022
Year
Happiness
Combined Moods (all 3)


Trends for Singapore
Happiness Score
−2
2
6
2018
2019
2020
2021
2022
Year
Happiness
Combined Moods (all 3)


Trends for Thailand
Happiness Score
0
4
8
2018
2019
2020
2021
2022
Year
Happiness
Combined Moods (all 3)


Trends for Turkey
Happiness Score
0
4
8
Happiness
Combined Moods (all 3)
2018
2019
2020
2021
2022
Year


Trends for United States
Happiness Score
−2
2
6
Happiness
Combined Moods (all 3)
2018
2019
2020
2021
2022
Year