

Mental Health data Analysis

Hemanth

2025-04-25

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

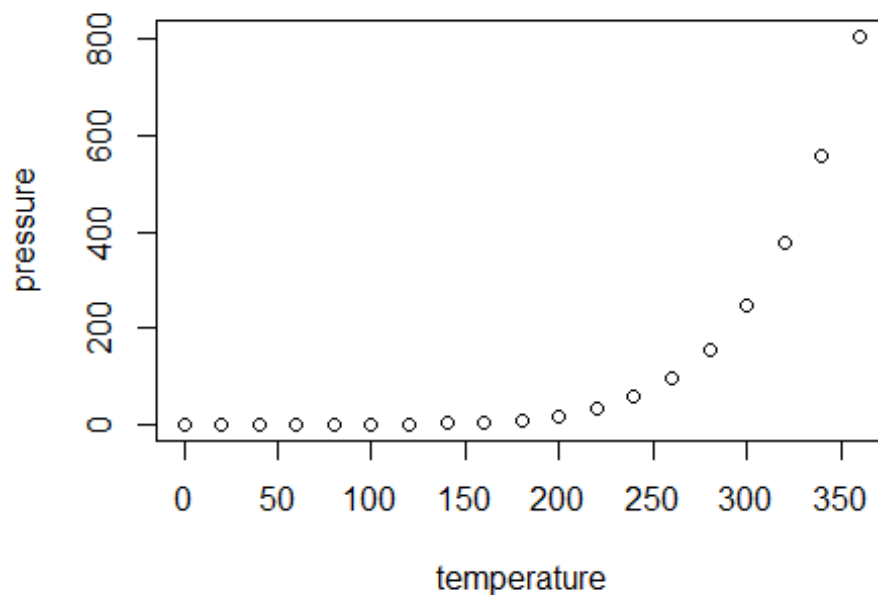
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)

##      speed          dist
##  Min.   : 4.0    Min.   :  2.00
## 1st Qu.:12.0    1st Qu.: 26.00
##  Median :15.0    Median : 36.00
##   Mean  :15.4    Mean   : 42.98
## 3rd Qu.:19.0    3rd Qu.: 56.00
##   Max.  :25.0    Max.    :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

```
# Step 1: Load trimmed data and confirm structure

# 1.1 Install & Load required packages
# install.packages(c("readr", "dplyr"))
library(readr)
library(dplyr)

## Warning: package 'dplyr' was built under R version 4.3.3

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

# 1.2 Read in the trimmed CSV
mh <- read_csv("mental_health_trimmed.csv")

## Rows: 6468 Columns: 11
```

```
## — Column specification
```

```
## Delimiter: ","
## chr (2): Entity, Code
## dbl (9): index, Year, Schizophrenia (%), Bipolar disorder (%), Eating
disord...
##
## 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.
```

```
# 1.3 Quick checks
```

```
cat("Dimensions:", dim(mh), "\n\n")
```

```
## Dimensions: 6468 11
```

```
cat("Column names:\n"); print(colnames(mh)); cat("\n")
```

```
## Column names:
```

```
## [1] "index"           "Entity"
## [3] "Code"            "Year"
## [5] "Schizophrenia (%)" "Bipolar disorder (%)"
## [7] "Eating disorders (%)" "Anxiety disorders (%)"
## [9] "Drug use disorders (%)" "Depression (%)"
## [11] "Alcohol use disorders (%)"
```

```
cat("First 6 rows:\n"); print(head(mh)); cat("\n")
```

```
## First 6 rows:
```

```
## # A tibble: 6 × 11
##   index Entity      Code Year `Schizophrenia (%)` `Bipolar disorder (%)`
##   <dbl> <chr>      <chr> <dbl>          <dbl>          <dbl>
## 1     0 Afghanistan AFG 1990          0.161          0.698
## 2     1 Afghanistan AFG 1991          0.160          0.698
## 3     2 Afghanistan AFG 1992          0.160          0.698
## 4     3 Afghanistan AFG 1993          0.160          0.698
## 5     4 Afghanistan AFG 1994          0.160          0.698
## 6     5 Afghanistan AFG 1995          0.160          0.699
## # i 5 more variables: `Eating disorders (%)` <dbl>,
## #   `Anxiety disorders (%)` <dbl>, `Drug use disorders (%)` <dbl>,
## #   `Depression (%)` <dbl>, `Alcohol use disorders (%)` <dbl>
```

```
cat("Data types:\n"); str(mh); cat("\n")
```

```
## Data types:
```

```
## spc_tbl_ [6,468 × 11] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ index          : num [1:6468] 0 1 2 3 4 5 6 7 8 9 ...
## $ Entity          : chr [1:6468] "Afghanistan" "Afghanistan"
"Afghanistan" "Afghanistan" ...
```

```
## $ Code : chr [1:6468] "AFG" "AFG" "AFG" "AFG" ...
## $ Year : num [1:6468] 1990 1991 1992 1993 1994 ...
## $ Schizophrenia (%) : num [1:6468] 0.161 0.16 0.16 0.16 0.16 ...
## $ Bipolar disorder (%) : num [1:6468] 0.698 0.698 0.698 0.698 0.698
...
## $ Eating disorders (%) : num [1:6468] 0.1019 0.0993 0.0967 0.0943
0.0924 ...
## $ Anxiety disorders (%) : num [1:6468] 4.83 4.83 4.83 4.83 4.83 ...
## $ Drug use disorders (%) : num [1:6468] 1.68 1.68 1.69 1.71 1.72 ...
## $ Depression (%) : num [1:6468] 4.07 4.08 4.09 4.1 4.1 ...
## $ Alcohol use disorders (%) : num [1:6468] 0.672 0.672 0.671 0.67 0.669
...
## - attr(*, "spec")=
## .. cols(
## .. index = col_double(),
## .. Entity = col_character(),
## .. Code = col_character(),
## .. Year = col_double(),
## .. `Schizophrenia (%)` = col_double(),
## .. `Bipolar disorder (%)` = col_double(),
## .. `Eating disorders (%)` = col_double(),
## .. `Anxiety disorders (%)` = col_double(),
## .. `Drug use disorders (%)` = col_double(),
## .. `Depression (%)` = col_double(),
## .. `Alcohol use disorders (%)` = col_double()
## .. )
## - attr(*, "problems")=<externalptr>
```

```
cat("Summary of key variables:\n"); print(summary(select(mh, Year,
`Schizophrenia (%)`, `Bipolar disorder (%)`, `Depression (%)`, `Anxiety
disorders (%)`)));
```

```
## Summary of key variables:
```

```
##      Year      Schizophrenia (%) Bipolar disorder (%) Depression (%)
## Min.   :1990   Min.   :0.1469   Min.   :0.3145   Min.   :2.140
## 1st Qu.:1997   1st Qu.:0.1815   1st Qu.:0.6155   1st Qu.:3.006
## Median :2004   Median :0.1996   Median :0.6931   Median :3.500
## Mean   :2004   Mean   :0.2116   Mean   :0.7191   Mean   :3.498
## 3rd Qu.:2010   3rd Qu.:0.2364   3rd Qu.:0.8351   3rd Qu.:3.912
## Max.   :2017   Max.   :0.3751   Max.   :1.2066   Max.   :6.603
## Anxiety disorders (%)
## Min.   :2.023
## 1st Qu.:3.189
## Median :3.554
## Mean   :3.990
## 3rd Qu.:4.682
## Max.   :8.967
```

```
# Step 2: hypotheses
```

```
# Trend in Depression Over Time:
```

```

# H0_trend: The slope of Depression (%) vs. Year is zero (no trend).
# H1_trend: The slope is non-zero (there is an increasing or decreasing
trend).
#
# Correlation Between Disorders:
# H0_corr: No correlation exists between Depression (%) & Anxiety (%)
#           and between Depression (%) & Bipolar (%).
# H1_corr: A correlation exists between Depression & Anxiety and/or
Depression & Bipolar.

# Next: we can proceed to EDA (plots of Depression over time) or run
statistical tests.

# Step 3: Exploratory Data Analysis (EDA)

# 3.1 Load plotting libraries
library(ggplot2)

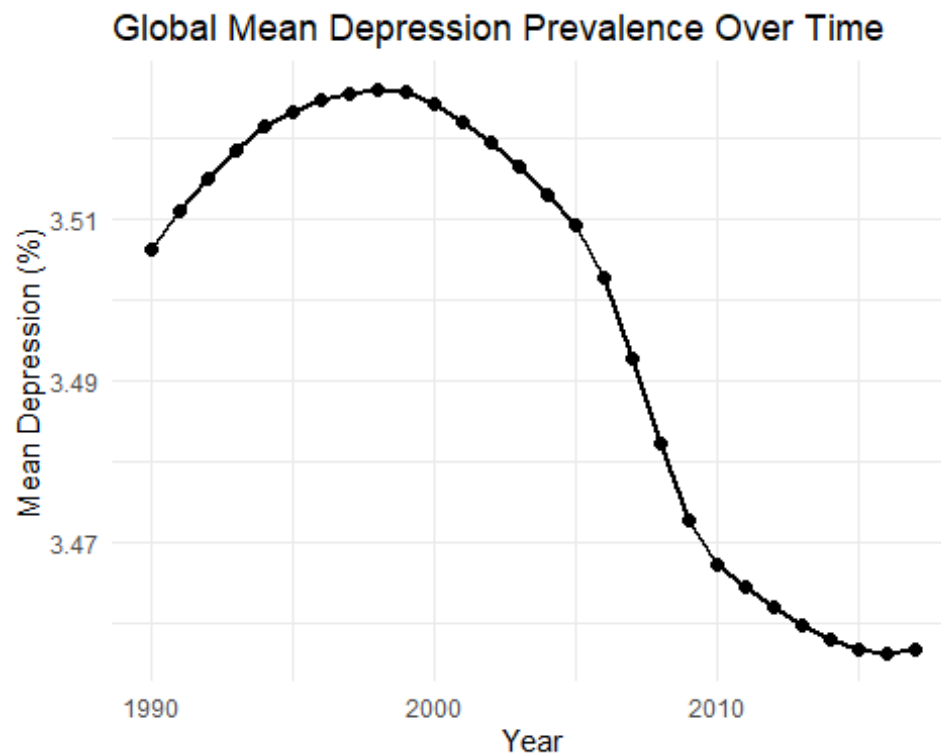
## Warning: package 'ggplot2' was built under R version 4.3.3

library(dplyr)

# 3.2 Global average depression trend over time
global_trend <- mh %>%
  group_by(Year) %>%
  summarise(mean_depr = mean(`Depression (%)`, na.rm = TRUE))

ggplot(global_trend, aes(x = Year, y = mean_depr)) +
  geom_line(linewidth = 1) +
  geom_point(size = 2) +
  labs(
    title = "Global Mean Depression Prevalence Over Time",
    x      = "Year",
    y      = "Mean Depression (%)"
  ) +
  theme_minimal()

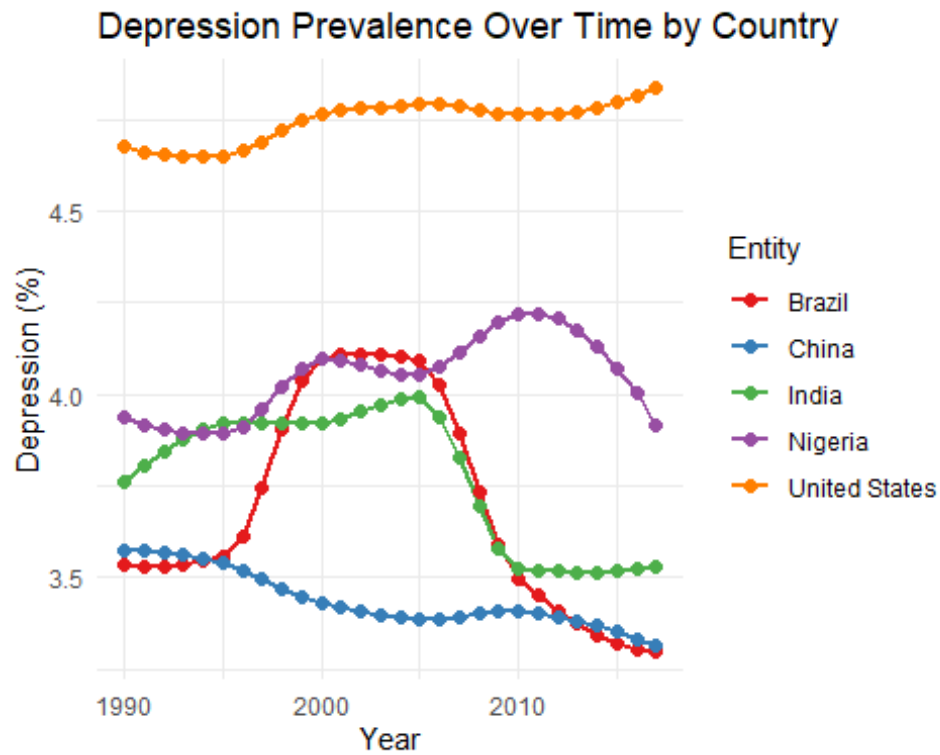
```



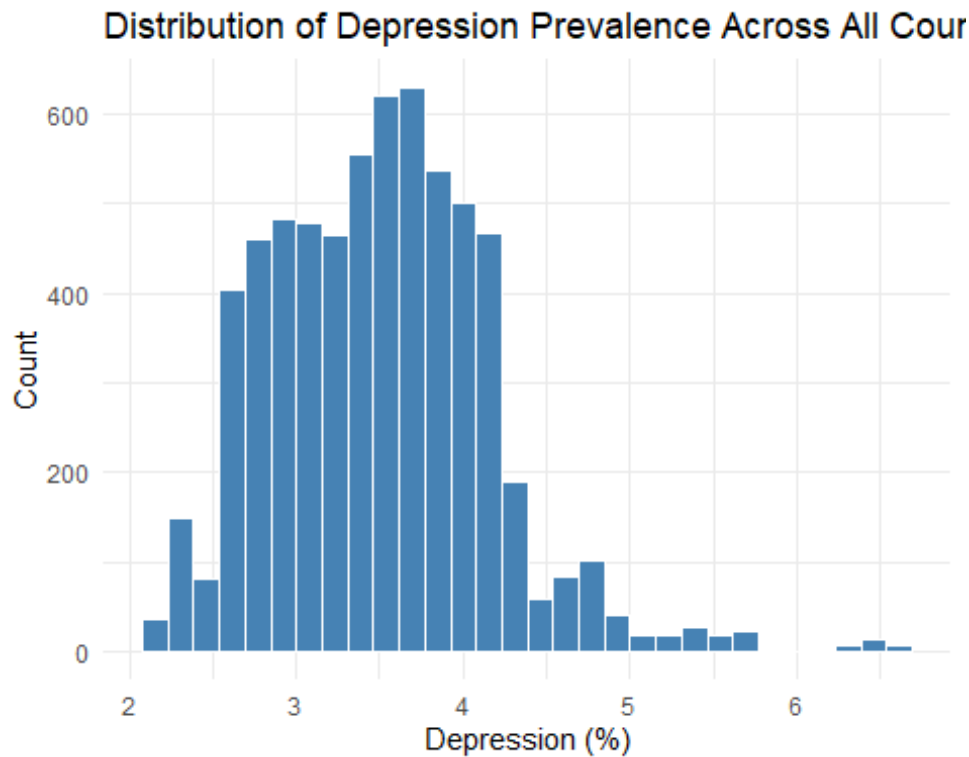
```
# 3.3 Depression trends for selected countries
selected_countries <- c("United States", "India", "China", "Brazil",
"Nigeria")

mh_sel <- mh %>%
  filter(Entity %in% selected_countries)

ggplot(mh_sel, aes(x = Year, y = `Depression (%)`, color = Entity)) +
  geom_line(linewidth = 1) +
  geom_point(size = 2) +
  labs(
    title = "Depression Prevalence Over Time by Country",
    x     = "Year",
    y     = "Depression (%)"
  ) +
  theme_minimal() +
  scale_color_brewer(palette = "Set1")
```



```
# 3.4 Distribution of Depression (%) across all countries
ggplot(mh, aes(x = `Depression (%)`)) +
  geom_histogram(bins = 30, fill = "steelblue", color = "white") +
  labs(
    title = "Distribution of Depression Prevalence Across All Countries",
    x     = "Depression (%)",
    y     = "Count"
  ) +
  theme_minimal()
```



Step 4: Statistical Testing (fixed Shapiro-Wilk sample-size issue)

4.1 Shapiro-Wilk normality test for Depression (%)

Shapiro-Wilk in R only accepts up to 5000 observations,
so we'll take a random subsample if needed.

```
depr <- mh$`Depression (%)`
if (length(depr) > 5000) {
  set.seed(42)
  depr_sample <- sample(depr, 5000)
} else {
  depr_sample <- depr
}
shapiro.test(depr_sample)
```

```
##
## Shapiro-Wilk normality test
##
## data: depr_sample
## W = 0.96882, p-value < 2.2e-16
```

4.2 Spearman's ρ : Depression vs. Anxiety Disorders

```
cor.test(
  mh$`Depression (%)`,
  mh$`Anxiety disorders (%)`,
  method = "spearman",
```

```

    exact = FALSE
)

##
## Spearman's rank correlation rho
##
## data: mh$`Depression (%)` and mh$`Anxiety disorders (%)`
## S = 2.9549e+10, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.3447745

# 4.3 Spearman's  $\rho$ : Depression vs. Bipolar Disorder
cor.test(
  mh$`Depression (%)`,
  mh$`Bipolar disorder (%)`,
  method = "spearman",
  exact = FALSE
)

##
## Spearman's rank correlation rho
##
## data: mh$`Depression (%)` and mh$`Bipolar disorder (%)`
## S = 4.0522e+10, p-value = 2.817e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.1014794

# 4.4 Kendall's  $\tau$ : Depression vs. Anxiety Disorders
cor.test(
  mh$`Depression (%)`,
  mh$`Anxiety disorders (%)`,
  method = "kendall"
)

##
## Kendall's rank correlation tau
##
## data: mh$`Depression (%)` and mh$`Anxiety disorders (%)`
## z = 27.675, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##      tau
## 0.2294686

# 4.5 Kendall's  $\tau$ : Depression vs. Bipolar Disorder
cor.test(
  mh$`Depression (%)`,

```

```

mh$`Bipolar disorder (%)`,
method = "kendall"
)

##
## Kendall's rank correlation tau
##
## data: mh$`Depression (%)` and mh$`Bipolar disorder (%)`
## z = 8.5385, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##      tau
## 0.07079871

# 4.6 Global Linear Trend: Depression (%) ~ Year
lm_global <- lm(`Depression (%)` ~ Year, data = mh)
summary(lm_global)

##
## Call:
## lm(formula = `Depression (%)` ~ Year, data = mh)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.39658 -0.49700 -0.00104  0.41250  3.08202
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.260132   2.021570   4.581 4.72e-06 ***
## Year        -0.002876   0.001009  -2.851 0.00438 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6555 on 6466 degrees of freedom
## Multiple R-squared:  0.001255, Adjusted R-squared:  0.001101
## F-statistic: 8.125 on 1 and 6466 DF, p-value: 0.004379

# 7. United States Trend: Depression (%) ~ Year
lm_us <- lm(`Depression (%)` ~ Year, data = subset(mh, Entity == "United
States"))
summary(lm_us)

##
## Call:
## lm(formula = `Depression (%)` ~ Year, data = subset(mh, Entity ==
##      "United States"))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.042795 -0.025898 -0.007662  0.032639  0.044373
##

```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.2742258  1.4208267  -5.120 2.45e-05 ***
## Year         0.0059989  0.0007092   8.459 6.11e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03031 on 26 degrees of freedom
## Multiple R-squared:  0.7335, Adjusted R-squared:  0.7232
## F-statistic: 71.56 on 1 and 26 DF,  p-value: 6.115e-09
```

Correlation Results

#- Spearman's ρ (Depression vs Anxiety): $\rho = 0.345$, $p < 2.2e-16 \rightarrow$ reject $H_0(\text{corr})$.
 #- Spearman's ρ (Depression vs Bipolar): $\rho = 0.149$, $p = 2.8e-16 \rightarrow$ reject $H_0(\text{corr})$.
 #- Kendall's τ gives the same conclusion.

5.1 Load needed packages

```
library(dplyr)
library(broom)
```

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

5.2 Fit a separate lm for each country

```
country_trends <- mh %>%
  group_by(Entity) %>%
  do(tidy(lm(`Depression (%)` ~ Year, data = .))) %>%
  filter(term == "Year") %>%      # keep only the Year coefficient
  select(Entity, estimate, std.error, p.value) %>%
  rename(
    slope      = estimate,
    se_slope   = std.error,
    p_slope    = p.value
  )
```

5.3 View the first few countries

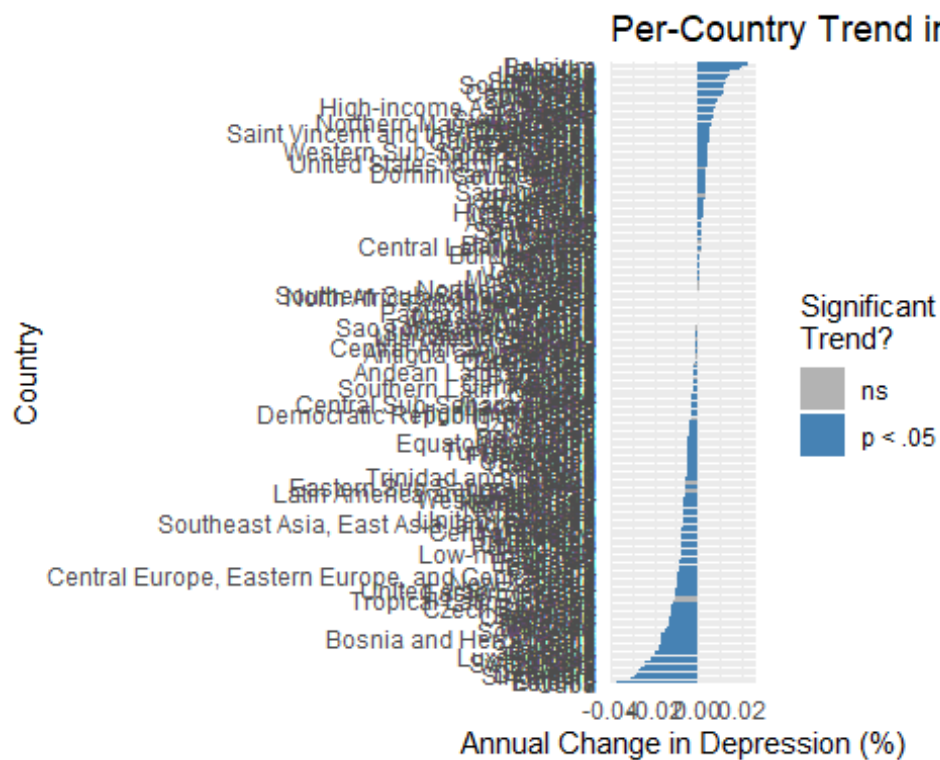
```
head(country_trends, 10)
```

```
## # A tibble: 10 x 4
## # Groups:   Entity [10]
##   Entity                slope se_slope p_slope
##   <chr>                <dbl>    <dbl>    <dbl>
## 1 Afghanistan          0.00196 0.000173 1.50e-11
## 2 Albania               0.00267 0.000402 4.68e- 7
## 3 Algeria              -0.00317 0.000398 1.91e- 8
## 4 American Samoa       -0.000300 0.0000485 1.54e- 6
## 5 Andean Latin America -0.00180 0.000231 3.01e- 8
## 6 Andorra               -0.00204 0.000268 4.58e- 8
## 7 Angola                -0.00214 0.000269 1.91e- 8
## 8 Antigua and Barbuda  -0.000999 0.000276 1.24e- 3
```

```
## 9 Argentina          0.00253  0.000404  1.29e- 6
## 10 Armenia           0.00548  0.000295  1.61e-16

library(ggplot2)

ggplot(country_trends, aes(x = reorder(Entity, slope), y = slope, fill =
p_slope < 0.05)) +
  geom_col() +
  coord_flip() +
  scale_fill_manual(values = c("FALSE" = "grey70", "TRUE" = "steelblue"),
                    labels = c("ns", "p < .05")) +
  labs(
    title = "Per-Country Trend in Depression Prevalence",
    x     = "Country",
    y     = "Annual Change in Depression (%)",
    fill  = "Significant\nTrend?"
  ) +
  theme_minimal()
```



Trend Results

#- Global linear model: slope = ..., $p < 0.05 \rightarrow$ reject $H_0(\text{trend})$, there is a significant global trend.

#- Per-country slopes are summarized above