

# EDA

2025-06-15

```
library(readxl)
library(ggplot2)
library(dplyr)
```

```
##
## 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
```

```
library(maps)
library(viridis)
```

```
## Loading required package: viridisLite

##
## Attaching package: 'viridis'

## The following object is masked from 'package:maps':
##
##   unemp
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v forcats   1.0.0      v stringr   1.5.1
## v lubridate 1.9.4      v tibble   3.2.1
## v purrr     1.0.4      v tidyr    1.3.1
## v readr     2.1.5
```

```
## -- Conflicts ----- tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## x purrr::map()     masks maps::map()
```

```
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(DataExplorer)
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
library(ggpubr)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
```

```
##   +.gg   ggplot2
library(reshape2)

##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##   smiths
library(scales)

##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##   discard
##
## The following object is masked from 'package:readr':
##
##   col_factor
##
## The following object is masked from 'package:viridis':
##
##   viridis_pal
data <- read.csv("C:/Users/s19444/Downloads/Nutrition.csv")
head(data)

##      Country Year UN_Region UN_sub_region Stunting_survey Stunting_model
## 1 Afghanistan 2000      Asia Southern Asia              NA             54.9
## 2 Afghanistan 2001      Asia Southern Asia              NA             55.0
## 3 Afghanistan 2002      Asia Southern Asia              NA             54.7
## 4 Afghanistan 2003      Asia Southern Asia              NA             54.3
## 5 Afghanistan 2004      Asia Southern Asia             59.3             53.8
## 6 Afghanistan 2005      Asia Southern Asia              NA             53.2
## Underweight_survey Wasting_survey Overweight_survey Overweight_model
## 1                NA                NA                NA             6.5
## 2                NA                NA                NA             6.4
## 3                NA                NA                NA             6.4
## 4                NA                NA                NA             6.5
## 5             31.7             9.1             6.5             6.5
## 6                NA                NA                NA             6.5
## WHZ_sample_size HAZ_sample_size WAZ_sample_size Under5_population..000.
## 1                NA                NA                NA             NA
## 2                NA                NA                NA             NA
## 3                NA                NA                NA             NA
## 4                NA                NA                NA             NA
## 5             946             946                NA          4705.37
## 6                NA                NA                NA             NA
## hcpi_a fcpi_a def_a Literacy_female Total_literacy
## 1      0.0    NA    NA                NA                NA
## 2    -43.4    NA    NA                NA                NA
## 3     51.9    NA    NA                NA                NA
```

```

## 4    35.7    NA  13.2            NA            NA
## 5    16.4    NA  11.2            NA            NA
## 6    10.6    NA  10.7            NA            NA
##    Current..health.expenditure....of.GDP.
## 1                                     NA
## 2                                     NA
## 3                                9.443391
## 4                                8.941258
## 5                                9.808474
## 6                                9.948289
##    Urban.population....of.total.population.
## 1                                22.078
## 2                                22.169
## 3                                22.261
## 4                                22.353
## 5                                22.500
## 6                                22.703
##    Rural.population....of.total.population. Mortality_rate_under.5
## 1                                77.922            131.6
## 2                                77.831            127.4
## 3                                77.739            123.0
## 4                                77.647            118.5
## 5                                77.500            114.0
## 6                                77.297            109.5
##    Low.birthweight.babies....of.births.
## 1                                NA
## 2                                NA
## 3                                NA
## 4                                NA
## 5                                NA
## 6                                NA
##    Unemployment..total....of.total.labor.force.
## 1                                7.955
## 2                                7.958
## 3                                7.939
## 4                                7.922
## 5                                7.914
## 6                                7.914

```

```
colSums(is.na(data))
```

```

##                                Country
##                                0
##                                Year
##                                0
##                                UN_Region
##                                0
##                                UN_sub_region
##                                0
##                                Stunting_survey
##                                819
##                                Stunting_model
##                                1
##                                Underweight_survey
##                                823

```

```

##           Wasting_survey
##           822
##       Overweight_survey
##           846
##       Overweight_model
##           0
##       WHZ_sample_size
##           865
##       HAZ_sample_size
##           860
##       WAZ_sample_size
##           892
##       Under5_population..000.
##           824
##           hcpi_a
##           0
##           fcpi_a
##           98
##           def_a
##           72
##       Literacy_female
##           825
##       Total_literacy
##           834
##       Current..health.expenditure....of.GDP.
##           67
##       Urban.population....of.total.population.
##           0
##       Rural.population....of.total.population.
##           0
##       Mortality_rate_under.5
##           0
##       Low.birthweight.babies....of.births.
##           365
##       Unemployment..total....of.total.labor.force.
##           194

colnames(data)[1:18] <- c("Country", "Year", "UN_Region", "UN_sub_region", "Stunting_survey", "Stunting_m

## Warning in colnames(data)[1:18] <- c("Country", "Year", "UN_Region",
## "UN_sub_region", : number of items to replace is not a multiple of replacement
## length

#Selected countries
selected_countries <- c("Afghanistan", "Bangladesh", "Bolivia", "Botswana", "Brazil", "Chilie", "Colombi

library(ggplot2)
library(rnaturalearth)
library(rnaturalearthdata)

##
## Attaching package: 'rnaturalearthdata'

## The following object is masked from 'package:rnaturalearth':
##
##     countries110

```

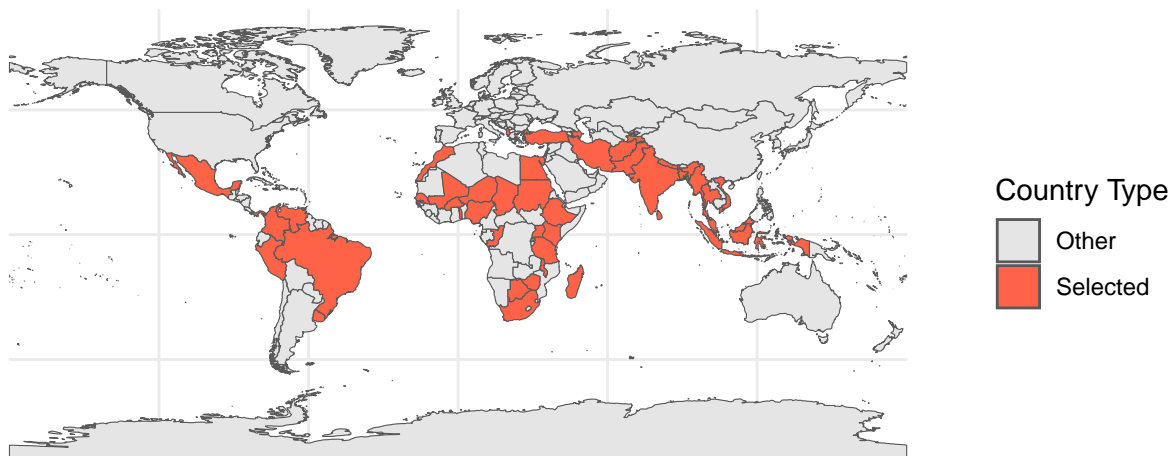
```
library(sf)

## Linking to GEOS 3.13.0, GDAL 3.10.1, PROJ 9.5.1; sf_use_s2() is TRUE
world <- ne_countries(scale = "medium", returnclass = "sf")

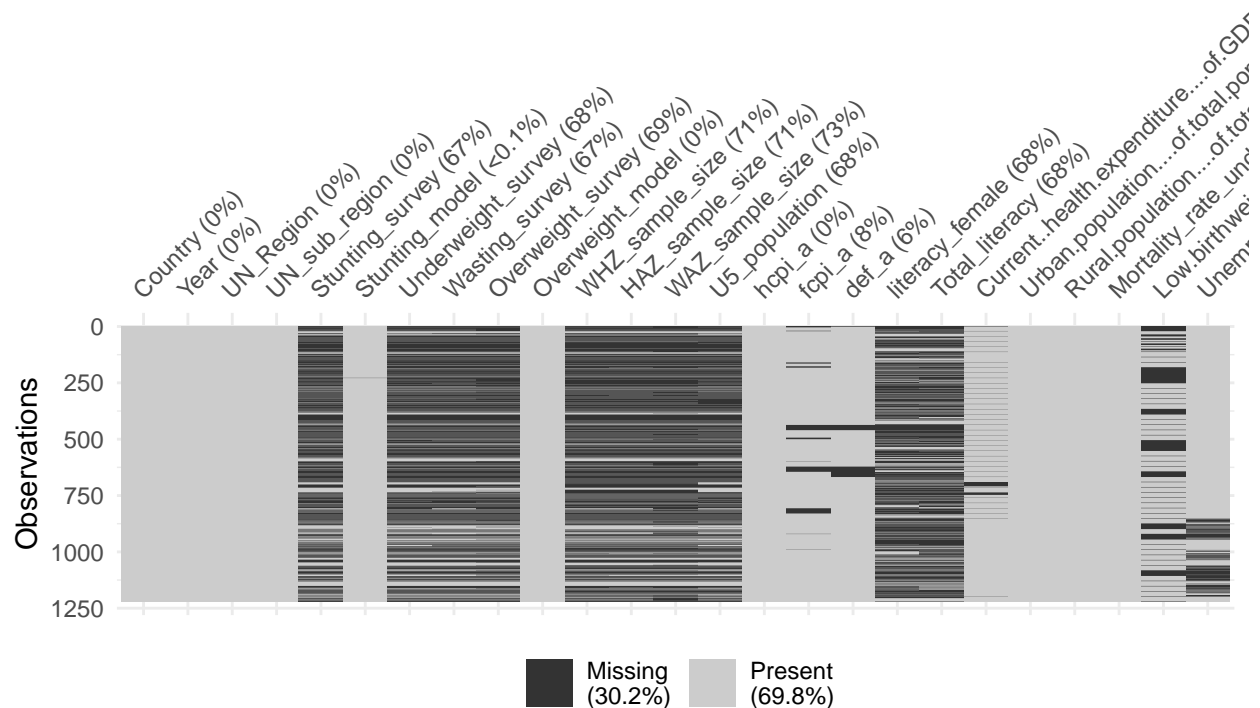
world$highlight <- ifelse(world$name %in% selected_countries, "Selected", "Other")

ggplot(data = world) +
  geom_sf(aes(fill = highlight)) +
  scale_fill_manual(values = c("Selected" = "tomato", "Other" = "gray90")) +
  theme_minimal() +
  labs(title = "Highlighted Countries on World Map",
       fill = "Country Type")
```

Highlighted Countries on World Map



```
#Missing values
library(naniar)
vis_miss(data)
```



```
miss_var_summary(data)
```

```
## # A tibble: 25 x 3
##   variable      n_miss pct_miss
##   <chr>      <int>   <num>
## 1 WAZ_sample_size      892    73.2
## 2 WHZ_sample_size      865    71.0
## 3 HAZ_sample_size      860    70.5
## 4 Overweight_survey    846    69.4
## 5 Total_literacy       834    68.4
## 6 literacy_female      825    67.7
## 7 U5_population        824    67.6
## 8 Underweight_survey   823    67.5
## 9 Wasting_survey       822    67.4
## 10 Stunting_survey     819    67.2
## # i 15 more rows
```

```
colMeans(is.na(data)) * 100
```

```
##           Country
##           0.0000000
##           Year
##           0.0000000
##           UN_Region
##           0.0000000
##           UN_sub_region
##           0.0000000
```

```
##           Stunting_survey
##           67.18621821
##           Stunting_model
##           0.08203445
##           Underweight_survey
##           67.51435603
##           Wasting_survey
##           67.43232158
##           Overweight_survey
##           69.40114848
##           Overweight_model
##           0.00000000
##           WHZ_sample_size
##           70.95980312
##           HAZ_sample_size
##           70.54963084
##           WAZ_sample_size
##           73.17473339
##           U5_population
##           67.59639048
##           hcpi_a
##           0.00000000
##           fcpi_a
##           8.03937654
##           def_a
##           5.90648072
##           literacy_female
##           67.67842494
##           Total_literacy
##           68.41673503
##           Current..health.expenditure....of.GDP.
##           5.49630845
##           Urban.population....of.total.population.
##           0.00000000
##           Rural.population....of.total.population.
##           0.00000000
##           Mortality_rate_under.5
##           0.00000000
##           Low.birthweight.babies....of.births.
##           29.94257588
##           Unemployment..total....of.total.labor.force.
##           15.91468417
```

By omitting the predictor variables more than 50% of missing values, the selected socioeconomic factors are:

hcpi\_a fcpi\_a def\_a Current health expenditure(% of GDP) urban population(% of total population) mortality rate under 5 Low birth weight babies unemployment rate

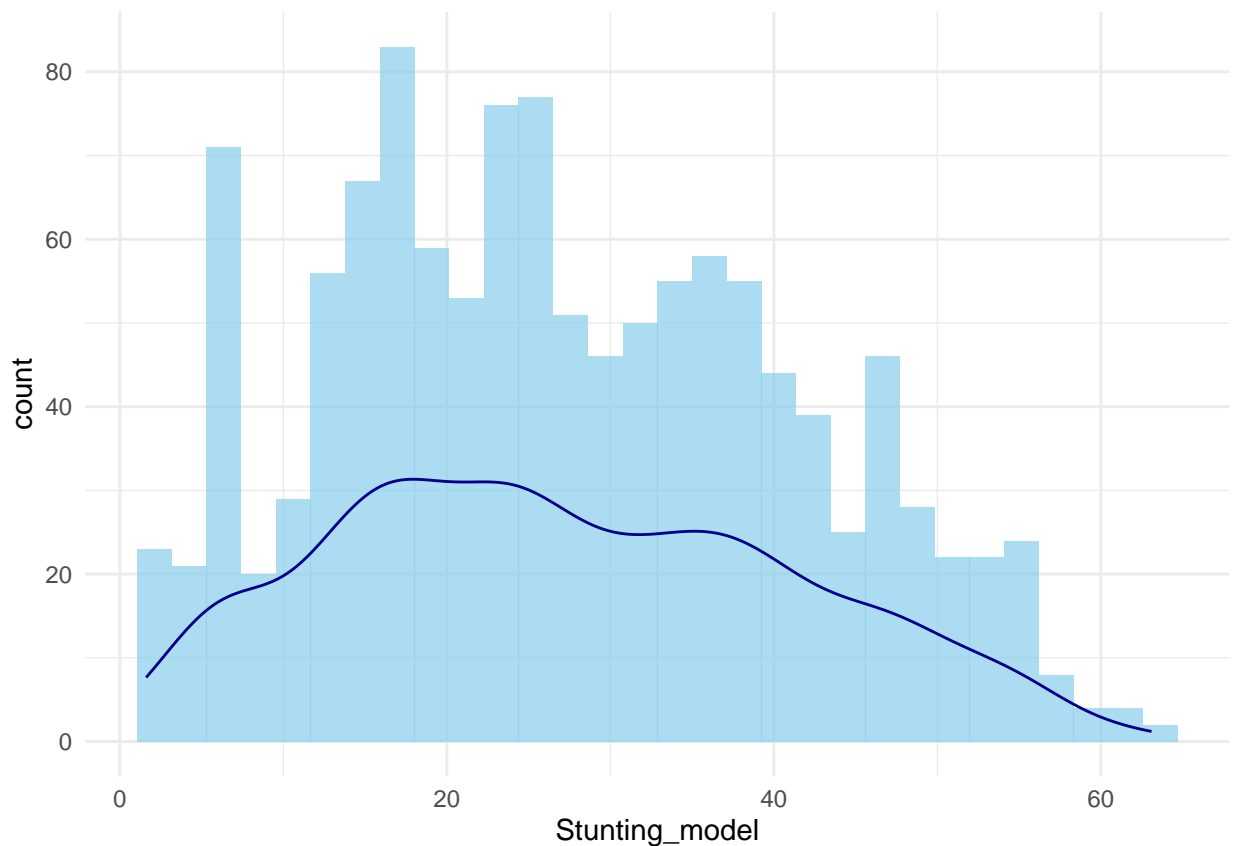
## Distribution of response variables

Stunting

```
# Histograms & density
ggplot(data, aes(x = Stunting_model)) +
  geom_histogram(bins = 30, fill = "skyblue", alpha = 0.7) +
```

```
geom_density(aes(y = ..count..), color = "darkblue") +  
theme_minimal()
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.  
## i Please use `after_stat(count)` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was  
## generated.  
  
## Warning: Removed 1 row containing non-finite outside the scale range  
## (`stat_bin()`).  
  
## Warning: Removed 1 row containing non-finite outside the scale range  
## (`stat_density()`).
```



The distribution is multi-modal (Multiple peaks) This indicates that different countries or time periods may cluster around different stunting levels.

The highest concentrations of observations appear between 10% and 40% with several mini-peaks

The distribution has a slightly right skew: there are some countries or years with very high stunting percentages, though most values are in the lower-to-mid range.

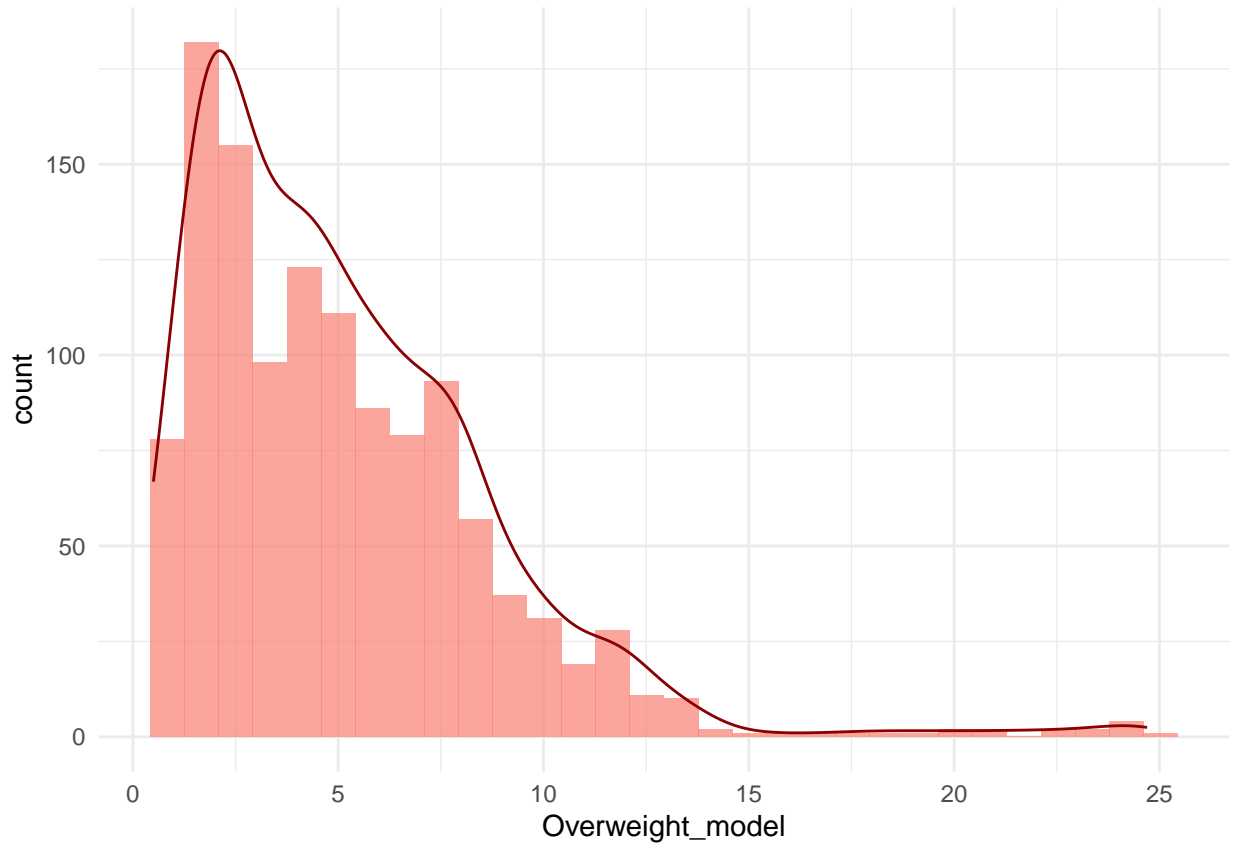
Long tail extending beyond 50%: high stunting in certain countries or time periods, which might act as outliers in modeling.

overweight

```
ggplot(data, aes(x = Overweight_model)) +  
geom_histogram(bins = 30, fill = "salmon", alpha = 0.7) +
```



```
geom_density(aes(y = ..count..), color = "darkred") +  
theme_minimal()
```



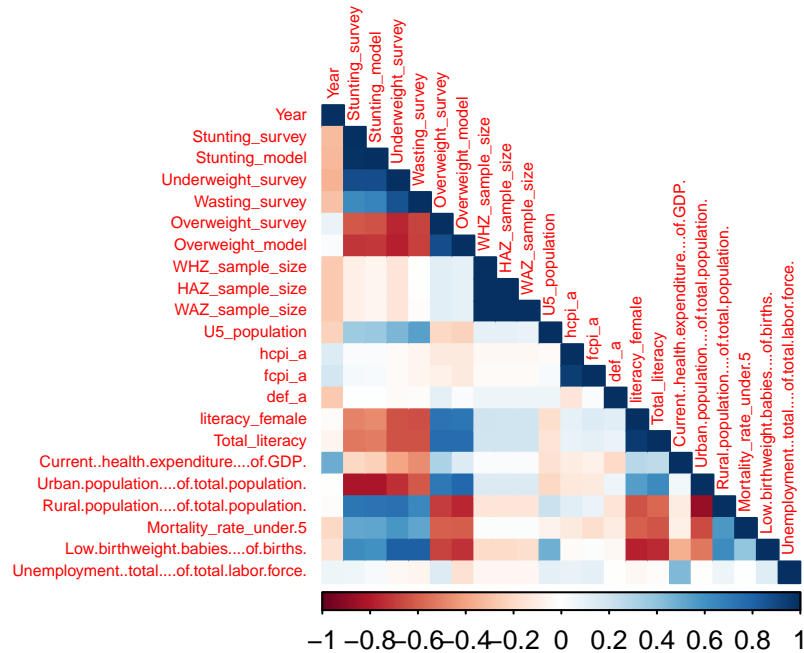
Right skewed

There is one clear peak around 0-2% overweight prevalence. (unimodal) not normally distributed the most values appear to be around 0-2%

#Correlation Analysis

```
df_numeric <- data %>%  
  select(where(is.numeric)) %>%  
  drop_na()
```

```
cor_matrix <- cor(df_numeric, use = "complete.obs")  
corrplot(cor_matrix, method = "color", type = "lower", tl.cex = 0.5)
```



According to the correlation plot,

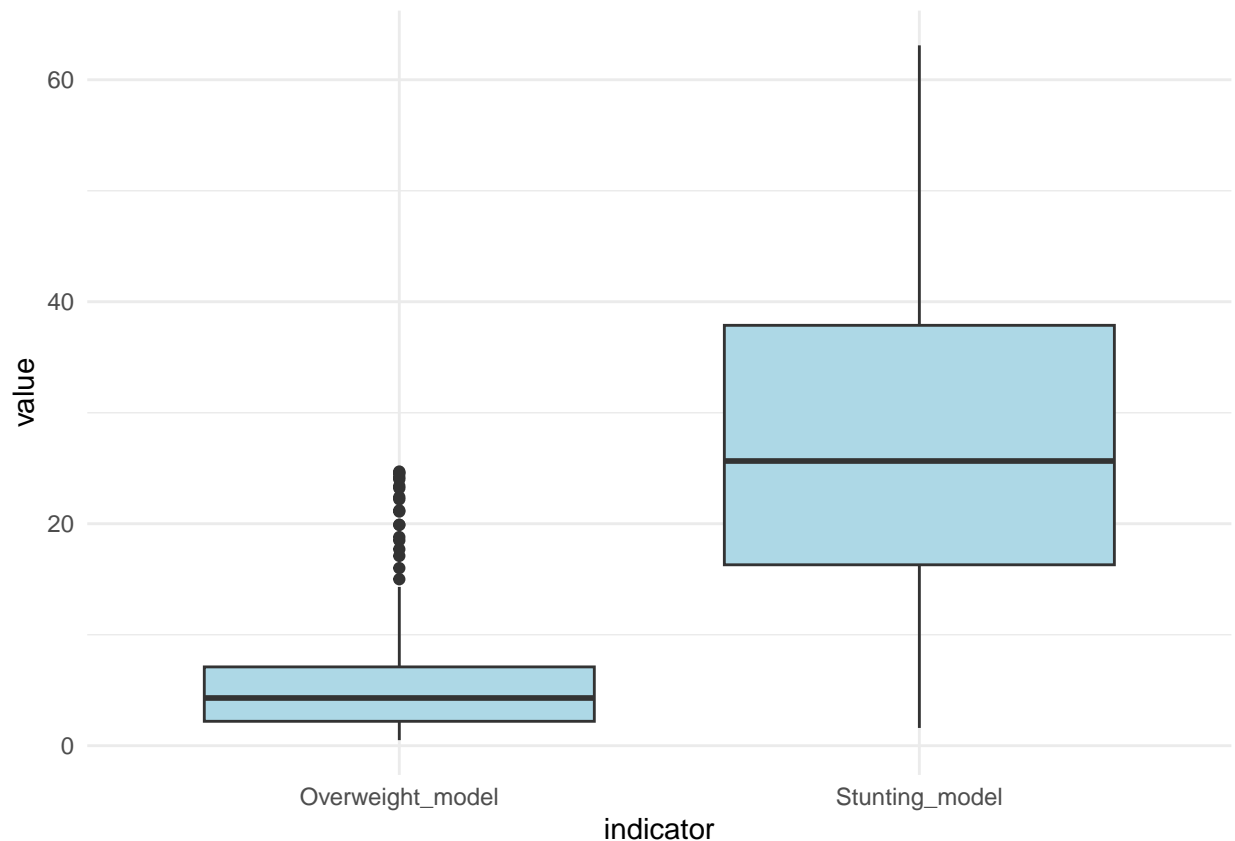
Overweight modeled values are positively correlated with socioeconomic indicators literacy rate, current health expenditure, urban population and negatively correlated with the variables hcpi\_a, fcpi\_a, def\_a, mortality rate under 5, rural population, low birth weight rate, unemployment rate

stunting modeled values are negatively correlated with fcpi\_a, hcpi\_a, def\_a, literacy, current health expenditure, urban population, rural population and positively correlated with mortality rate, low birth weight babies, unemployment rate

#Outlier detection

```
data %>%
  select(Country, Stunting_model, Overweight_model) %>%
  pivot_longer(cols = c(Stunting_model, Overweight_model), names_to = "indicator") %>%
  ggplot(aes(x = indicator, y = value)) +
  geom_boxplot(fill = "lightblue") +
  theme_minimal()
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_boxplot()`).
```



Box plots compare the distributions of overweight and stunting rates across countries.

overweight:

median is around 5-6% and the IQR is narrow multiple outliers clustered around 15-25%, indicating several countries with unusually high overweight rates

stunting:

median is around 25-27% and the IQR is much wider Whiskers extend from about 2% to roughly 62% no visible outliers, suggesting the whiskers capture the full range more normally distributed

patterns:

most countries face low overweight but moderate-to-high stunting few countries have severe overweight problems the dual burden exists but is asymmetric across the global landscape

##By country

Bar chart of average by country

```
data %>%
  group_by(Country) %>%
  summarise(across(c(Stunting_model, Overweight_model), mean, na.rm = TRUE)) %>%
  pivot_longer(cols = -Country, names_to = "indicator", values_to = "mean_val") %>%
  ggplot(aes(x = reorder(Country, mean_val), y = mean_val, fill = indicator)) +
  geom_bar(stat = "identity", position = "dodge") +
  coord_flip() +
  theme_minimal()
```

## Warning: There was 1 warning in `summarise()`.

```
## i In argument: `across(c(Stunting_model, Overweight_model), mean, na.rm =
## TRUE)`.
```

```
## i In group 1: `Country = "Afganistan"`.
```

```
## Caused by warning:
```

```
## ! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.
```

```
## Supply arguments directly to `.fns` through an anonymous function instead.
```

```
##
```

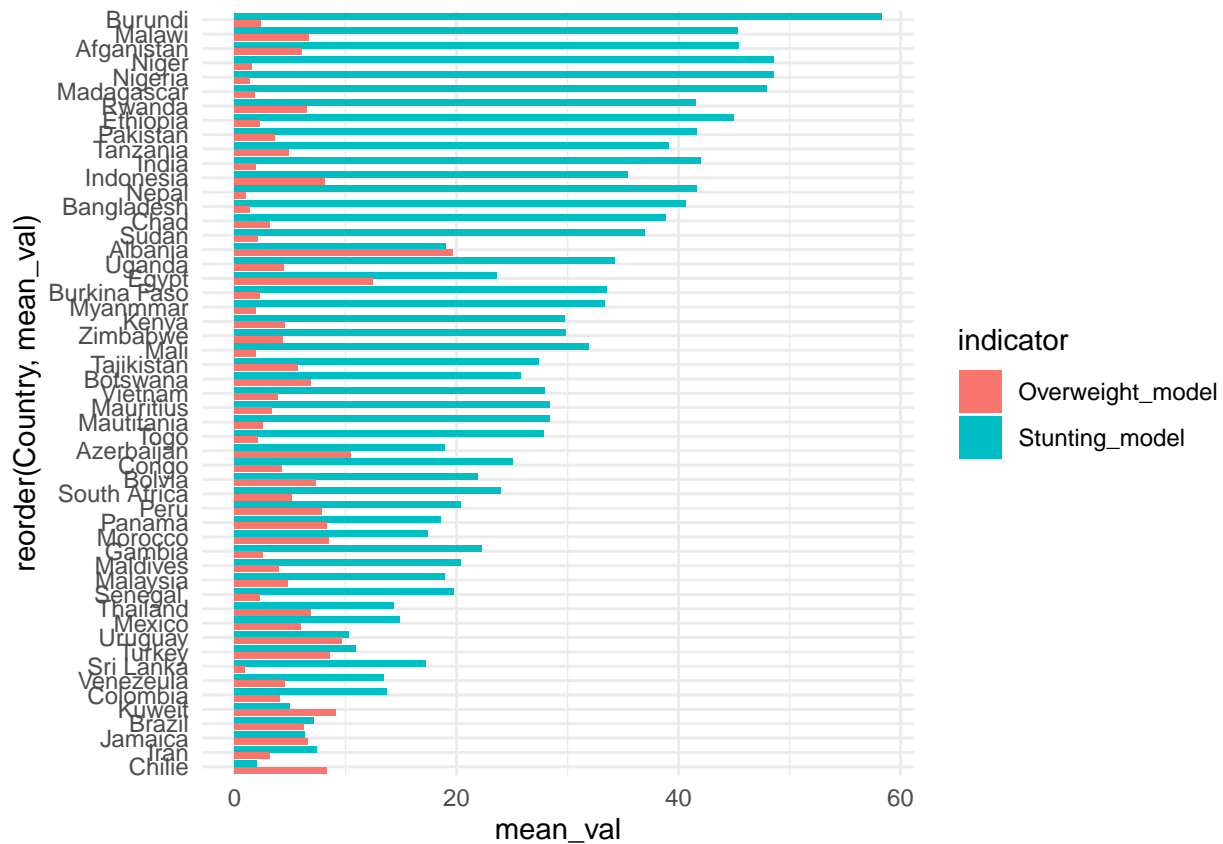
```
## # Previously
```

```
## across(a:b, mean, na.rm = TRUE)
```

```
##
```

```
## # Now
```

```
## across(a:b, \(x) mean(x, na.rm = TRUE))
```



High prevalence in stunting than the overweight across the selected countries. Countries at the top (Burundi, Afghanistan... ) have high combined rates of stunting and overweight.

Dual burden pattern can be seen: some countries show significant levels of both indicators.Ex: Egypt, South Africa. This indicates the prevalence of double burden of malnutrition.

Transition pattern: As move down the chart, there is generally a shift from stunting-dominant to more balanced or overweight-dominant patterns, but this is not perfectly linear

This suggests different countries face different malnutrition challenges.

some countries are primarily dealing with undernutrition (stunting), others with a mixed burden.

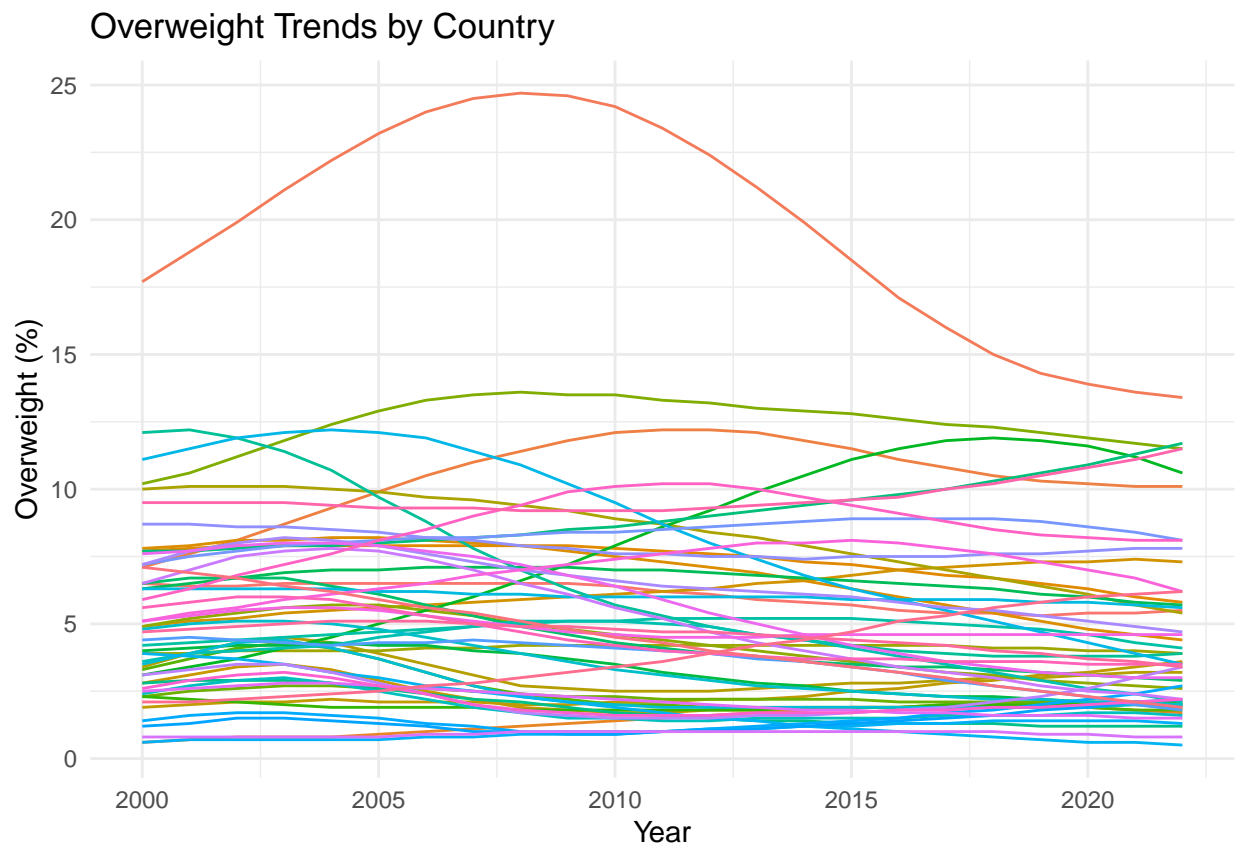
#Overweight

```

overweight_data <- data %>%
  select(Country, Year, Overweight_model) %>%
  filter(!is.na(Overweight_model))
#overweight_data

ggplot(data, aes(x = Year, y = Overweight_model, group = Country, color = Country)) +
  geom_line(show.legend = FALSE) +
  theme_minimal() +
  labs(title = "Overweight Trends by Country", y = "Overweight (%)")

```



```

countries <- unique(overweight_data$Country)

for(country in countries) {
  country_data <- overweight_data %>% filter(Country == country)

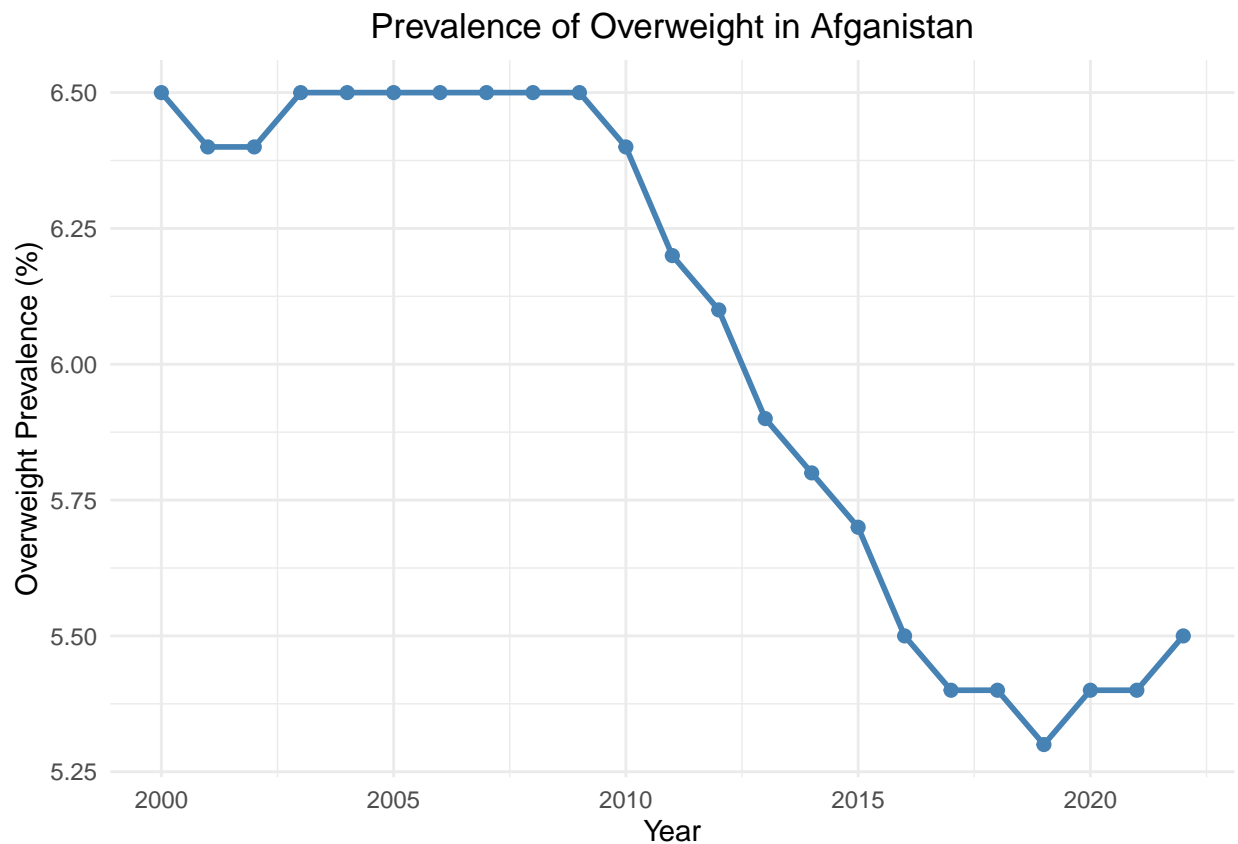
  p <- ggplot(country_data, aes(x = Year, y = Overweight_model)) +
    geom_line(color = "steelblue", size = 1) +
    geom_point(color = "steelblue", size = 2) +
    labs(title = paste("Prevalence of Overweight in", country),
         x = "Year",
         y = "Overweight Prevalence (%)") +
    theme_minimal() +
    theme(plot.title = element_text(hjust = 0.5))

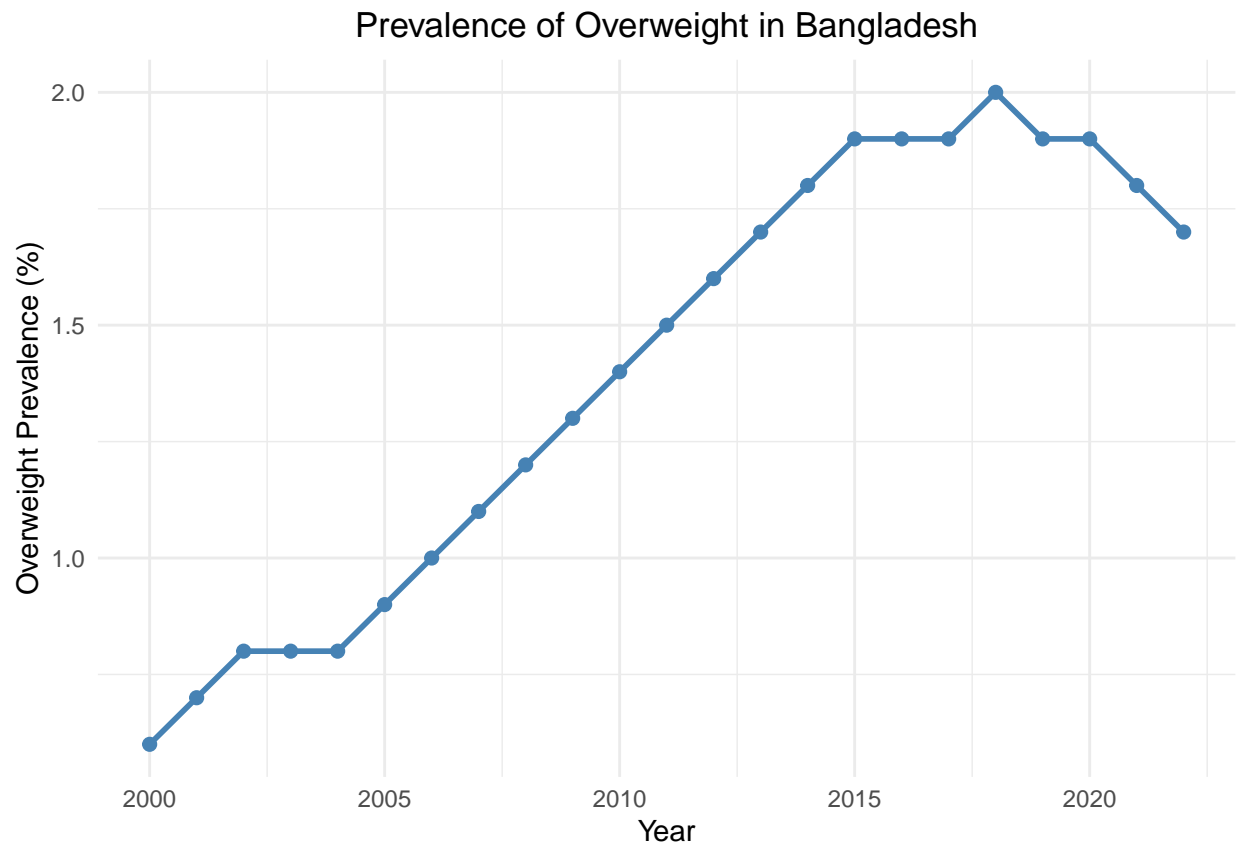
  print(p)
}

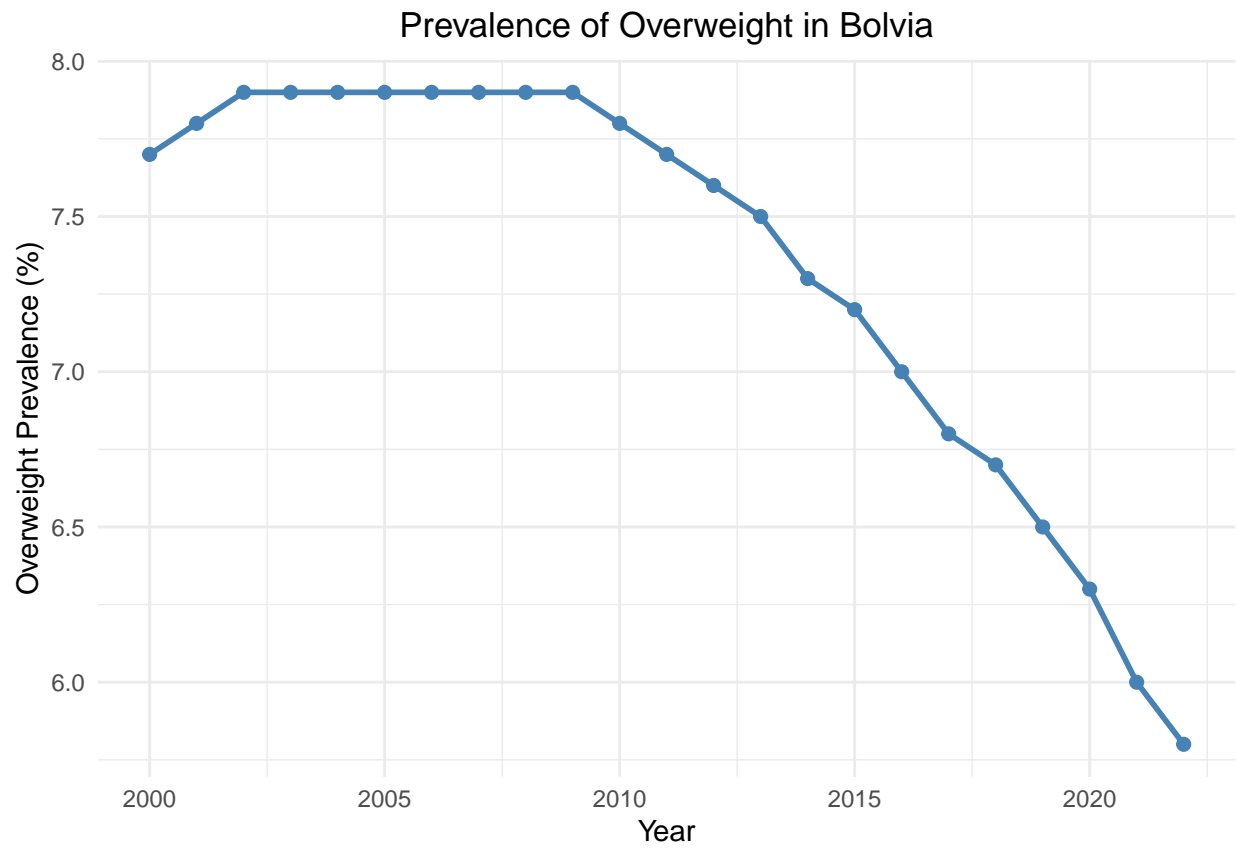
```

```
ggsave(paste0("overweight_", gsub(" ", "_", country), ".png"), plot = p, width = 8, height = 6)
}
```

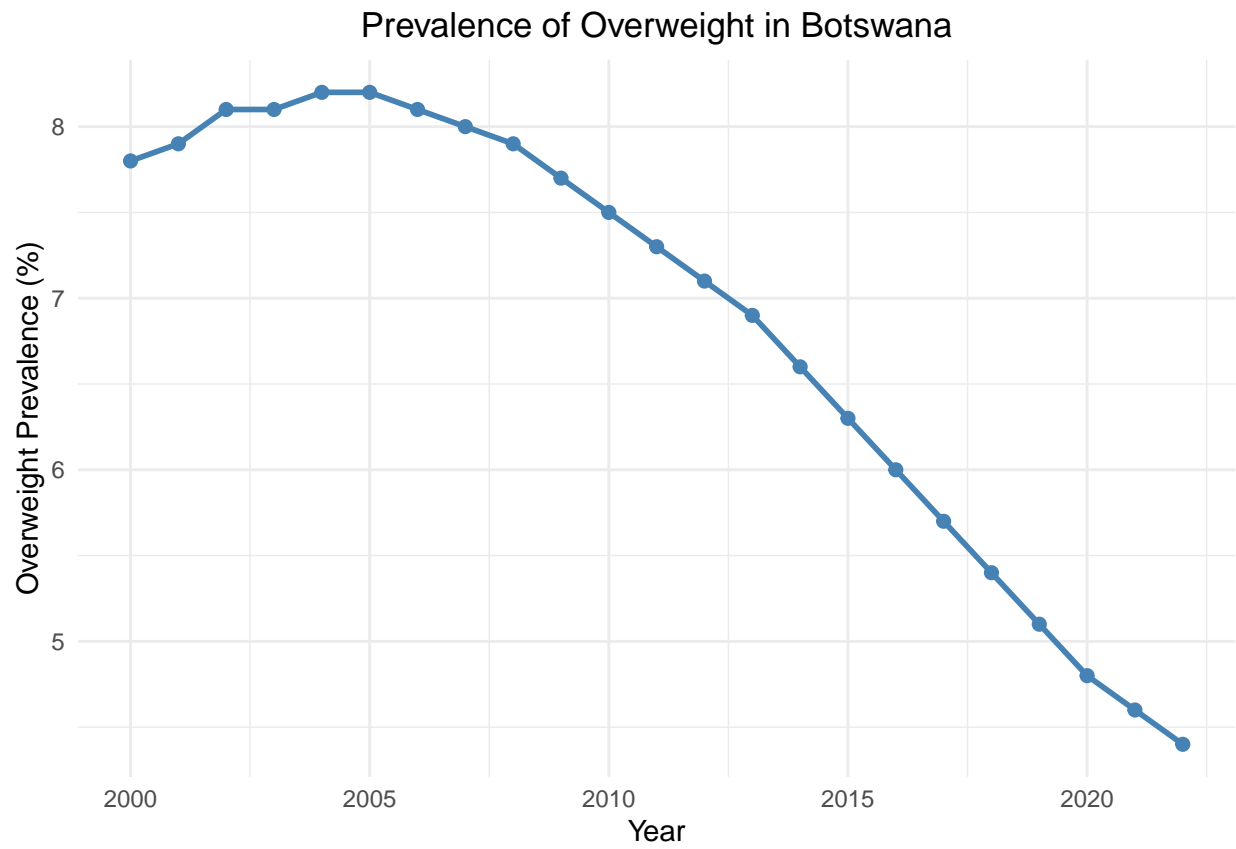
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

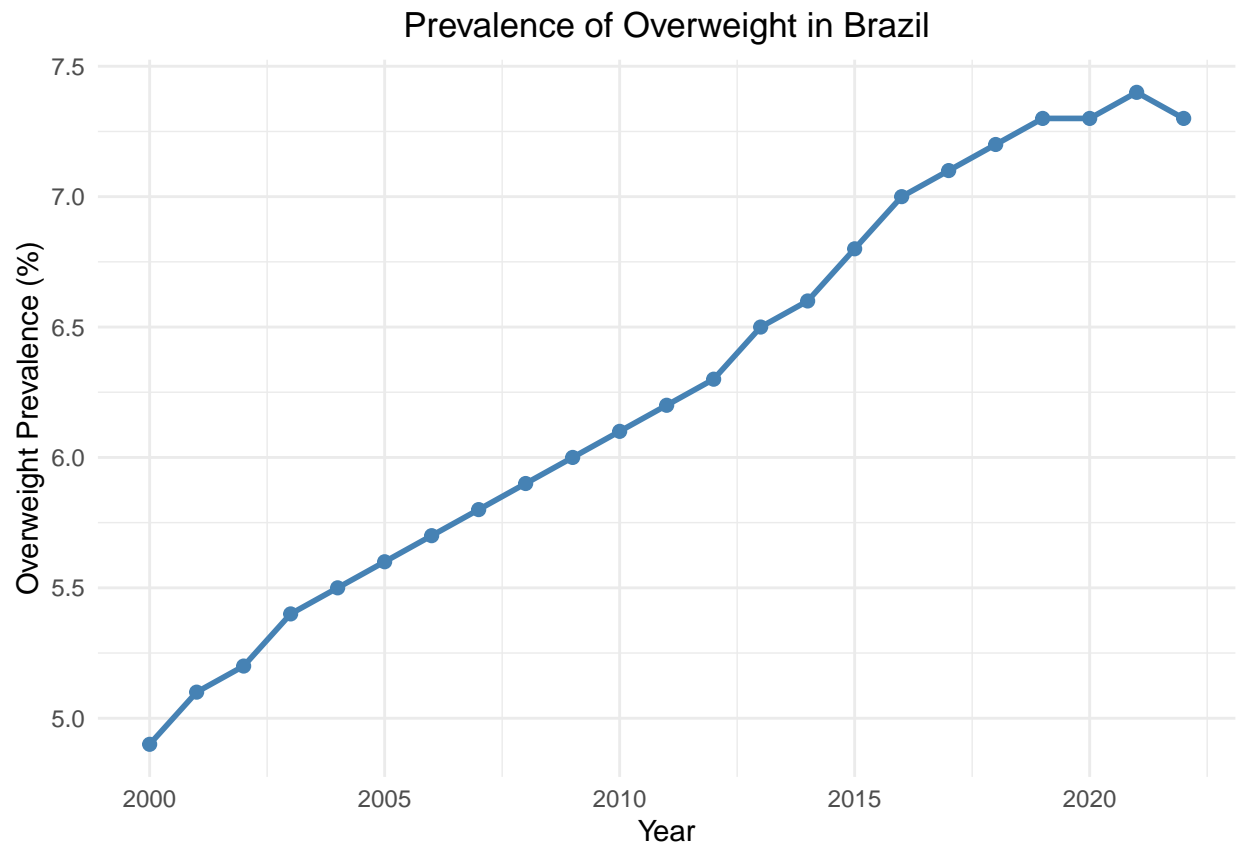


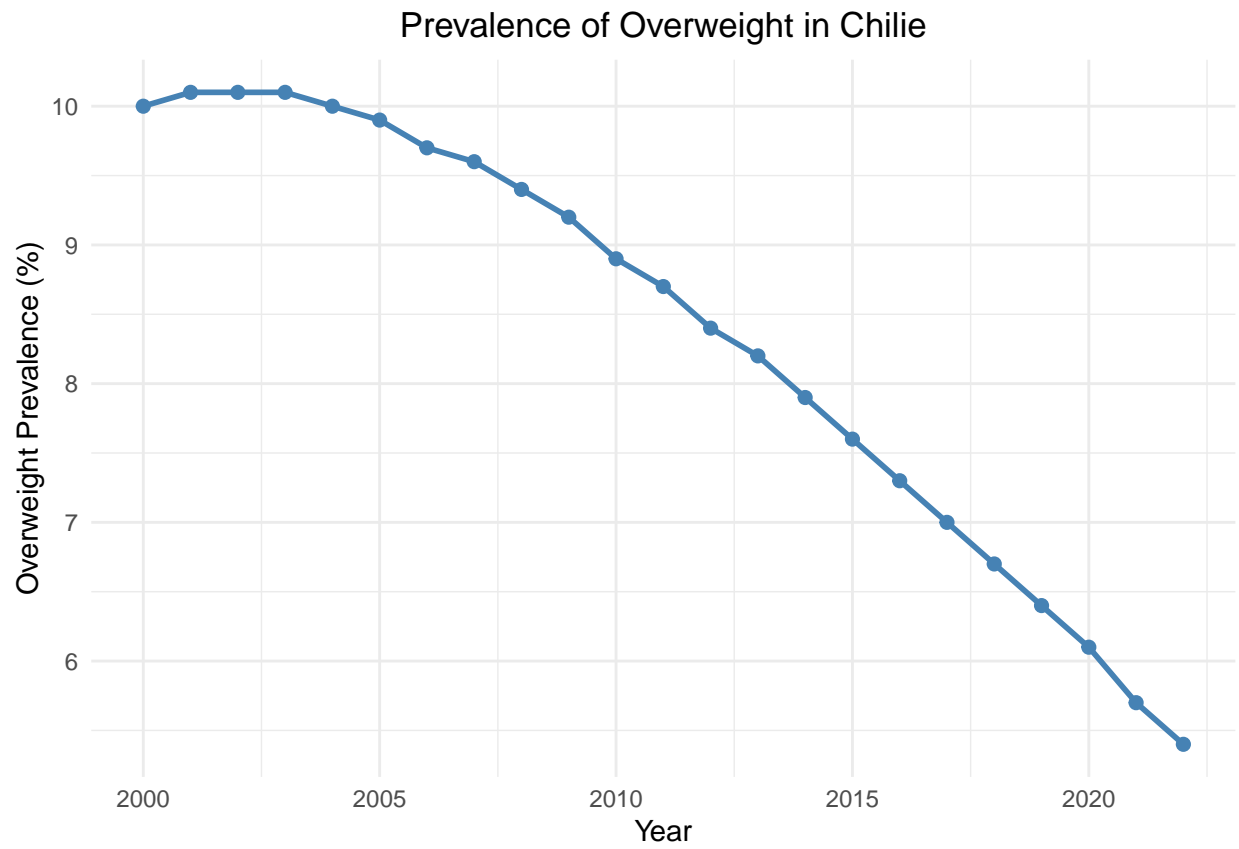


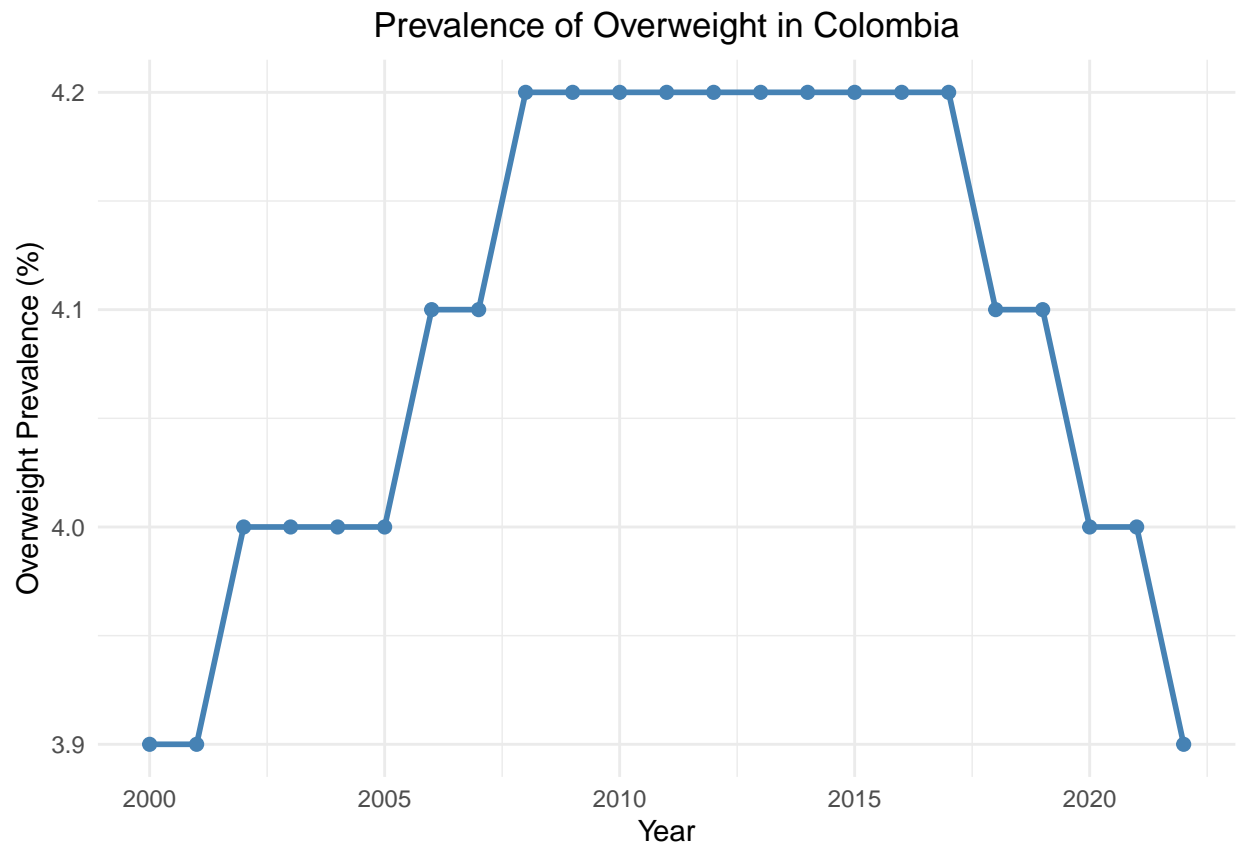


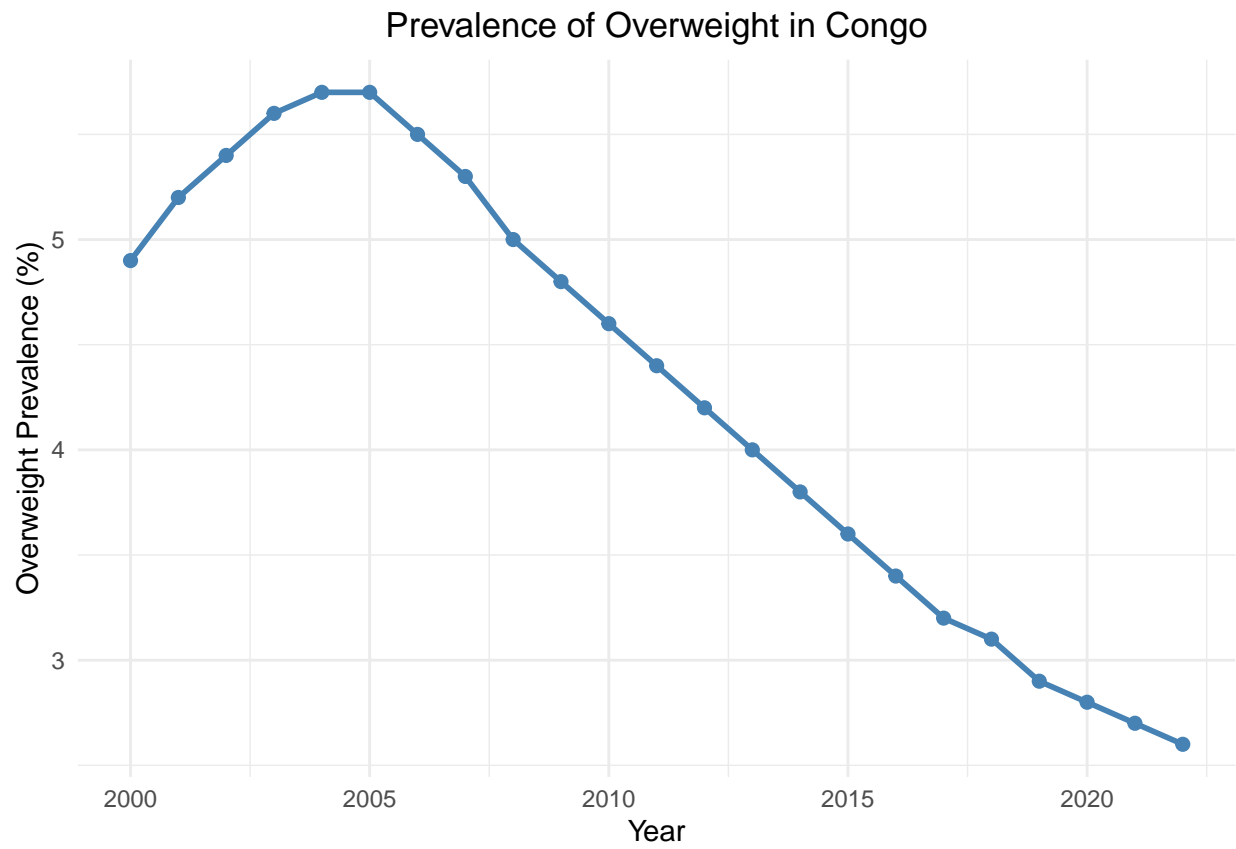


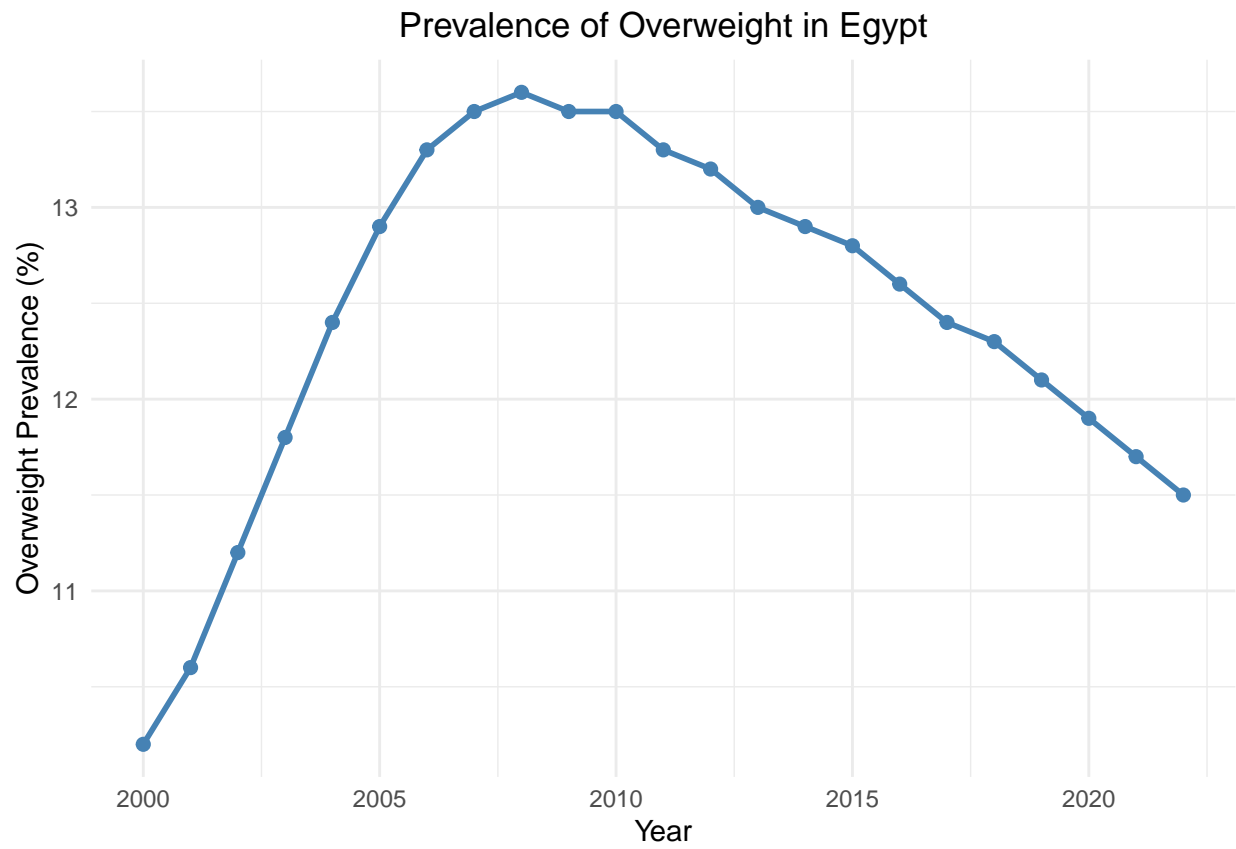


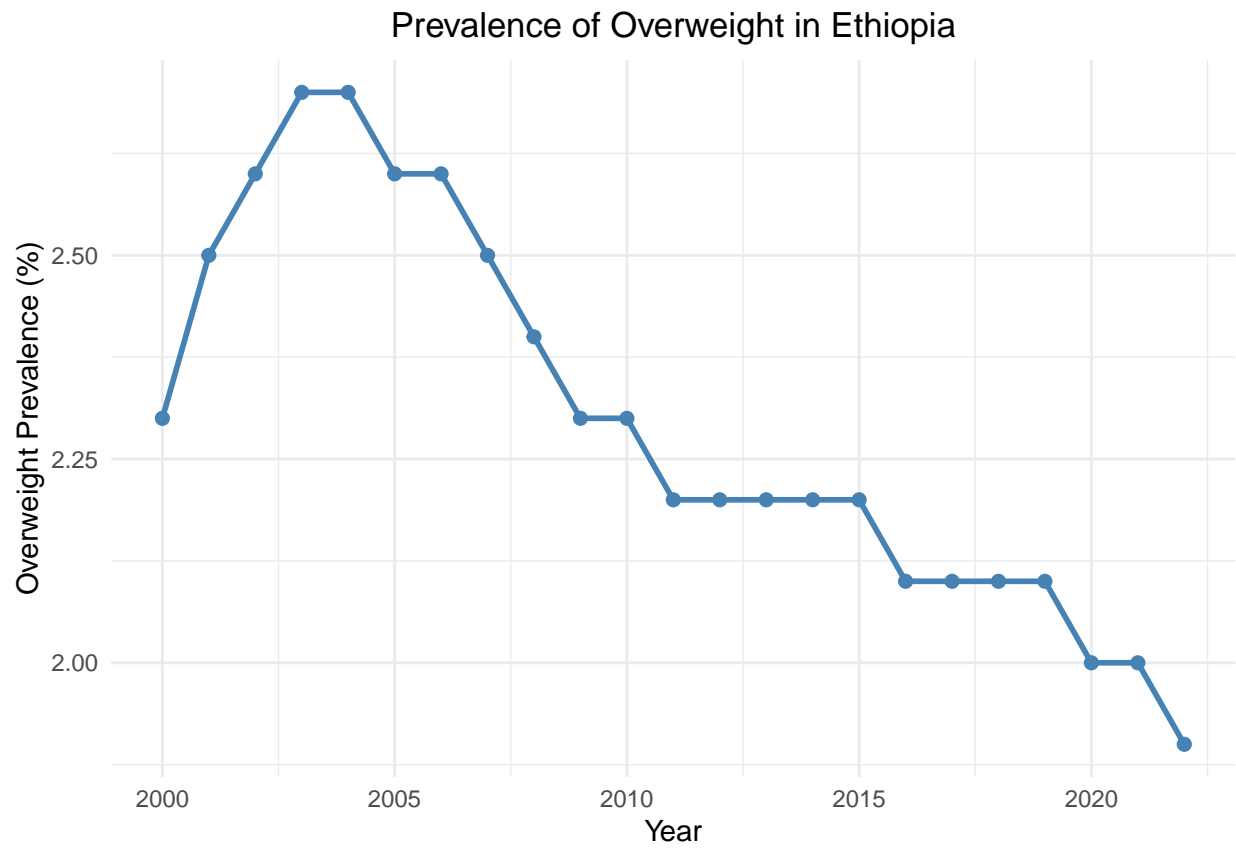


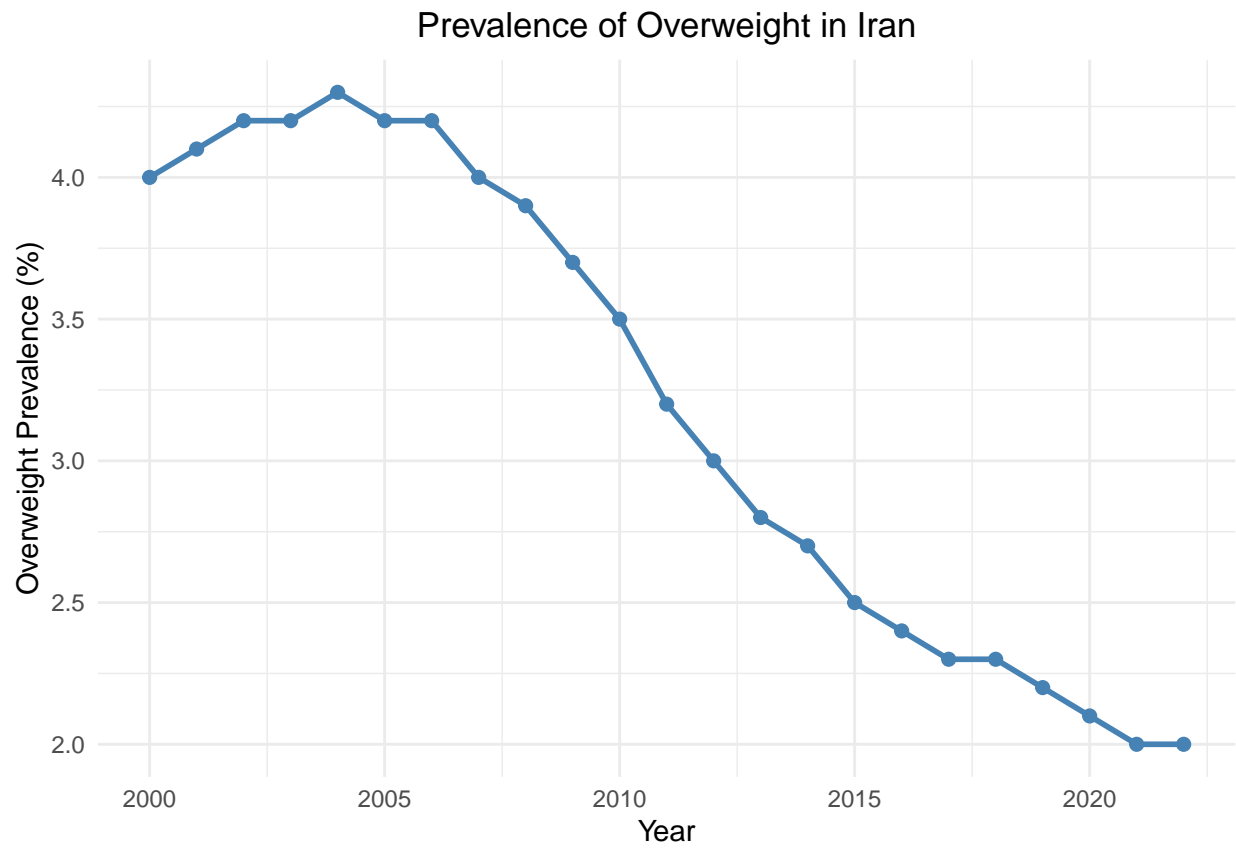




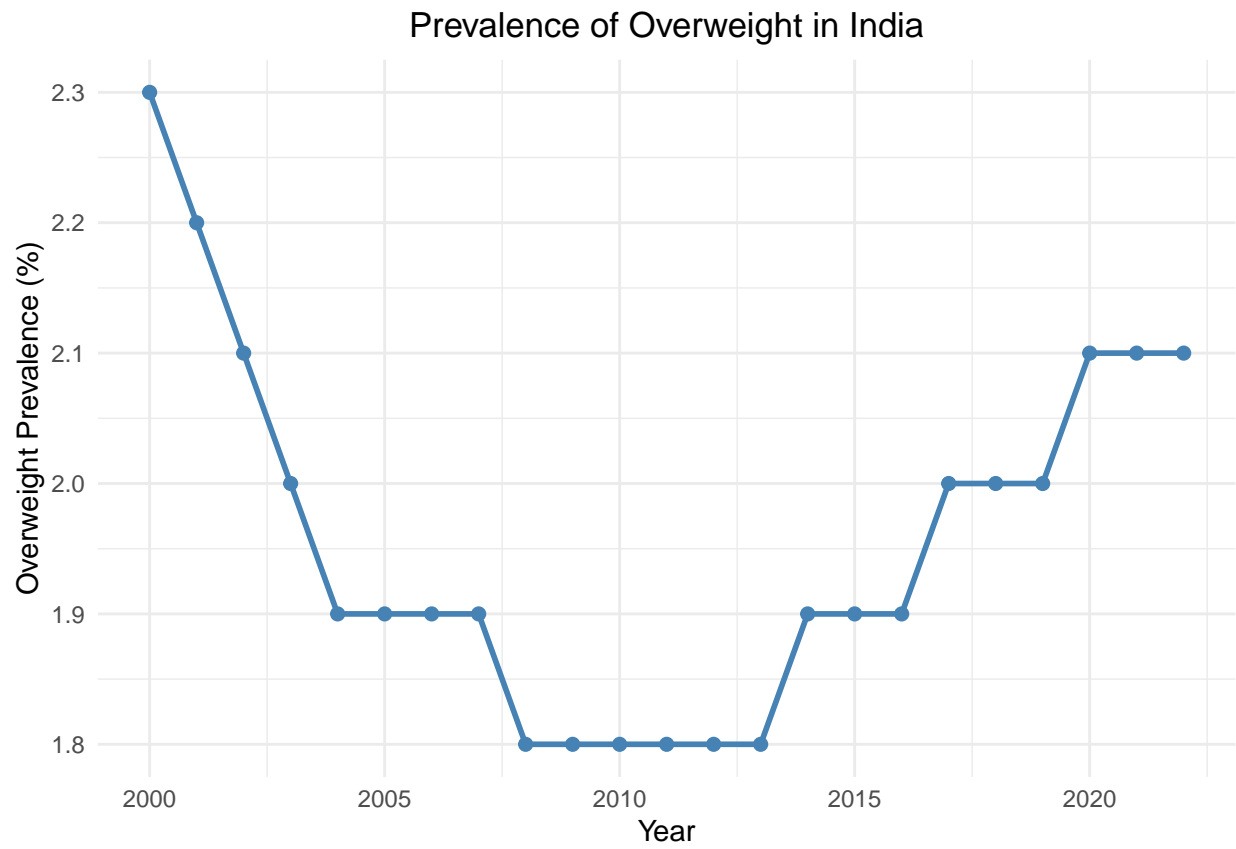


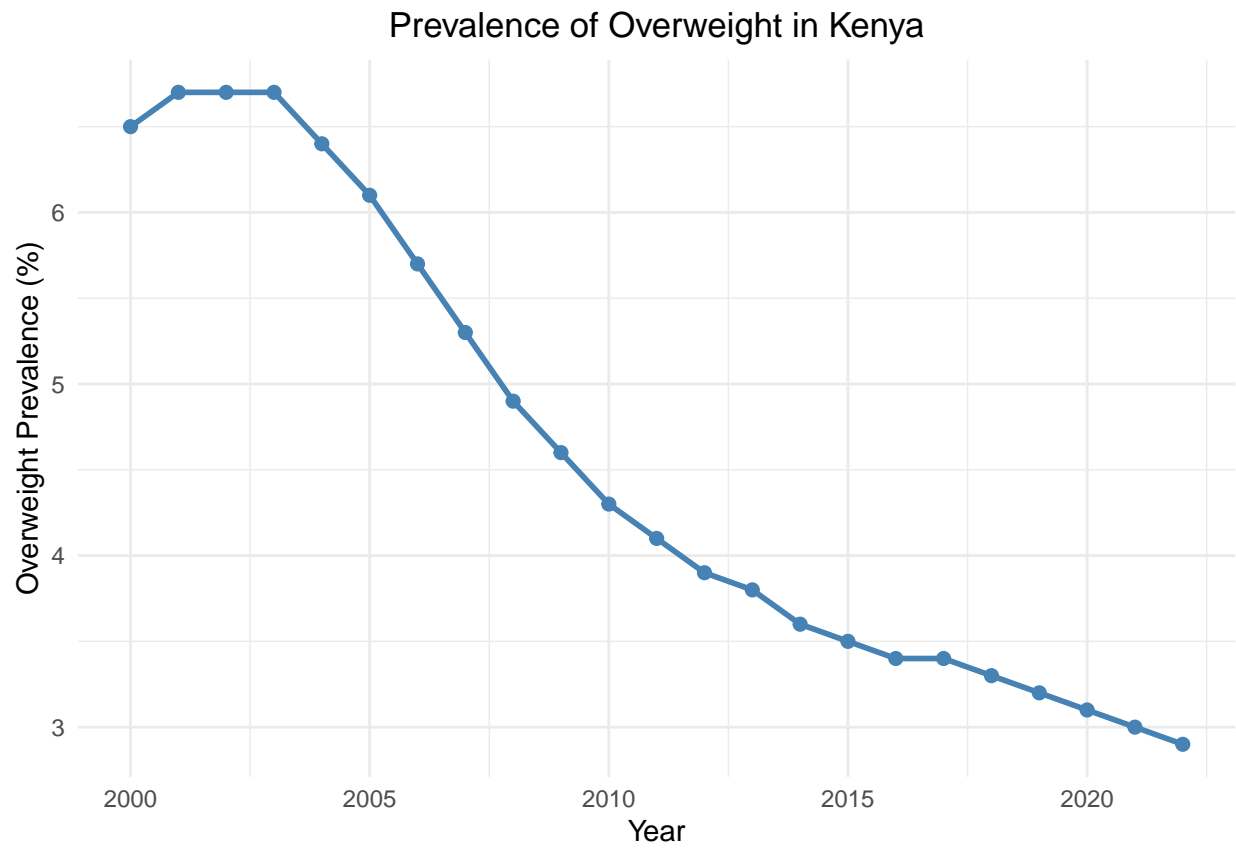


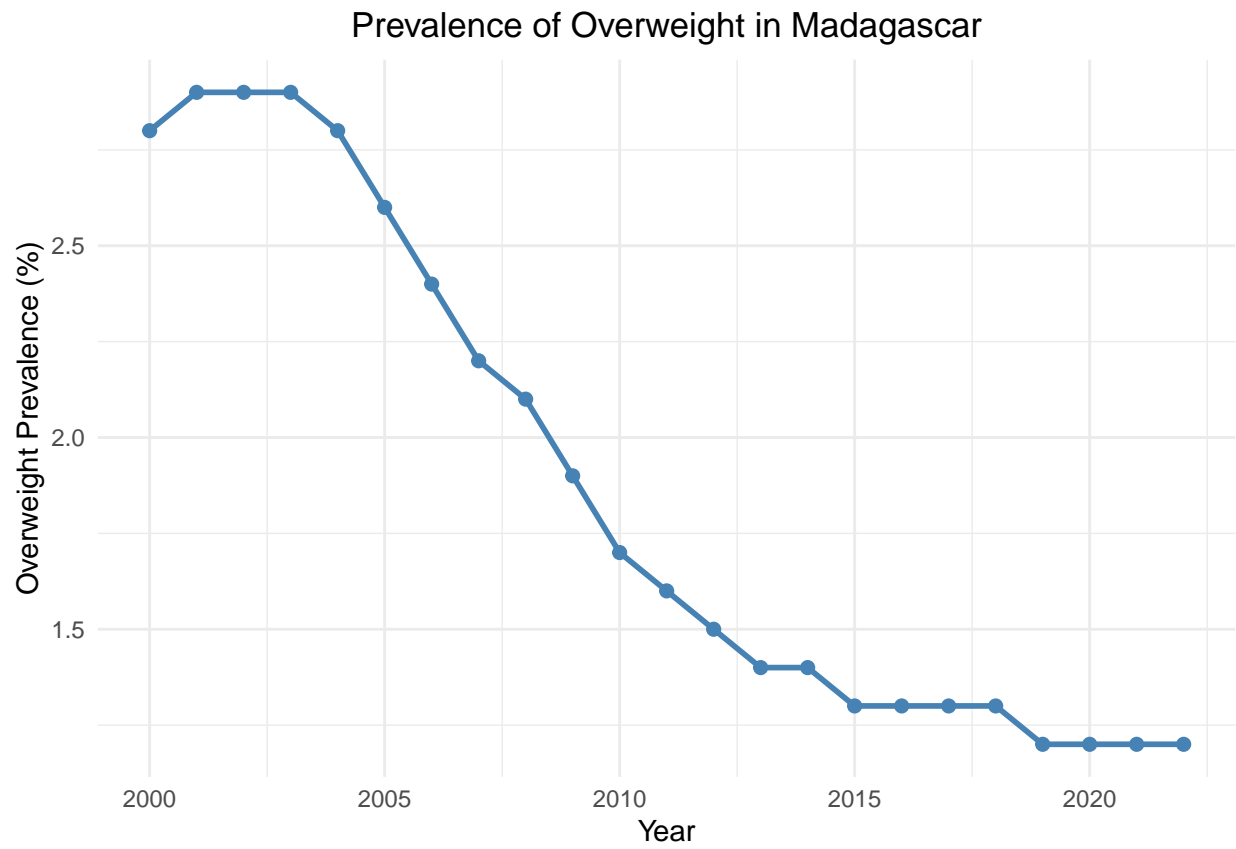


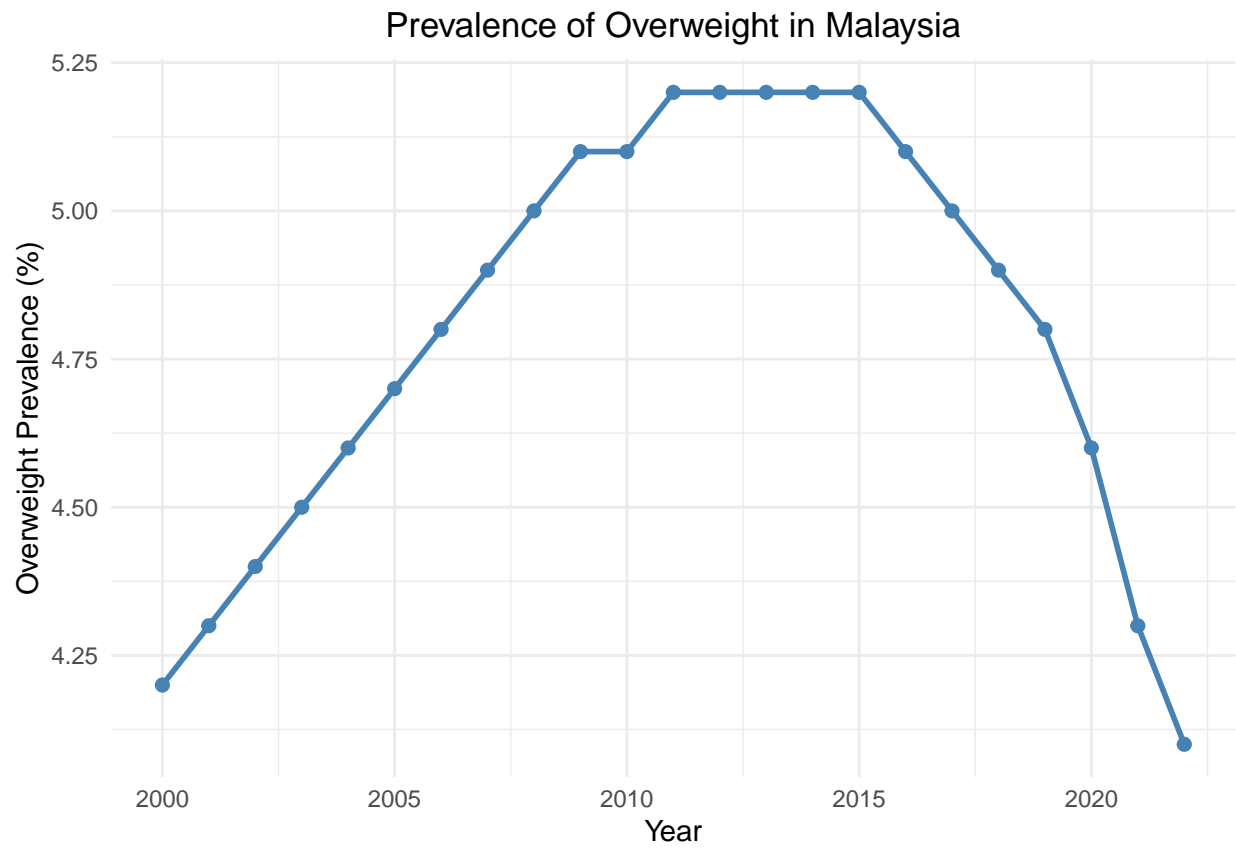


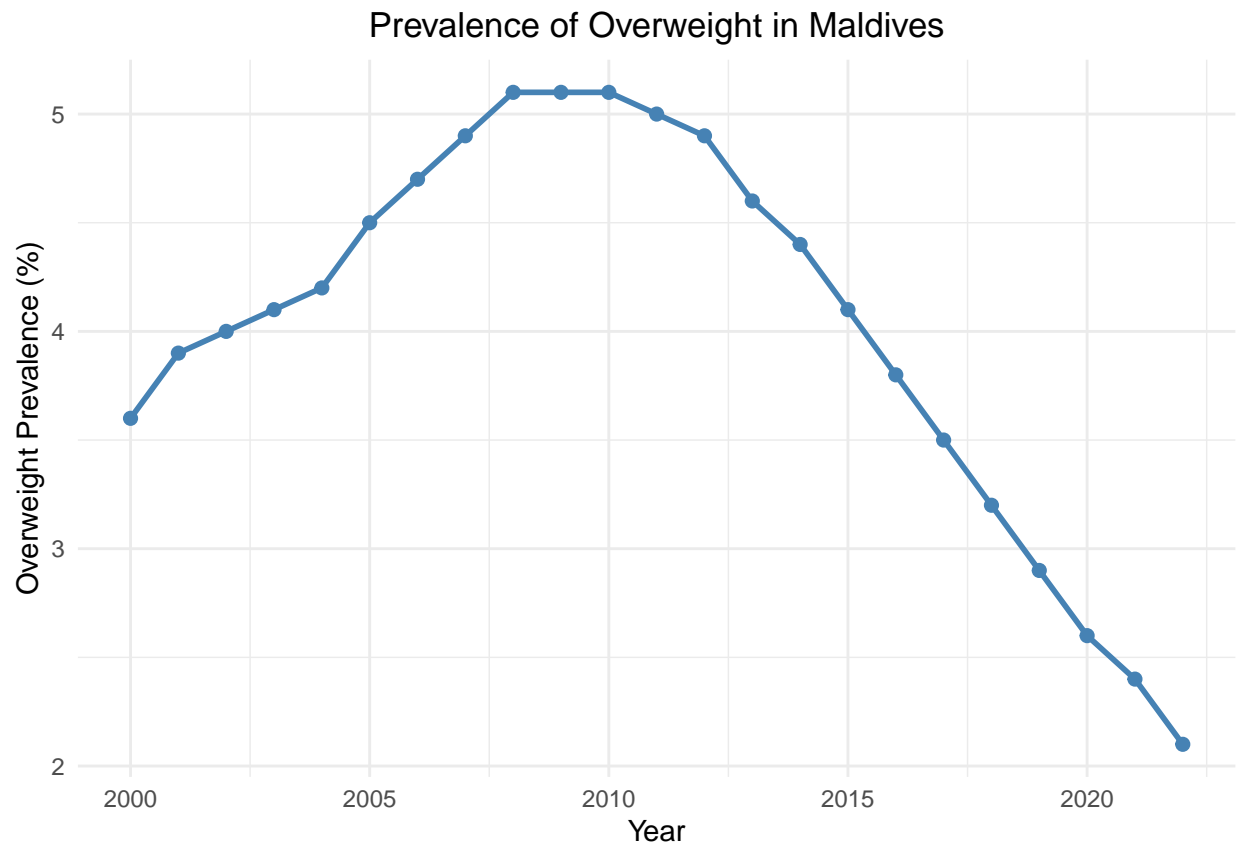


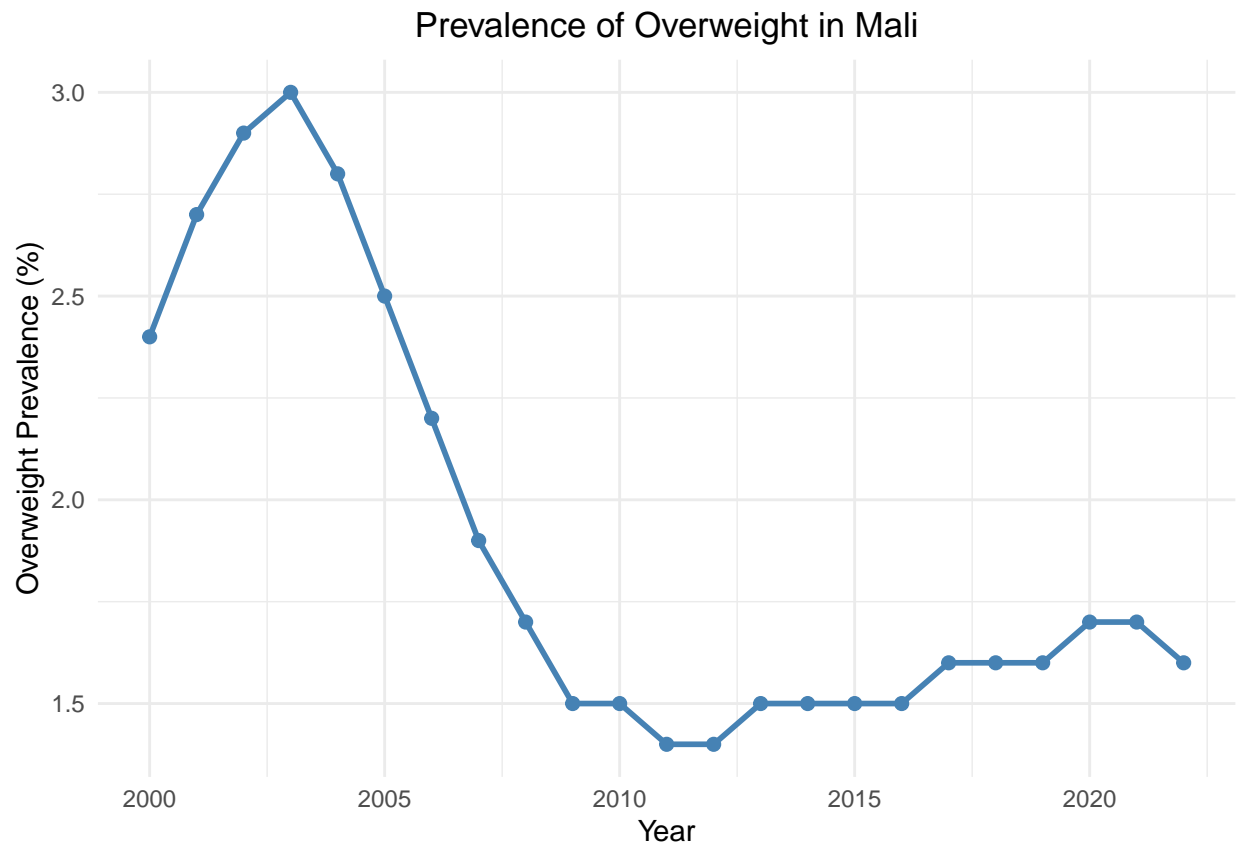


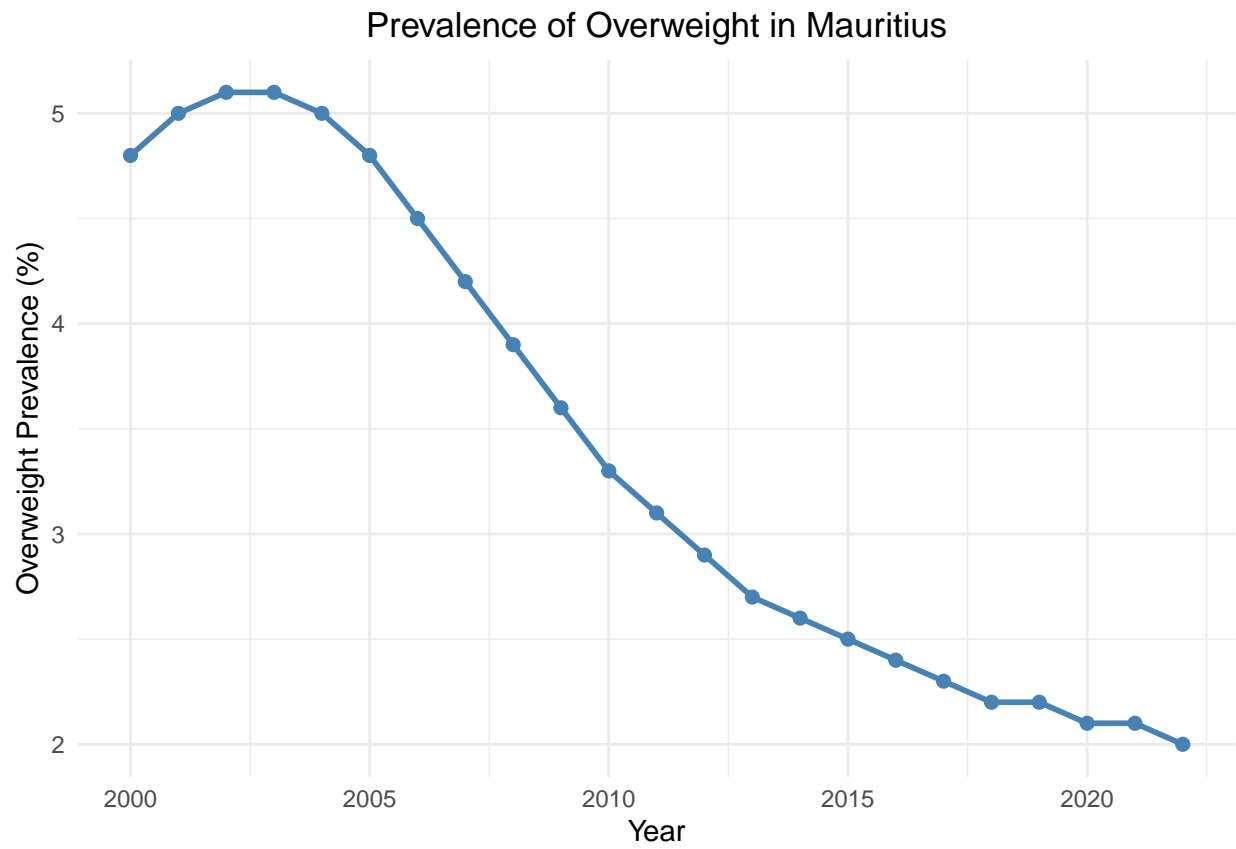


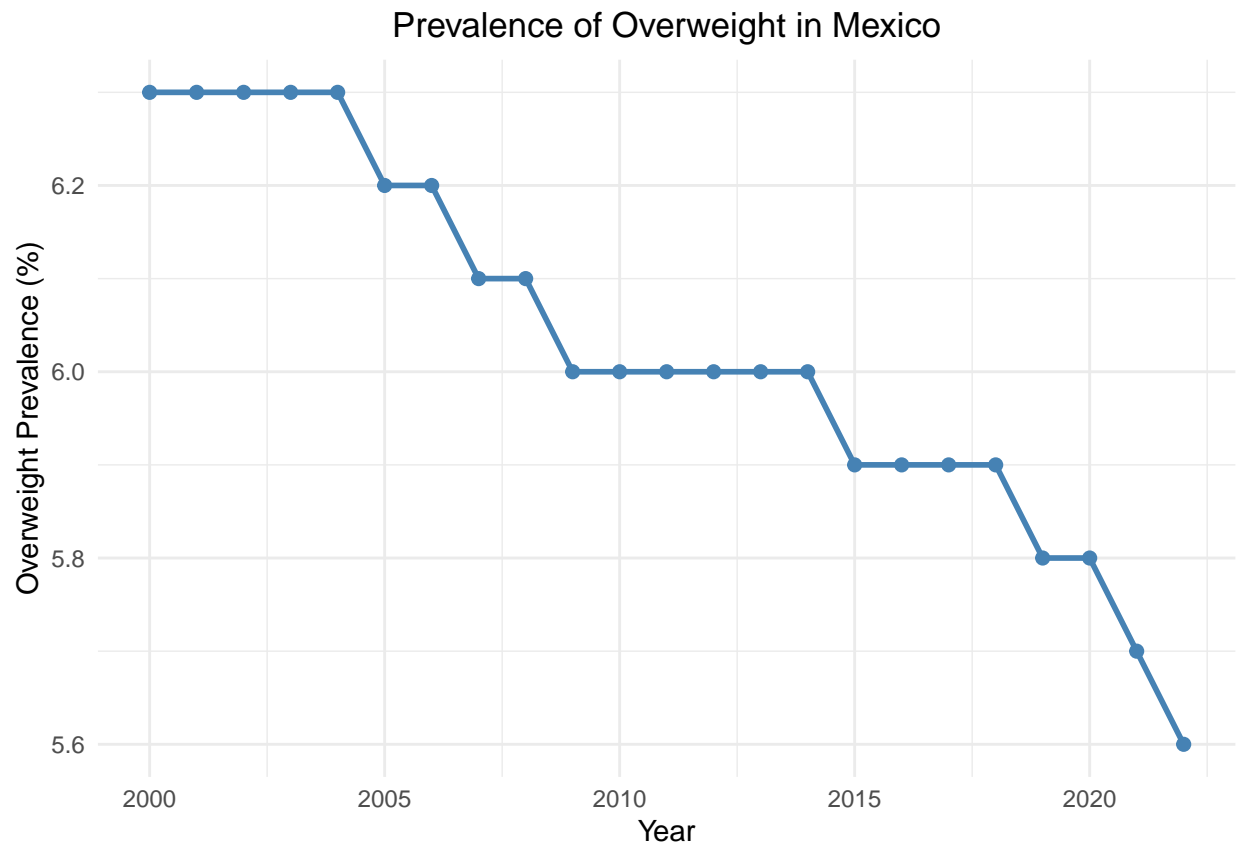




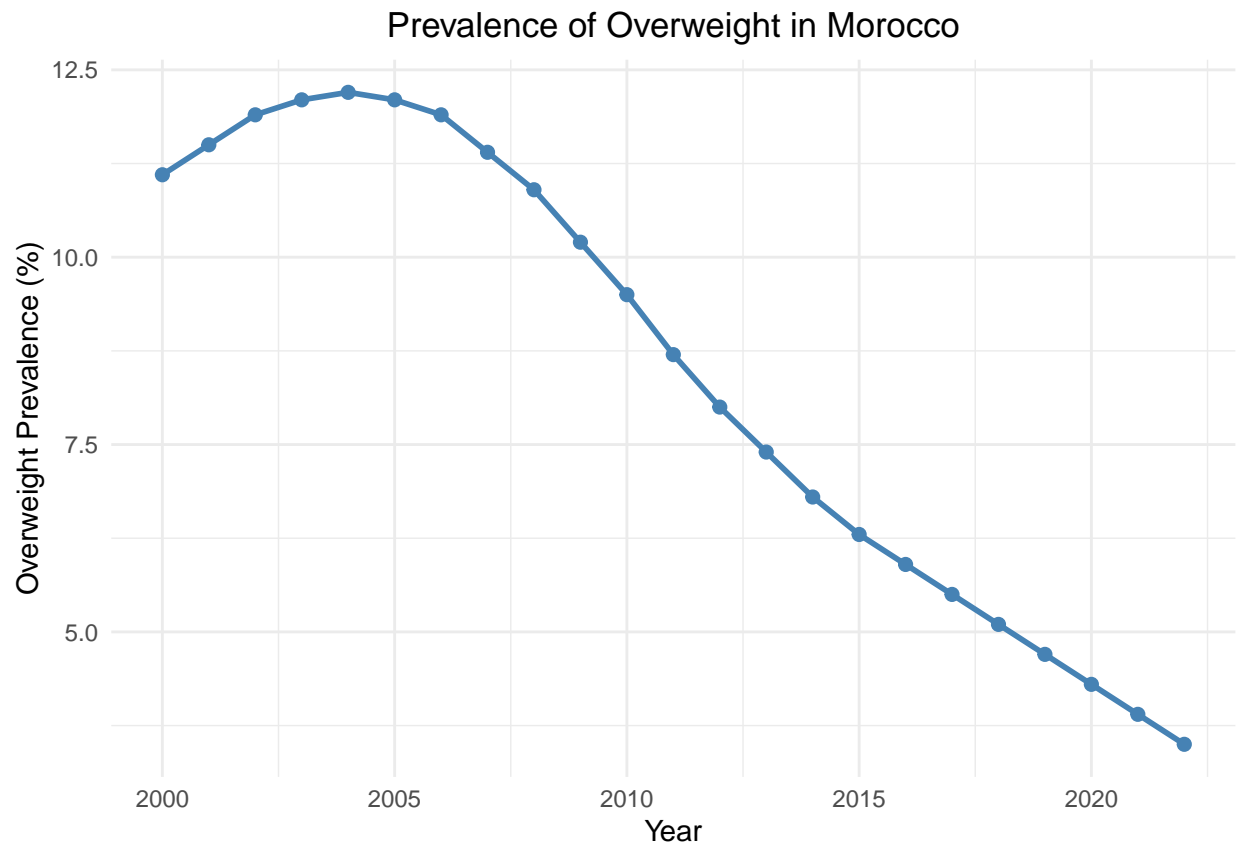


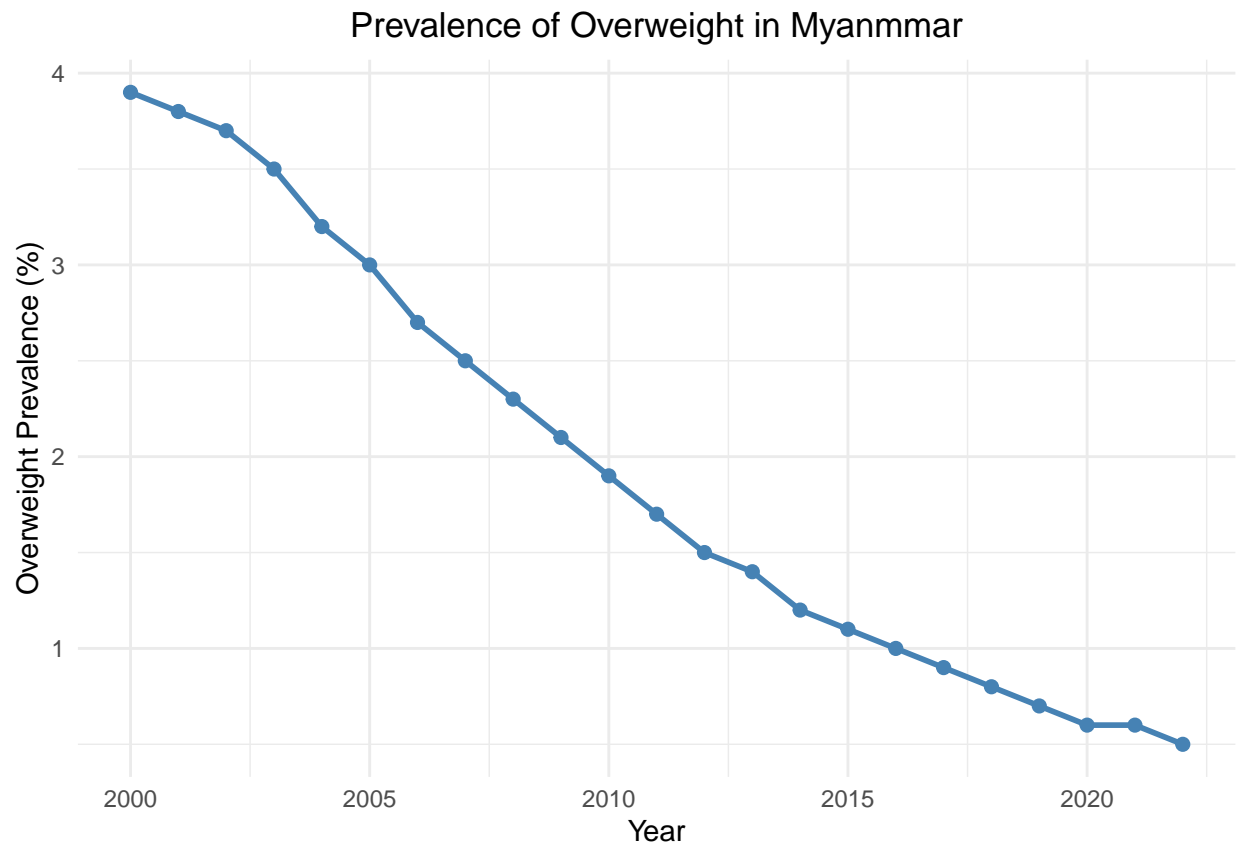


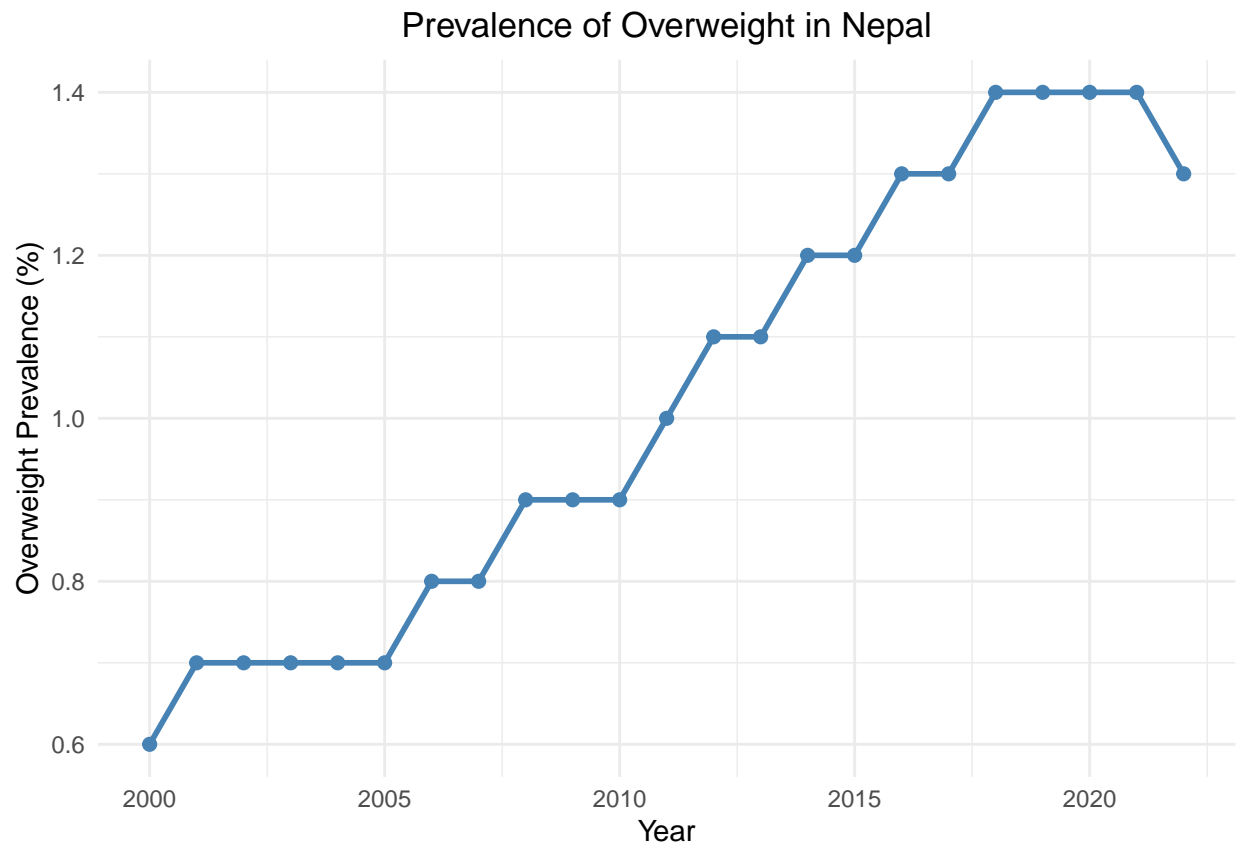


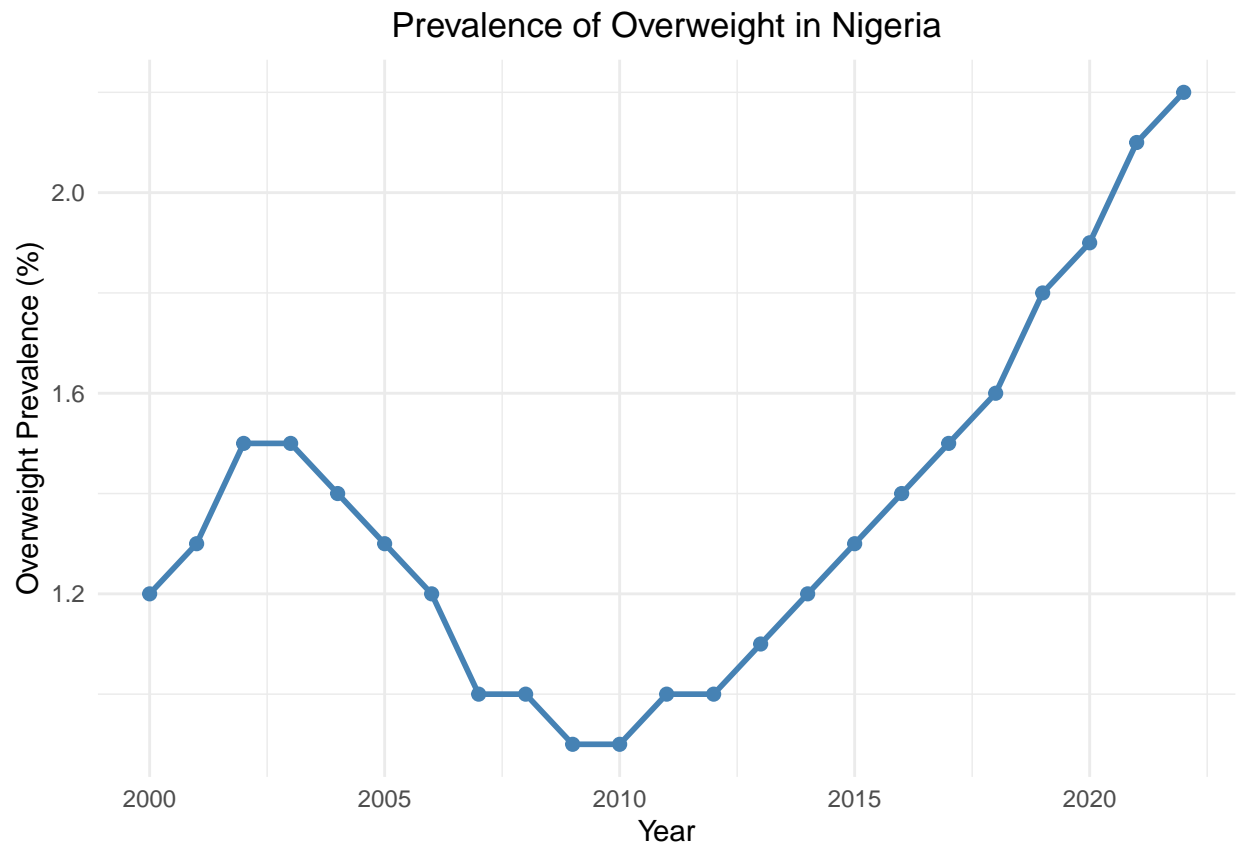


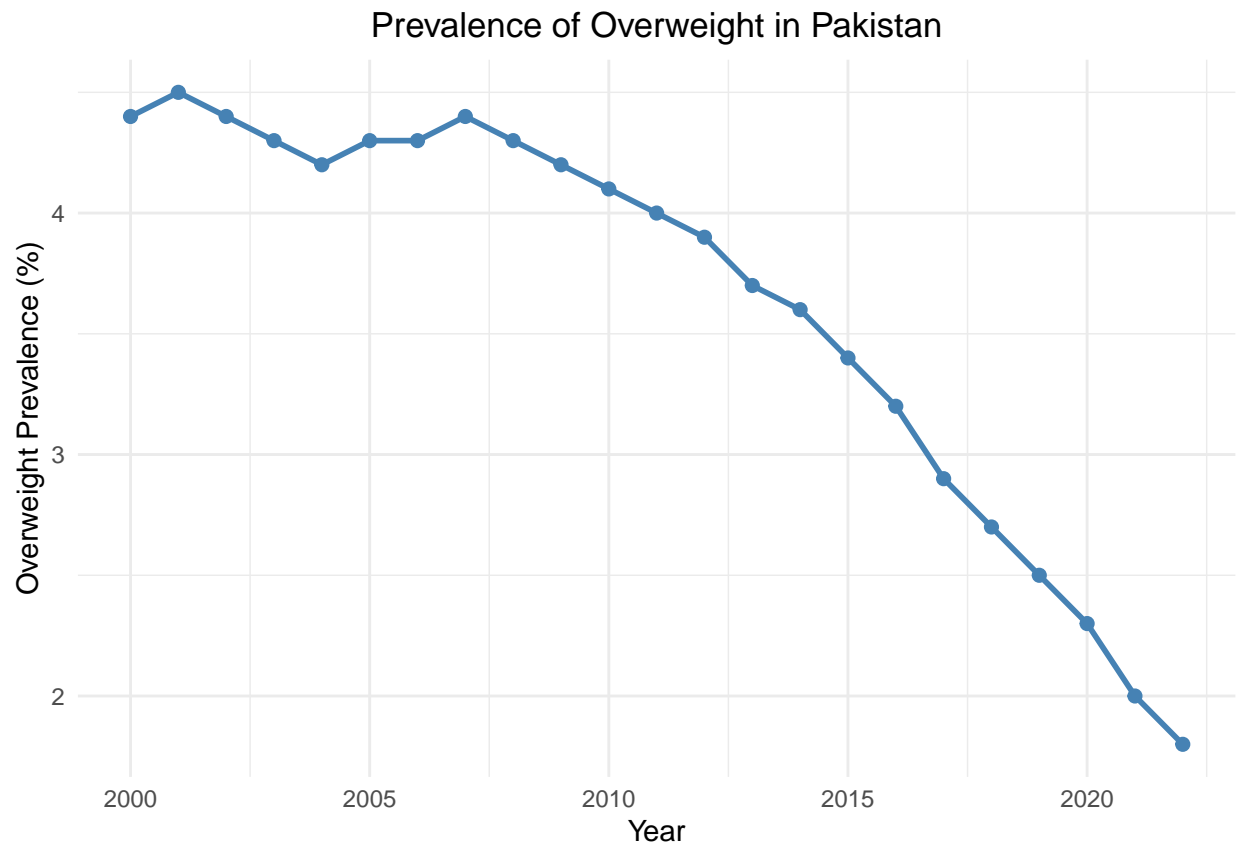


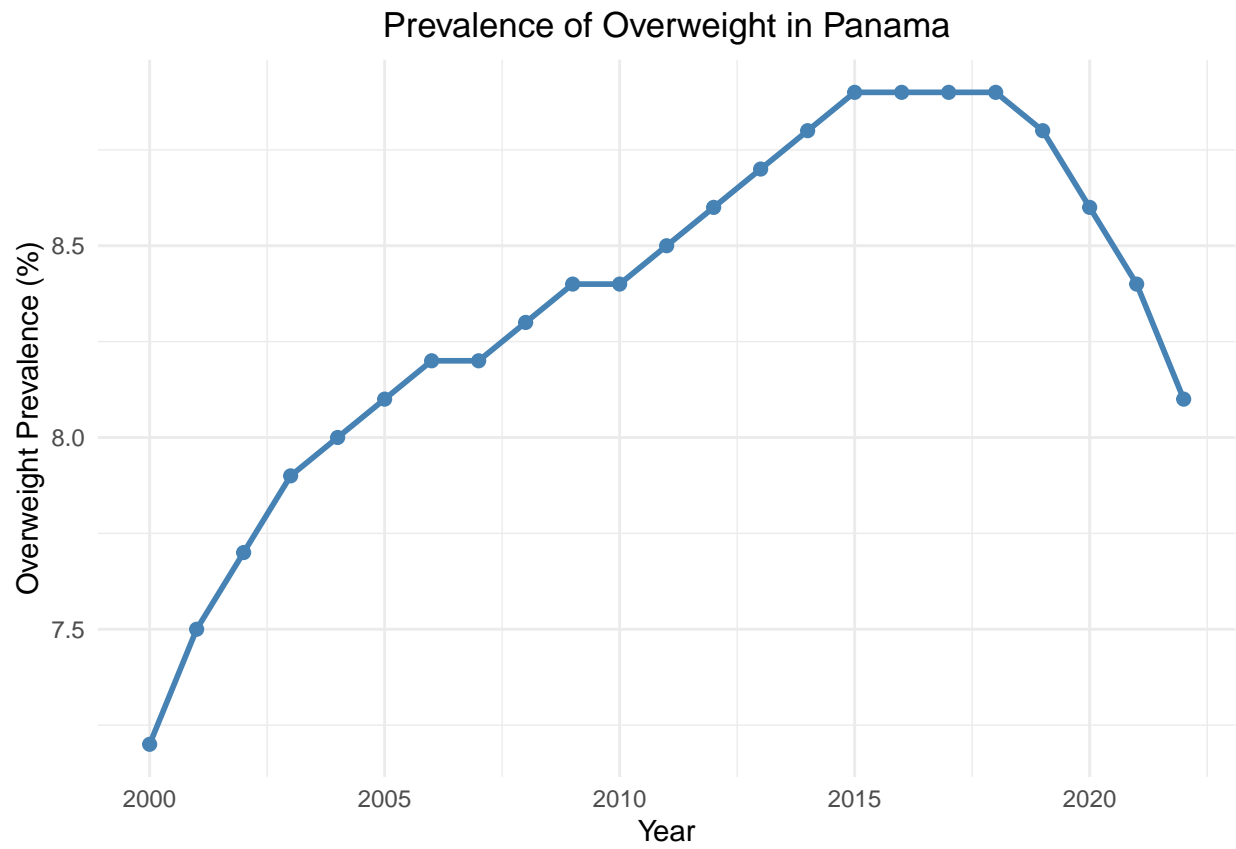


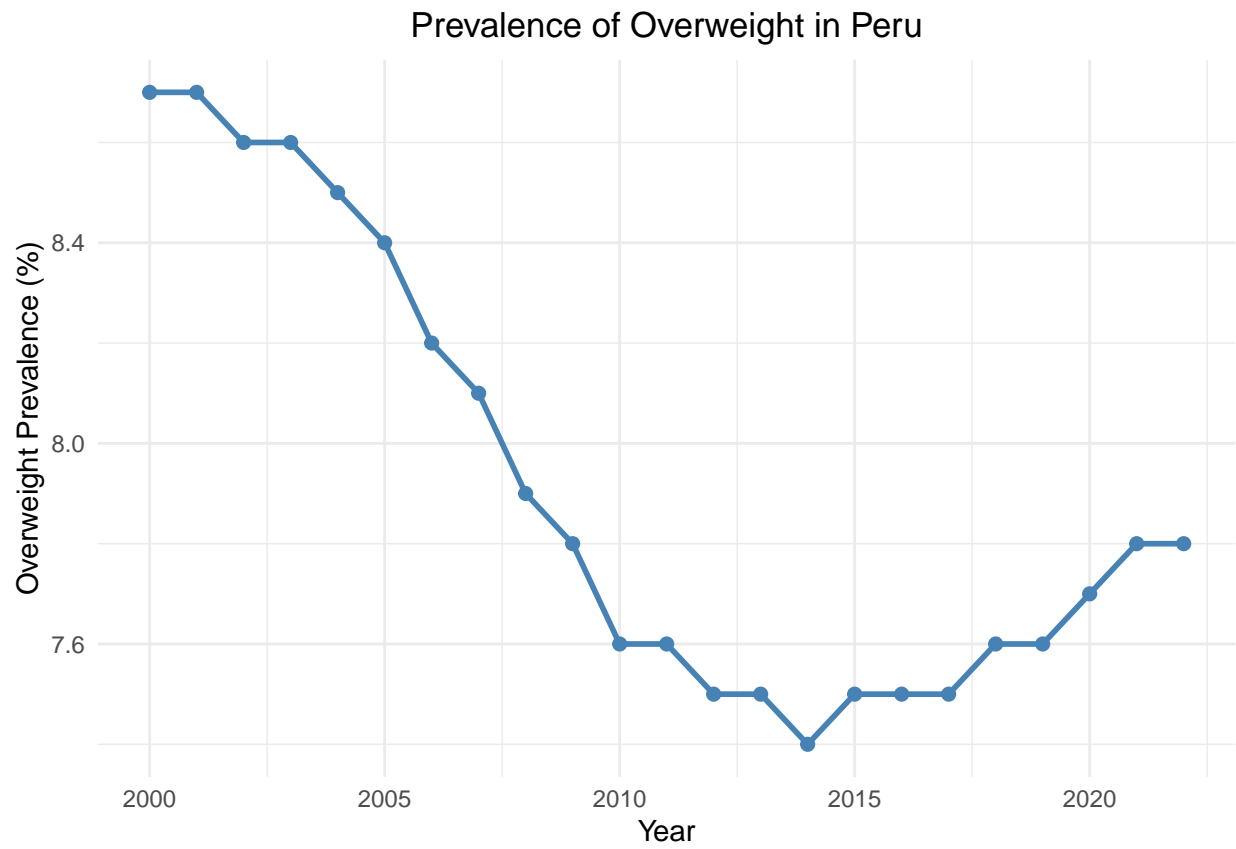


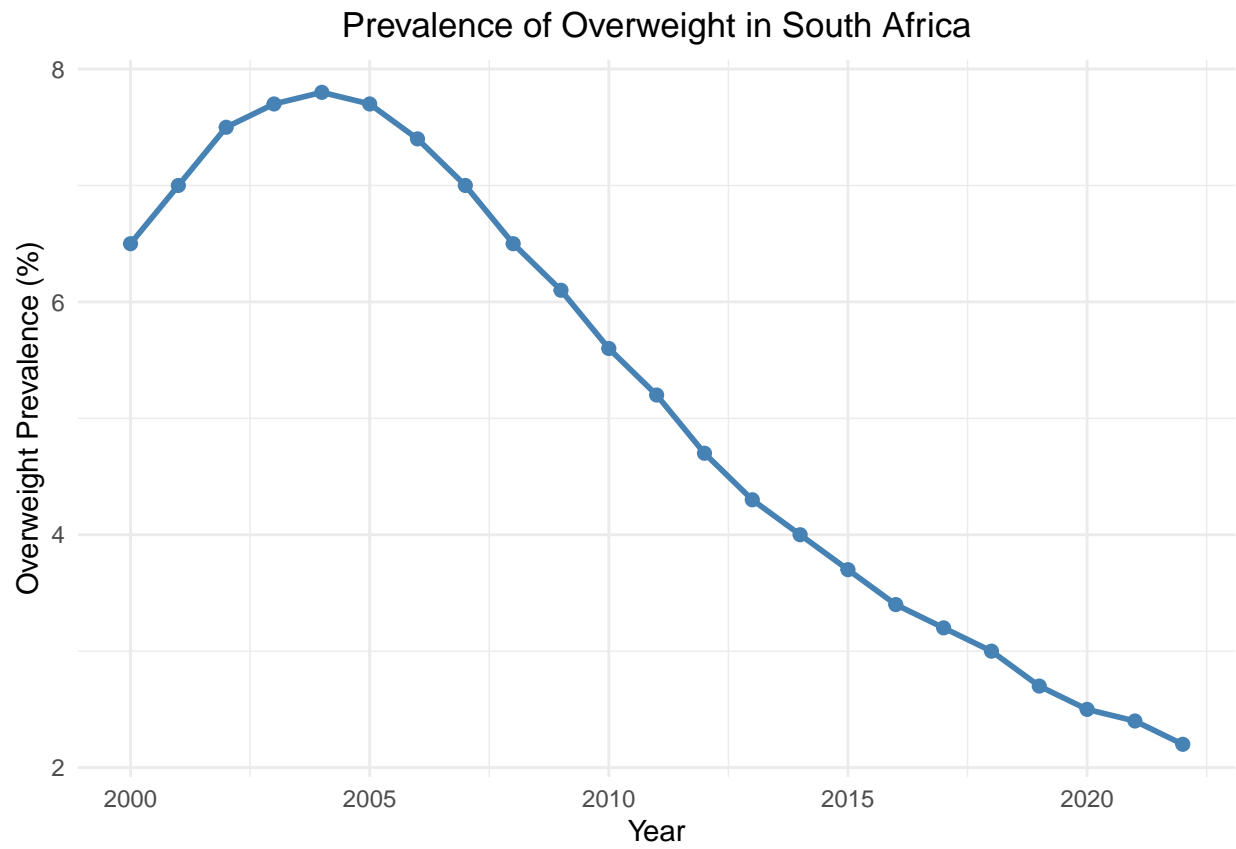




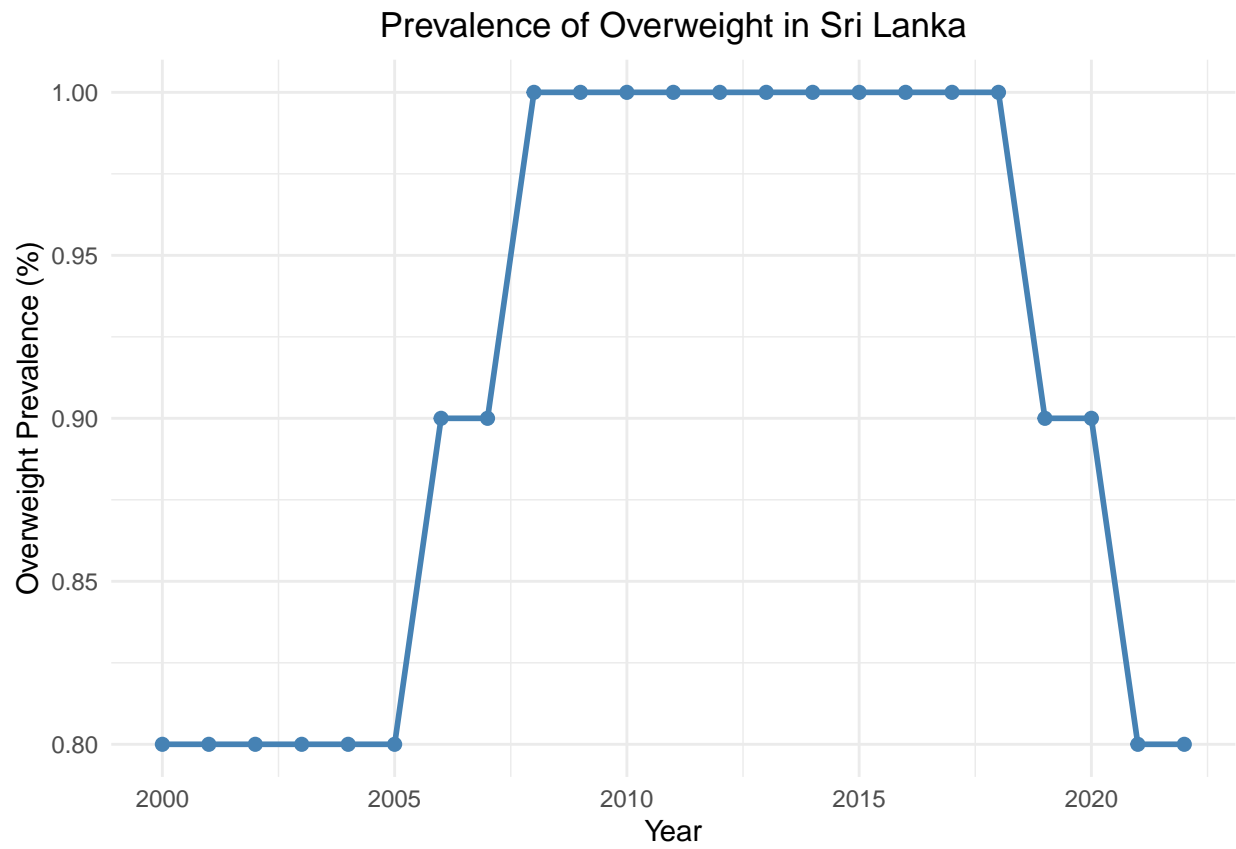


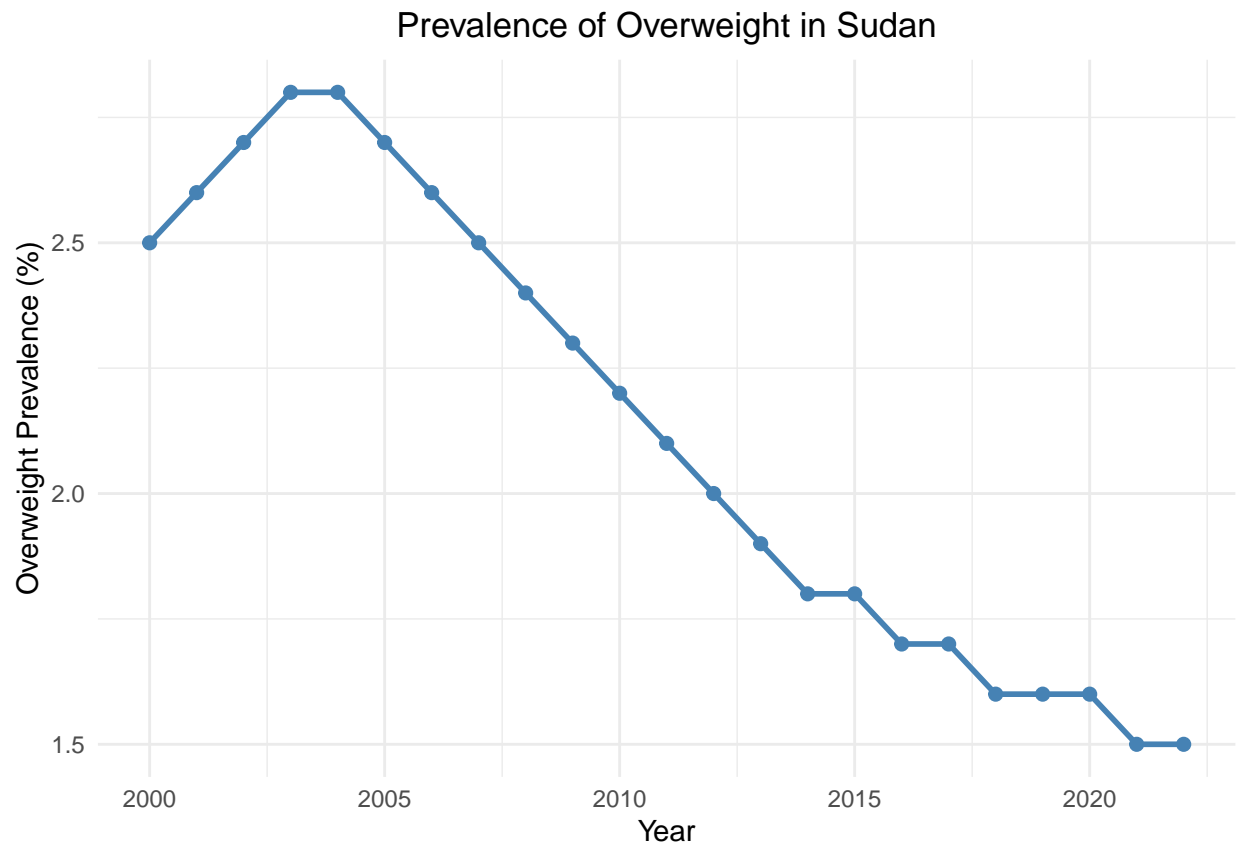


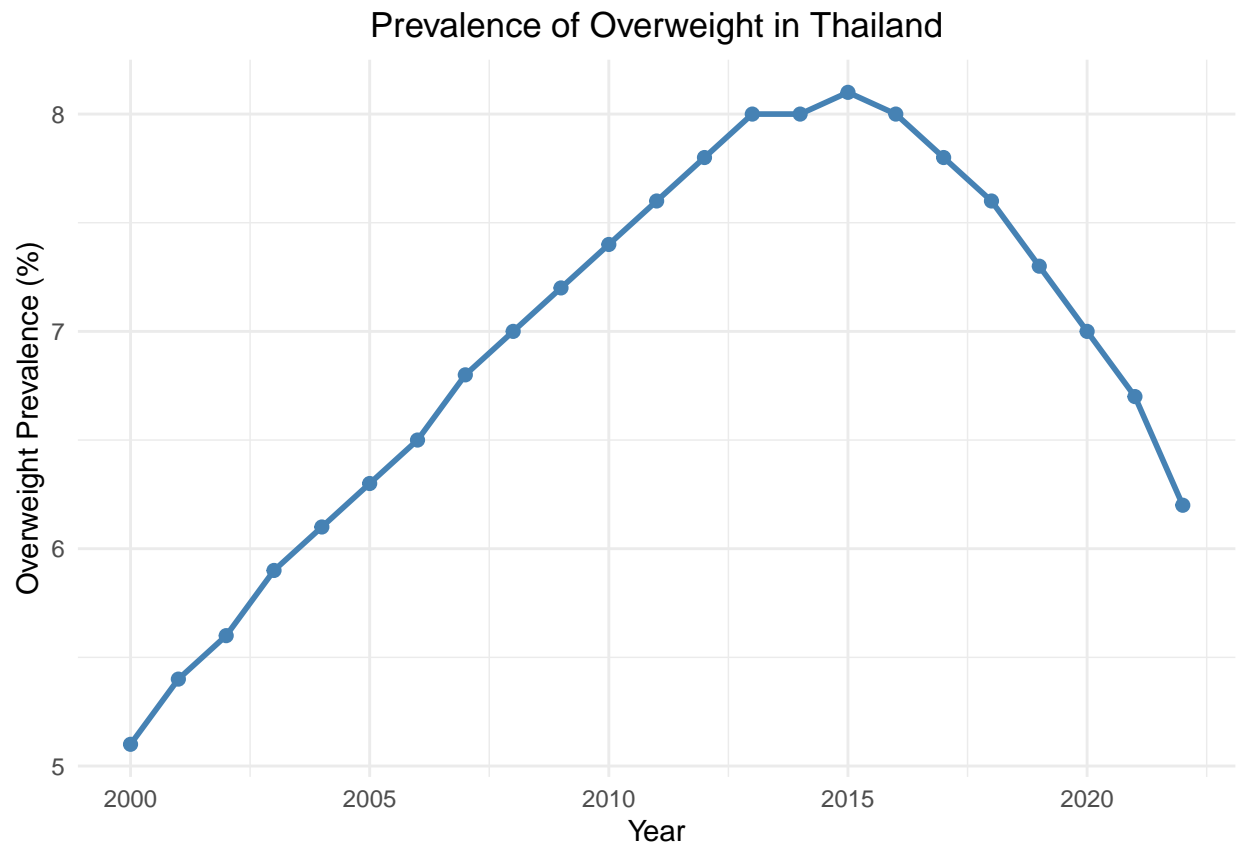


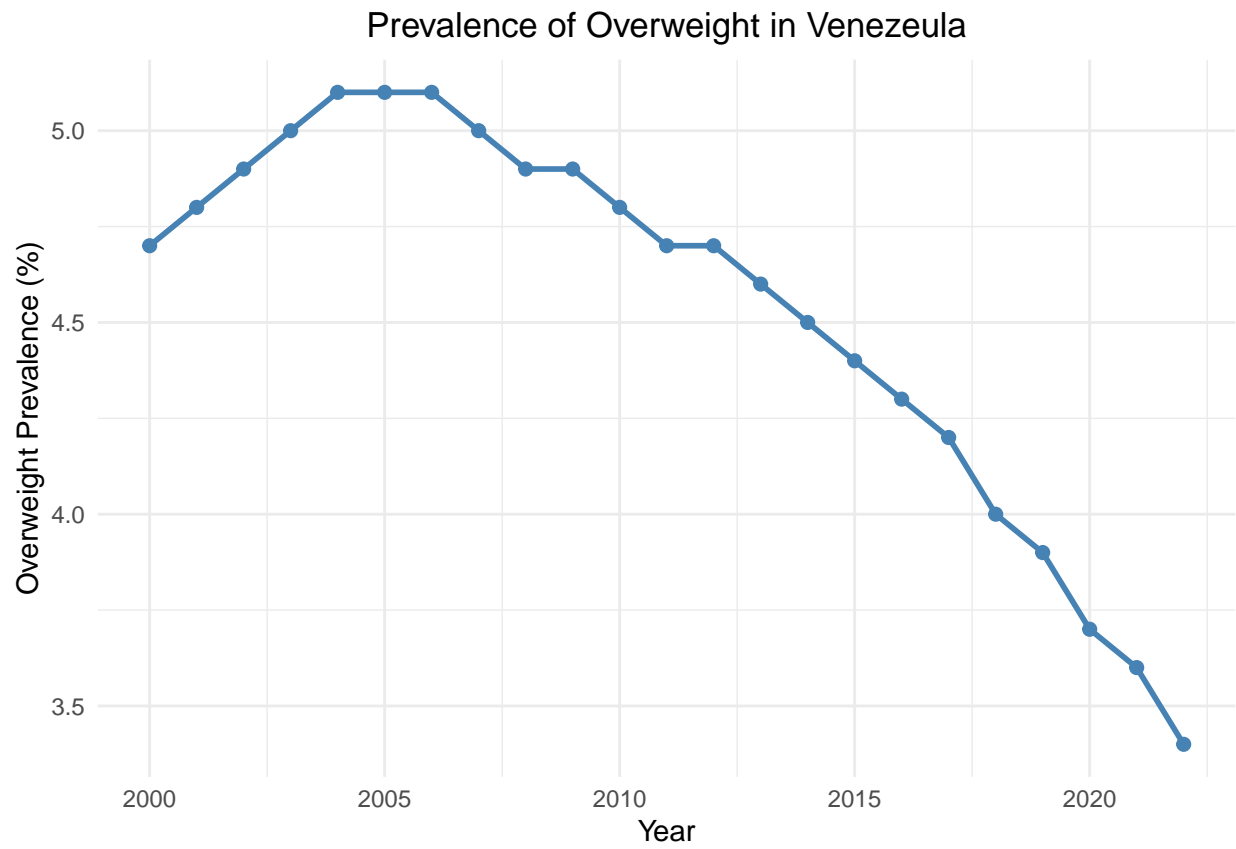


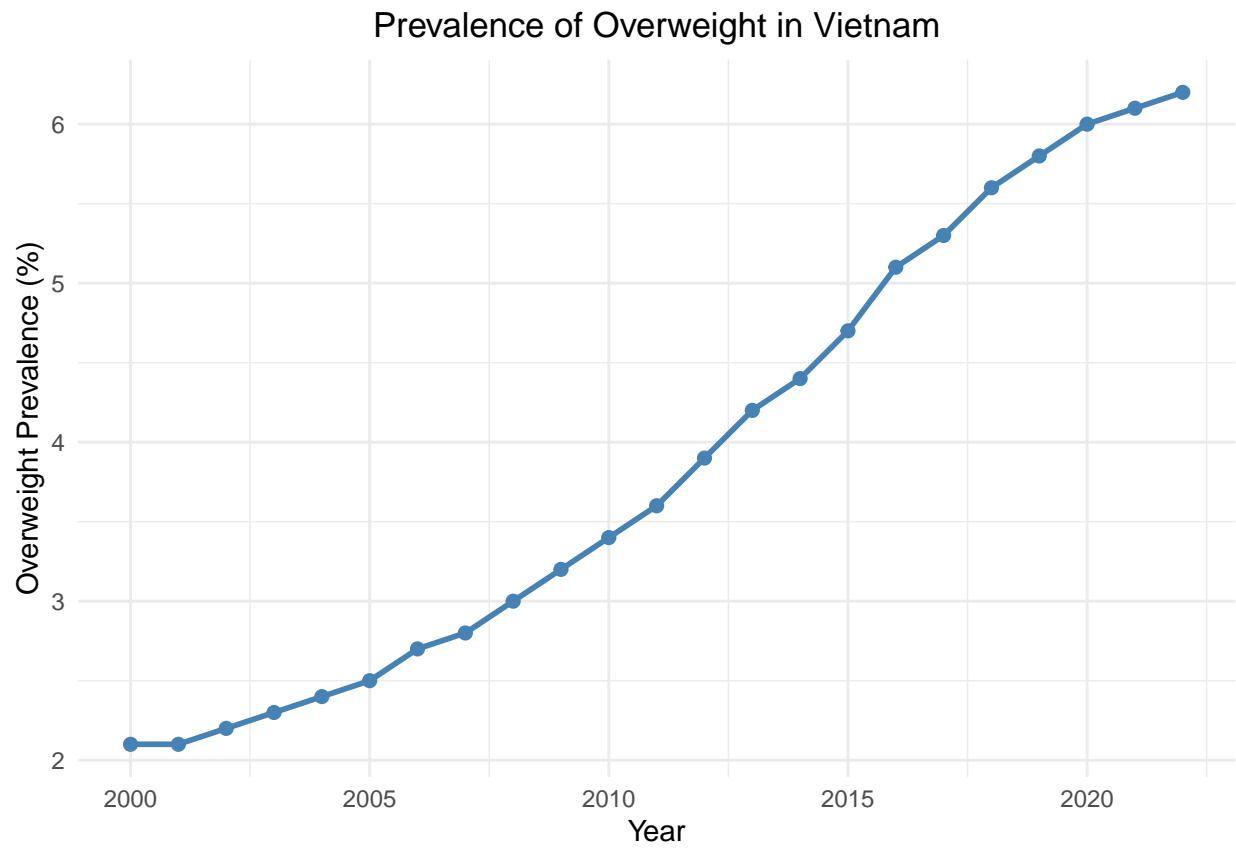


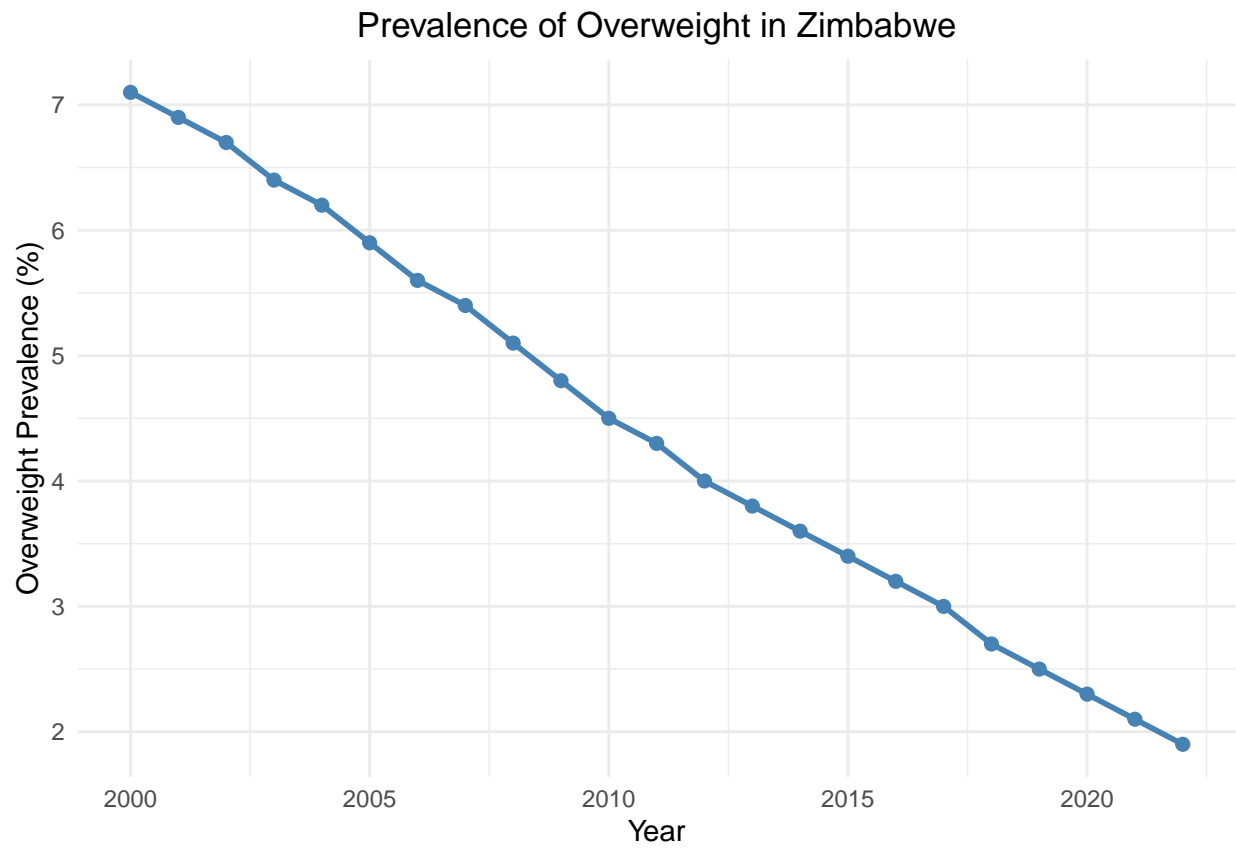


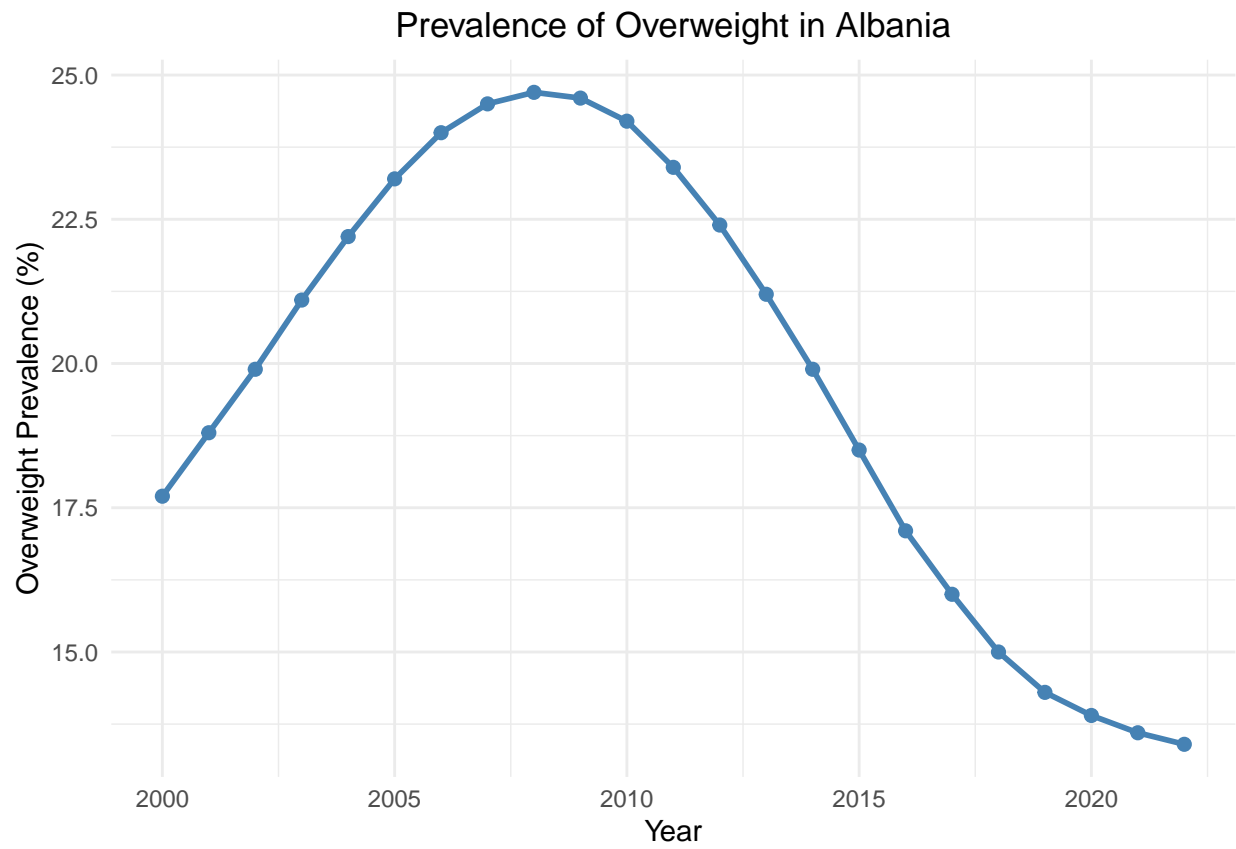


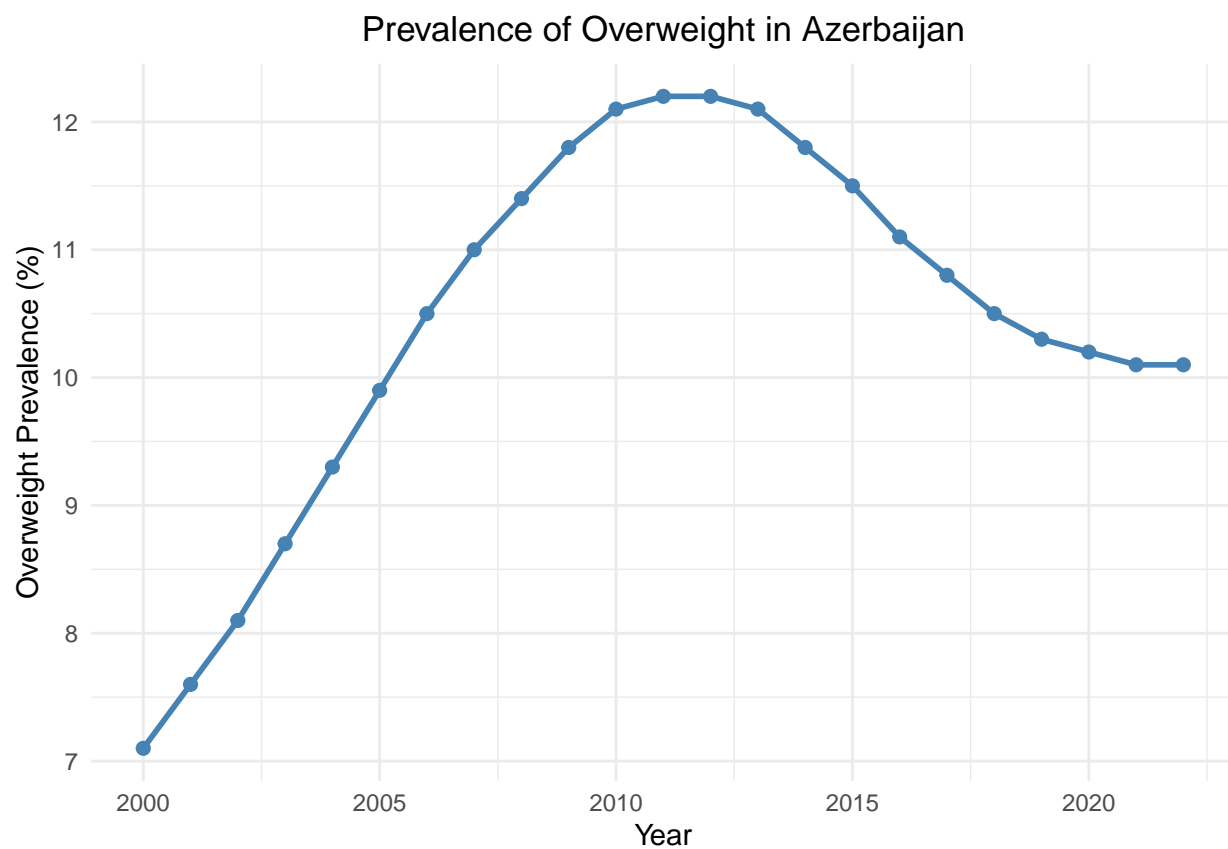




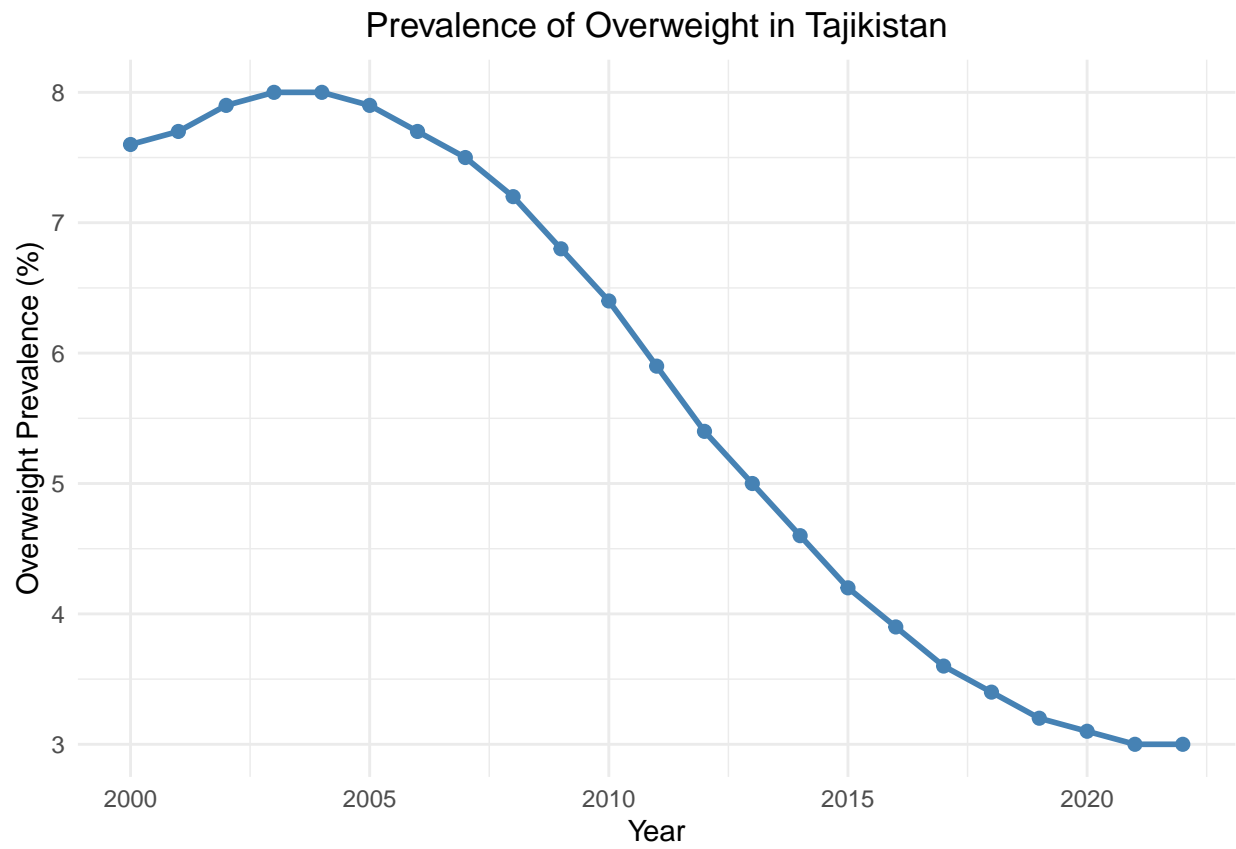


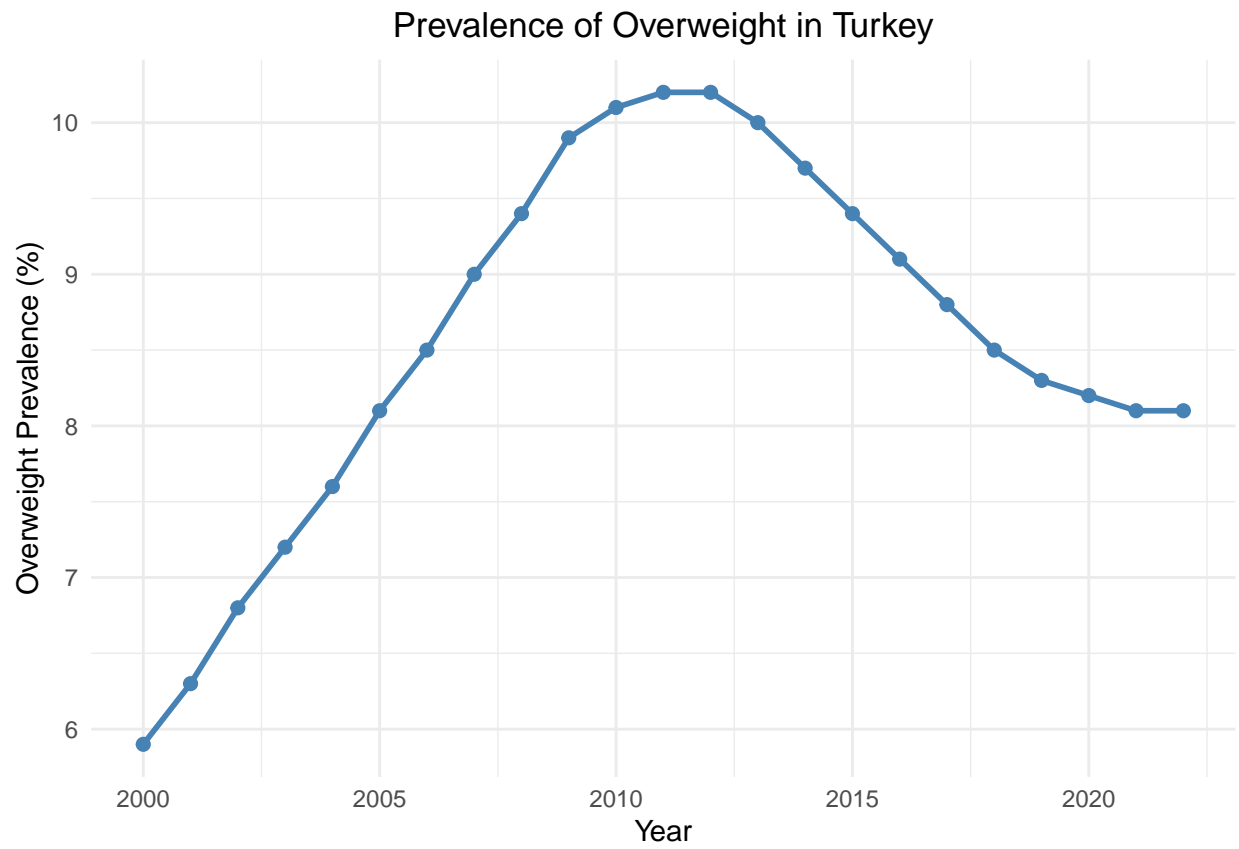


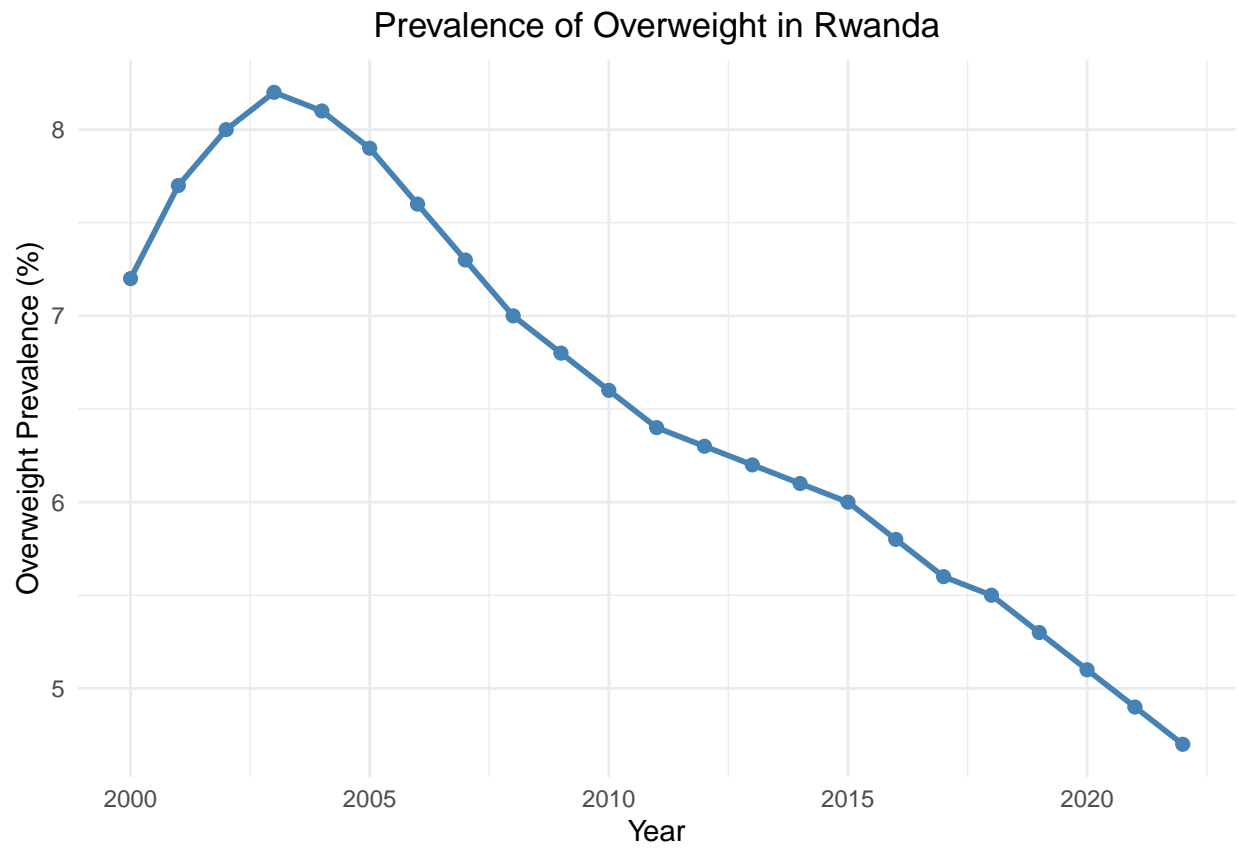


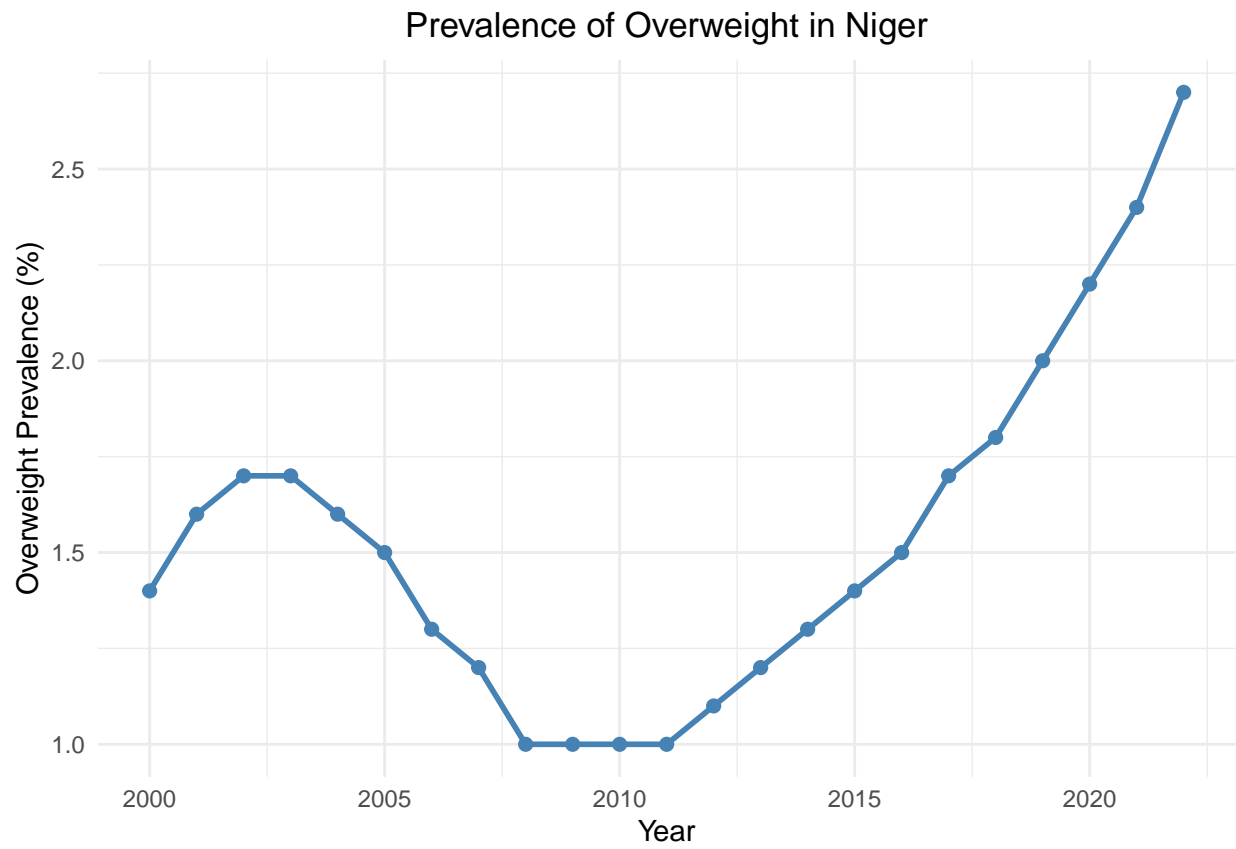


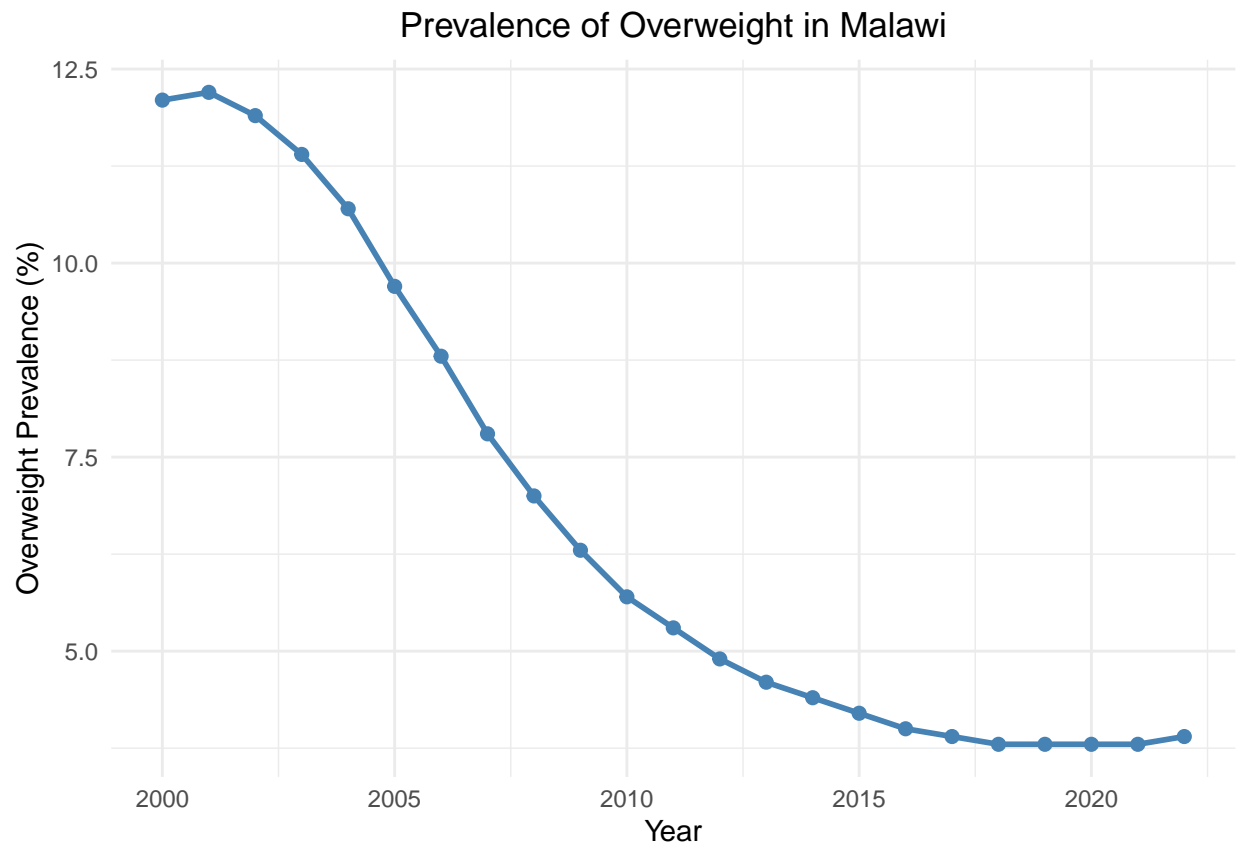


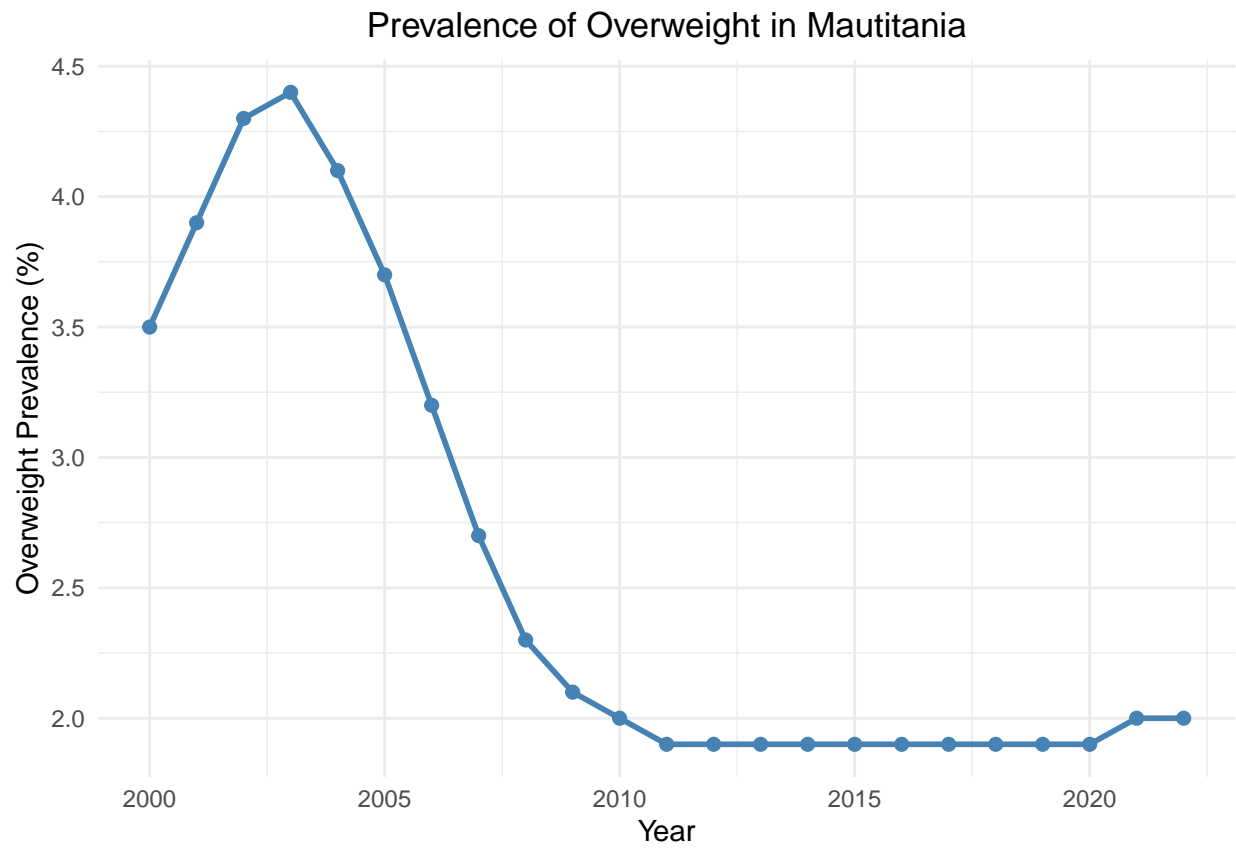


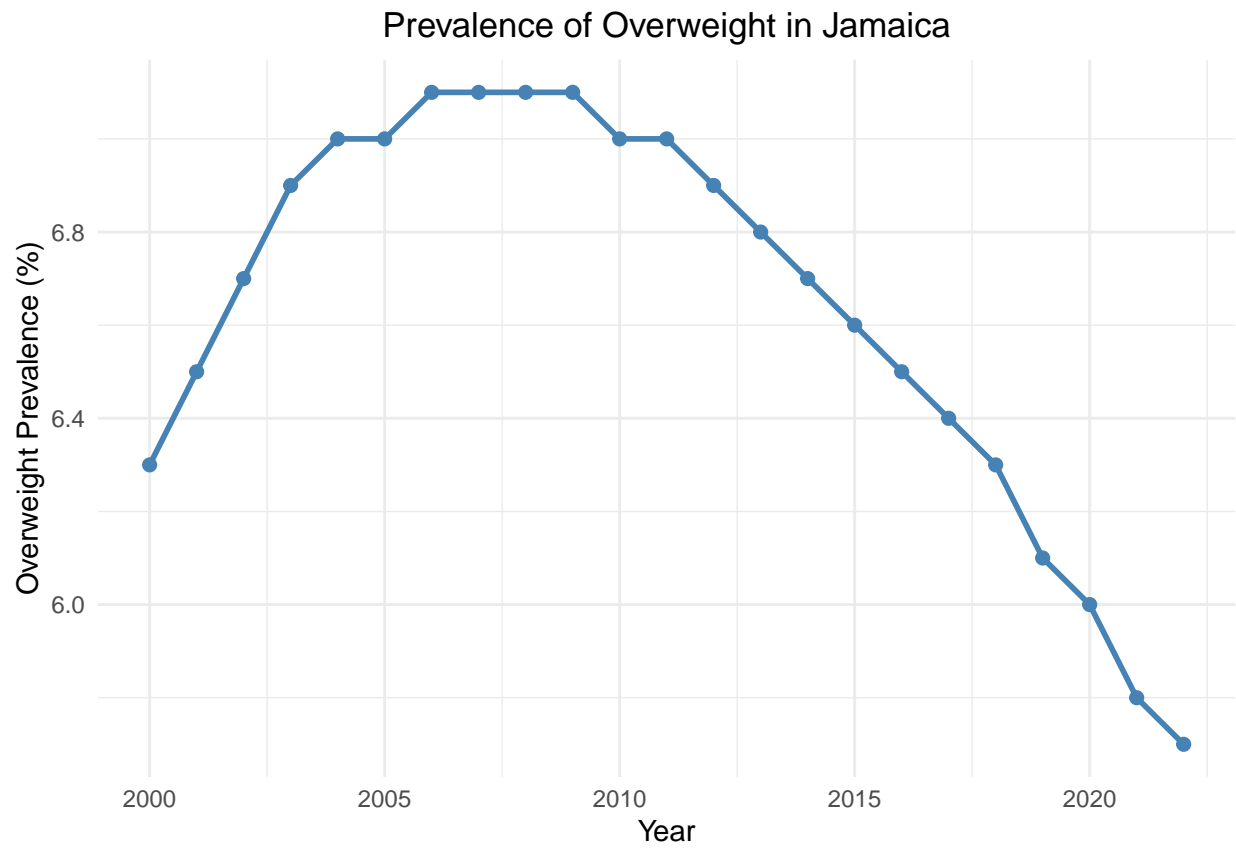


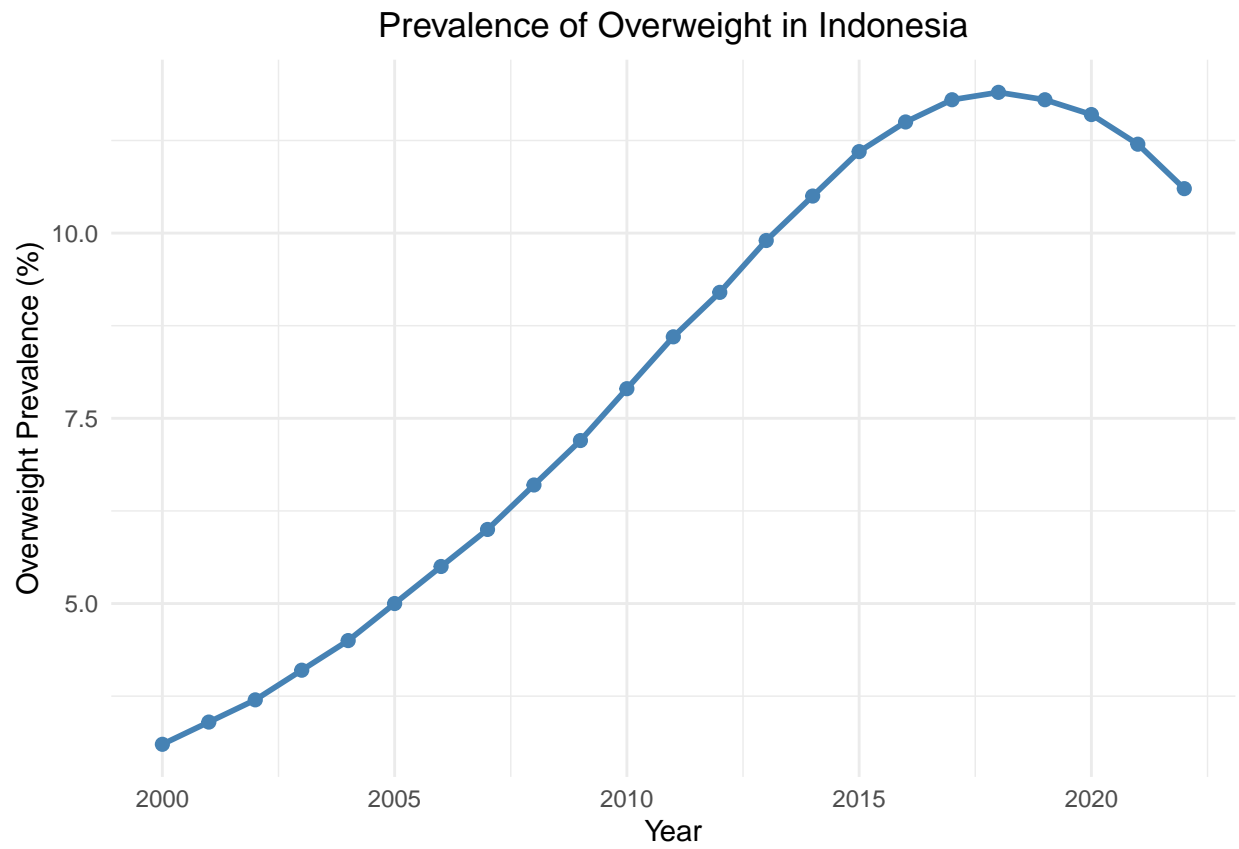




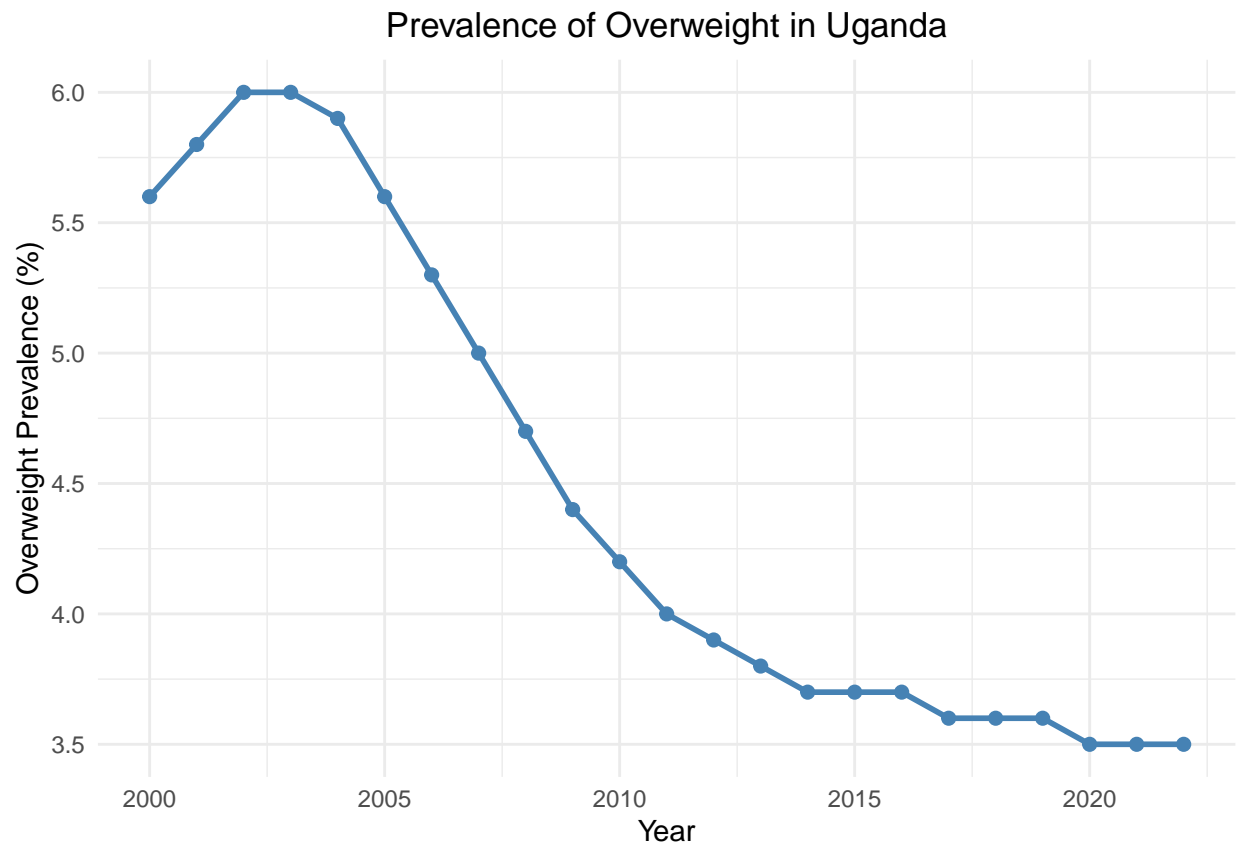


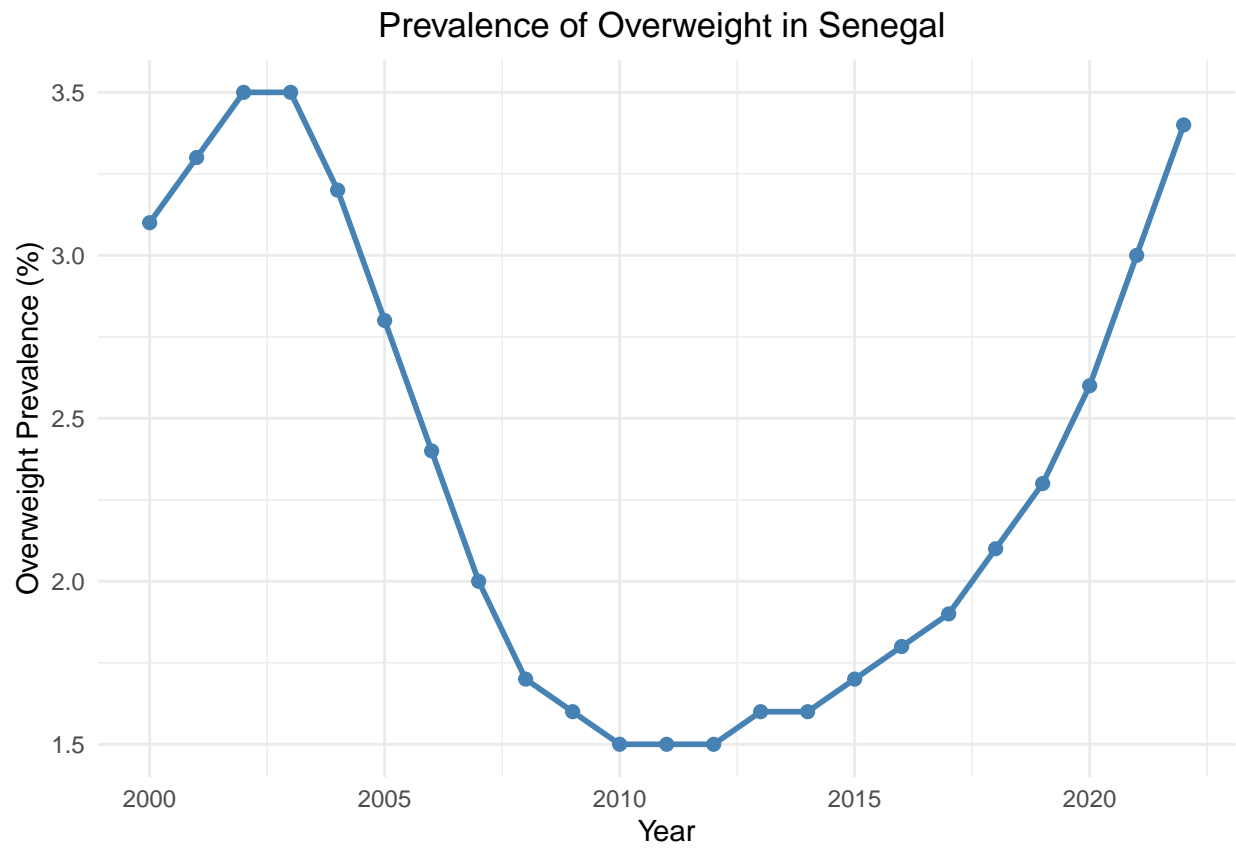


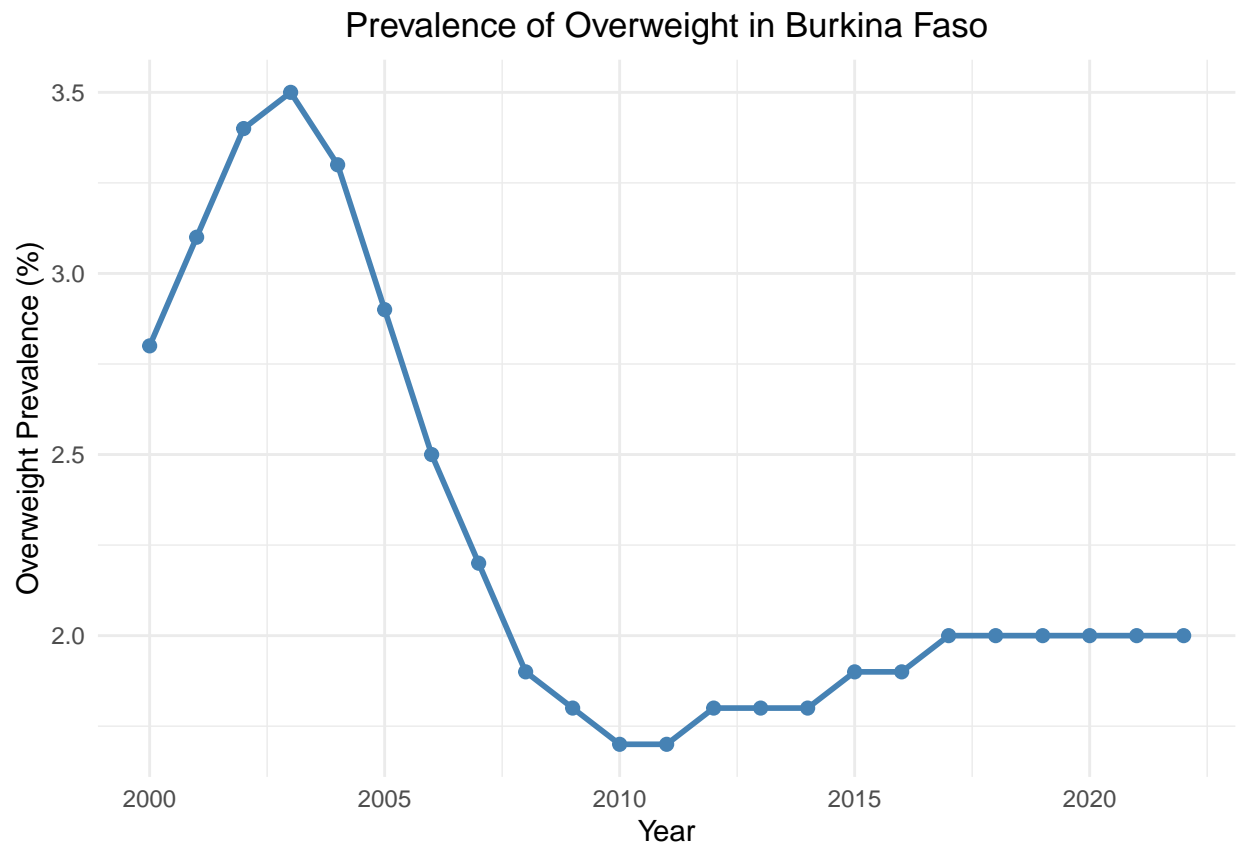


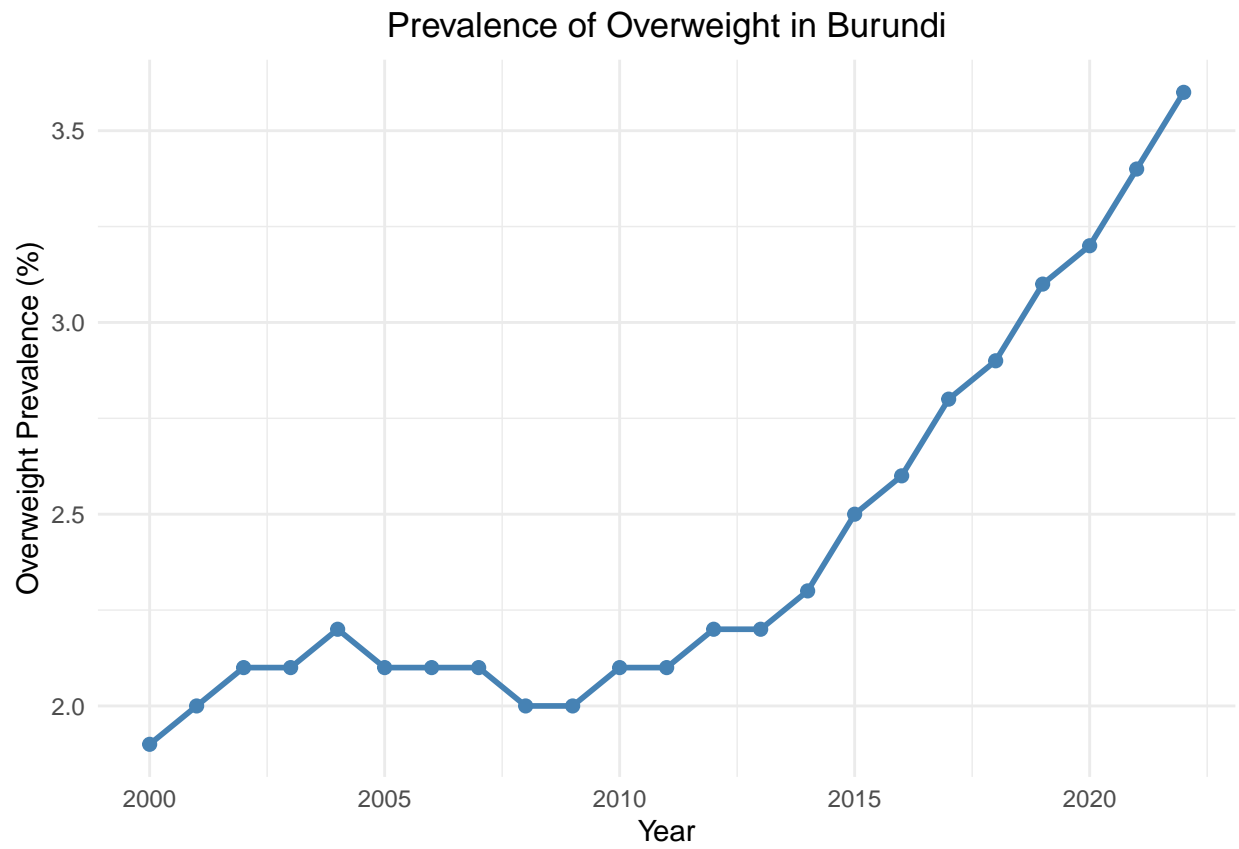


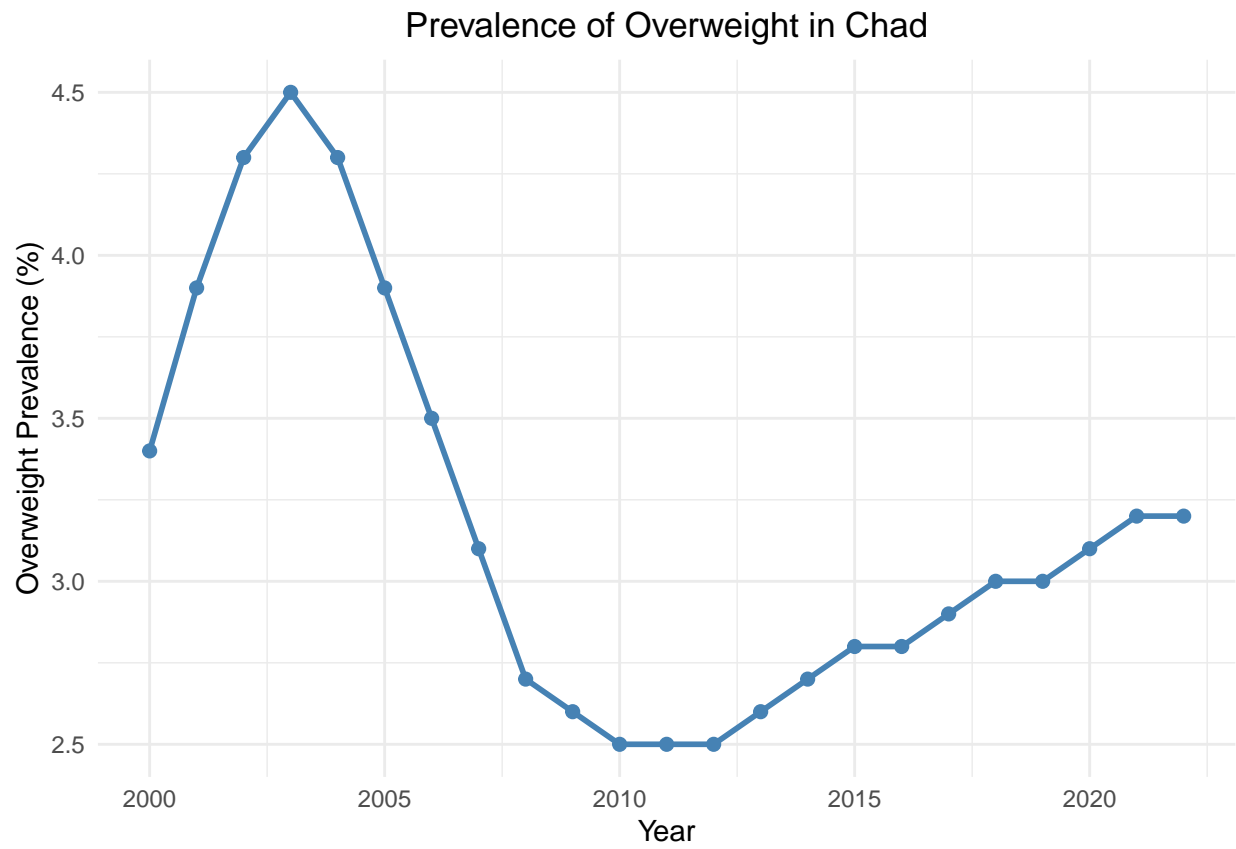


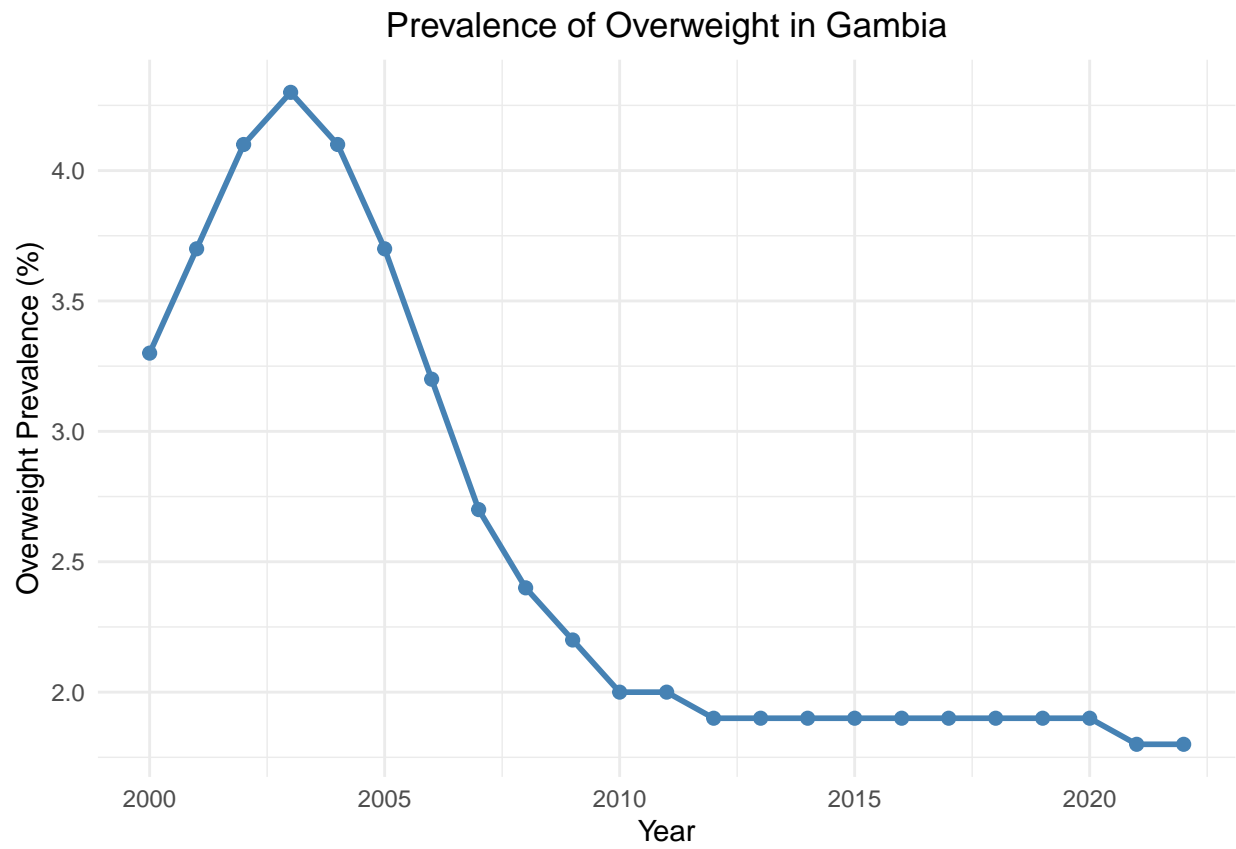




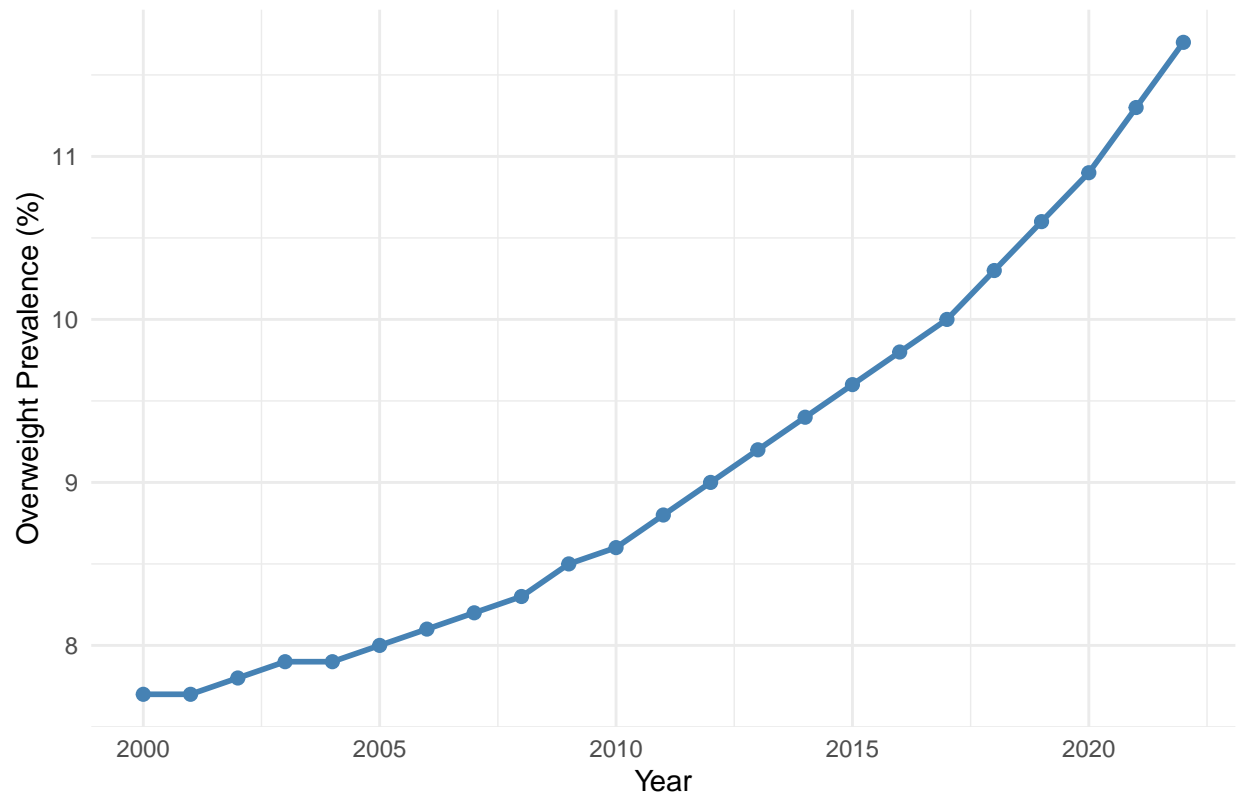


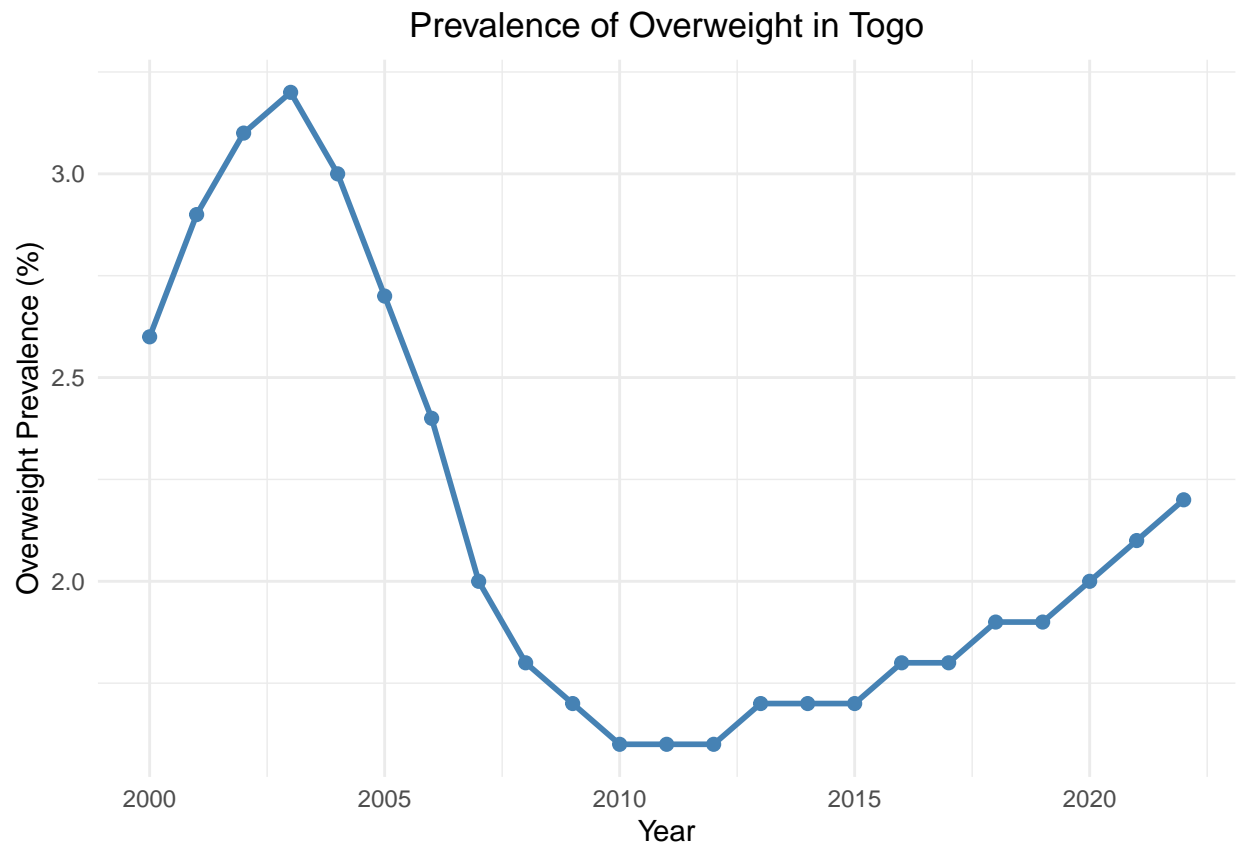




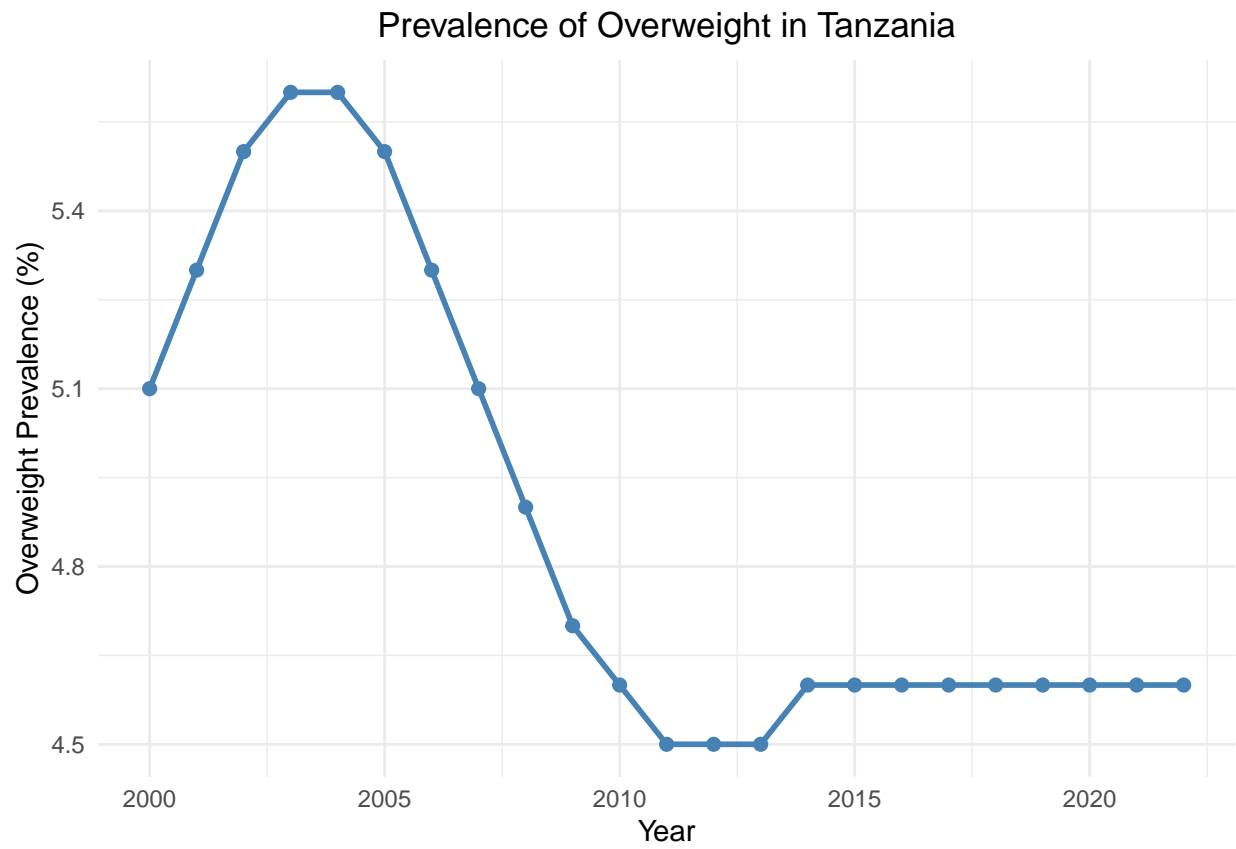


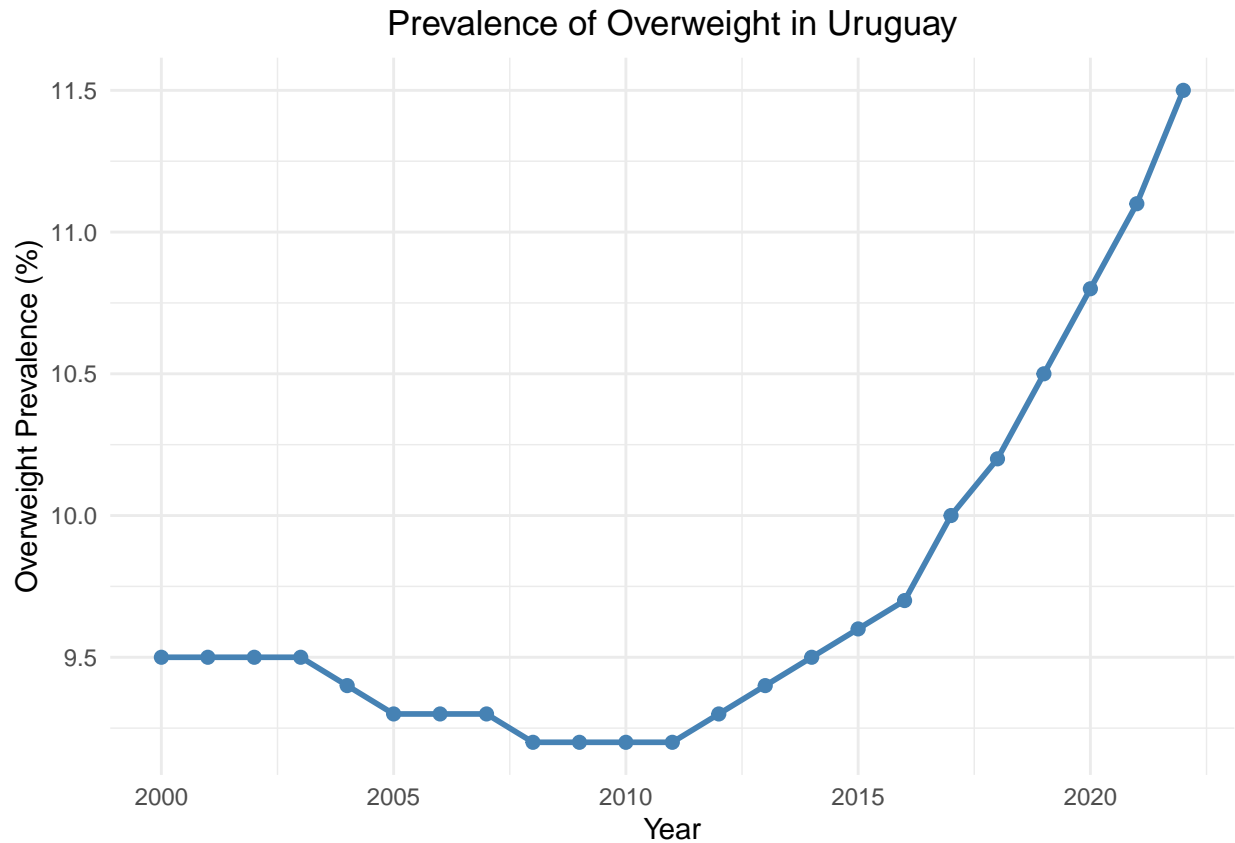
Prevalence of Overweight in Kuwait











countries from the same region and the same sub-region shows different prevalence in overweight. ex: Uruguay, Bolivia, Brazil

most of the countries have declining patterns

```
data$Year <- as.numeric(data$Year)
```

```
data_filtered <- data %>%
  filter(Country %in% selected_countries & !is.na(Overweight_model)) %>%
  group_by(Country) %>%
  filter(Year == max(Year, na.rm = TRUE)) %>%
  select(Country, Year, Overweight_model)
```

```
world_map <- map_data("world")
```

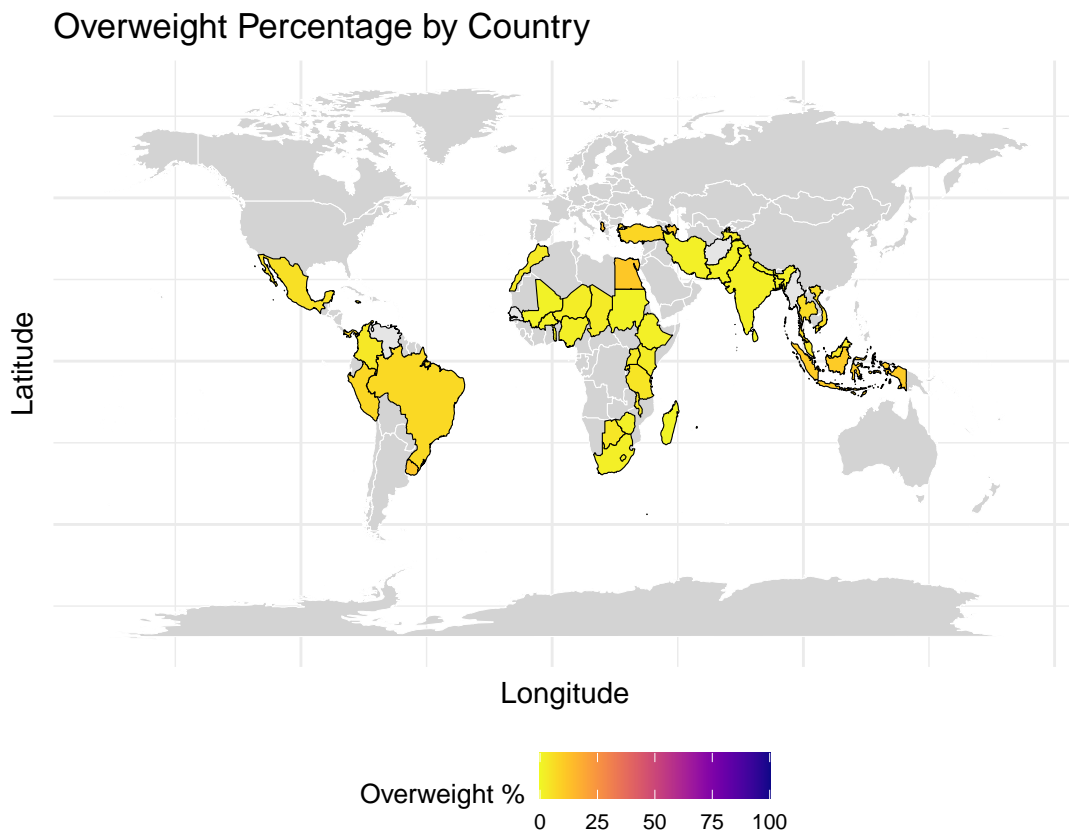
```
highlighted_map <- world_map %>%
  filter(region %in% selected_countries) %>%
  left_join(data_filtered, by = c("region" = "Country"))
```

```
ggplot() +
  geom_polygon(
    data = world_map,
    aes(x = long, y = lat, group = group),
    fill = "lightgray",
    color = "white",
    linewidth = 0.1
  ) +
  geom_polygon(
```

```

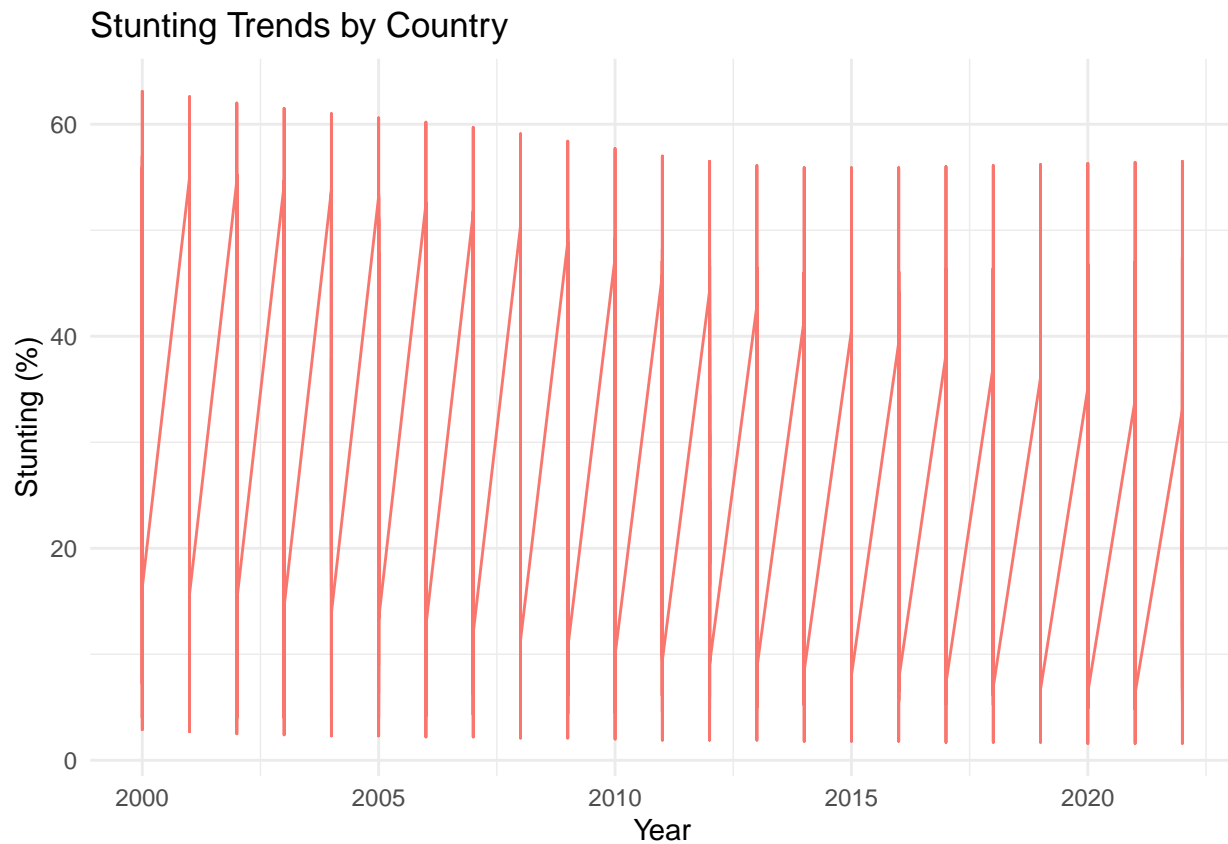
data = highlighted_map %>% mutate(Overweight = as.numeric(Overweight_model)),
aes(x = long, y = lat, group = group, fill = Overweight),
color = "black",
linewidth = 0.2
) +
scale_fill_viridis(
  name = "Overweight %",
  option = "plasma",
  direction = -1,
  na.value = "grey90",
  limits = c(0, 100)
) +
coord_fixed(ratio = 1.3) +
theme_minimal() +
labs(
  title = "Overweight Percentage by Country",
  x = "Longitude",
  y = "Latitude"
) +
theme(
  legend.position = "bottom",
  legend.title = element_text(size = 10),
  legend.text = element_text(size = 8),
  axis.text = element_blank(),
  axis.ticks = element_blank()
)

```



```
#Stunting
```

```
ggplot(data, aes(x = Year, y = Stunting_model, group = country, color = country)) +  
  geom_line(show.legend = FALSE) +  
  theme_minimal() +  
  labs(title = "Stunting Trends by Country", y = "Stunting (%)")
```



```
stunting_data <- data %>%  
  select(Country, Year, Stunting_model) %>%  
  filter(!is.na(Stunting_model))  
  
stunting_data$Stunting <- as.numeric(stunting_data$Stunting)  
stunting_data$Year <- as.numeric(as.character(stunting_data$Year))  
  
countries <- unique(stunting_data$Country)  
  
y_max <- max(stunting_data$Stunting, na.rm = TRUE)  
  
for (country in countries) {  
  country_data <- stunting_data %>% filter(Country == country)  
  
  if (nrow(country_data) < 2) next # Skip if not enough data  
  
  p <- ggplot(country_data, aes(x = Year, y = Stunting_model)) +
```

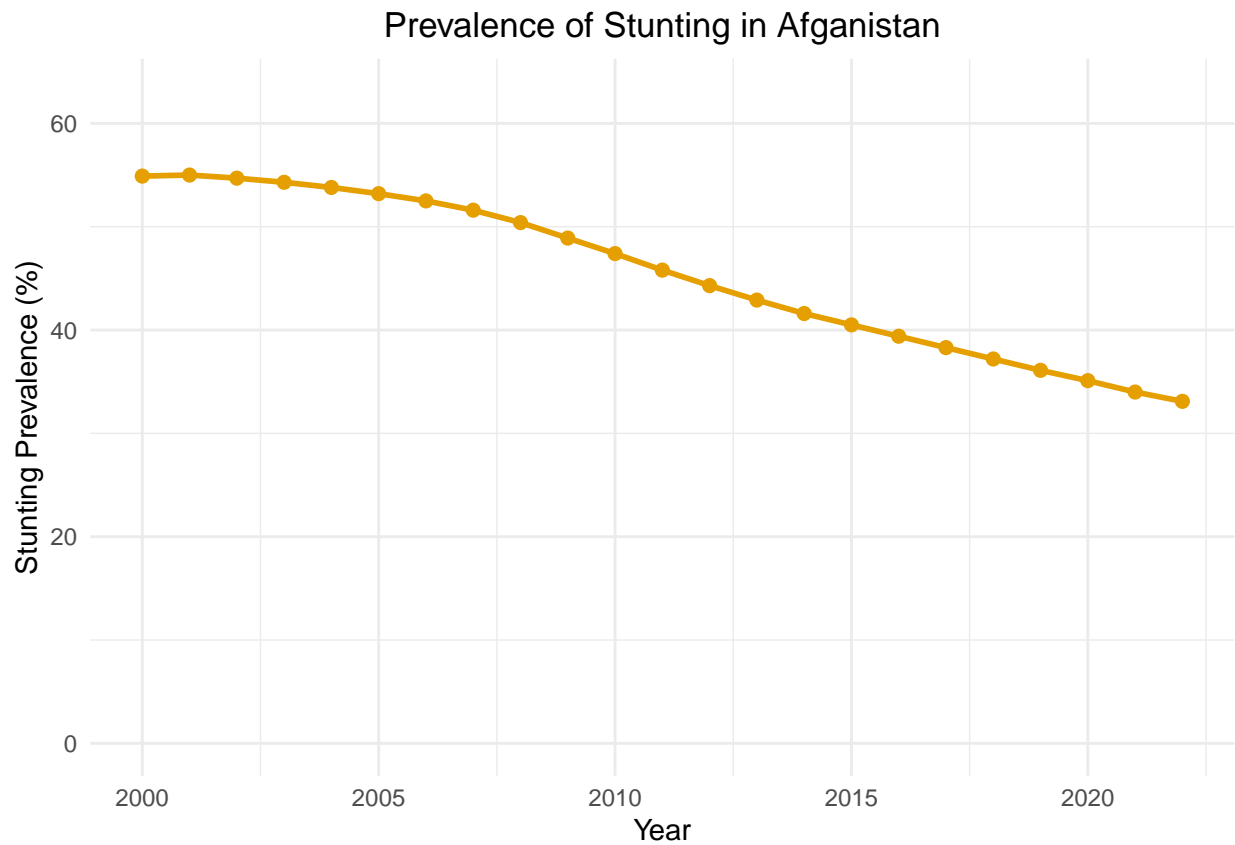
```

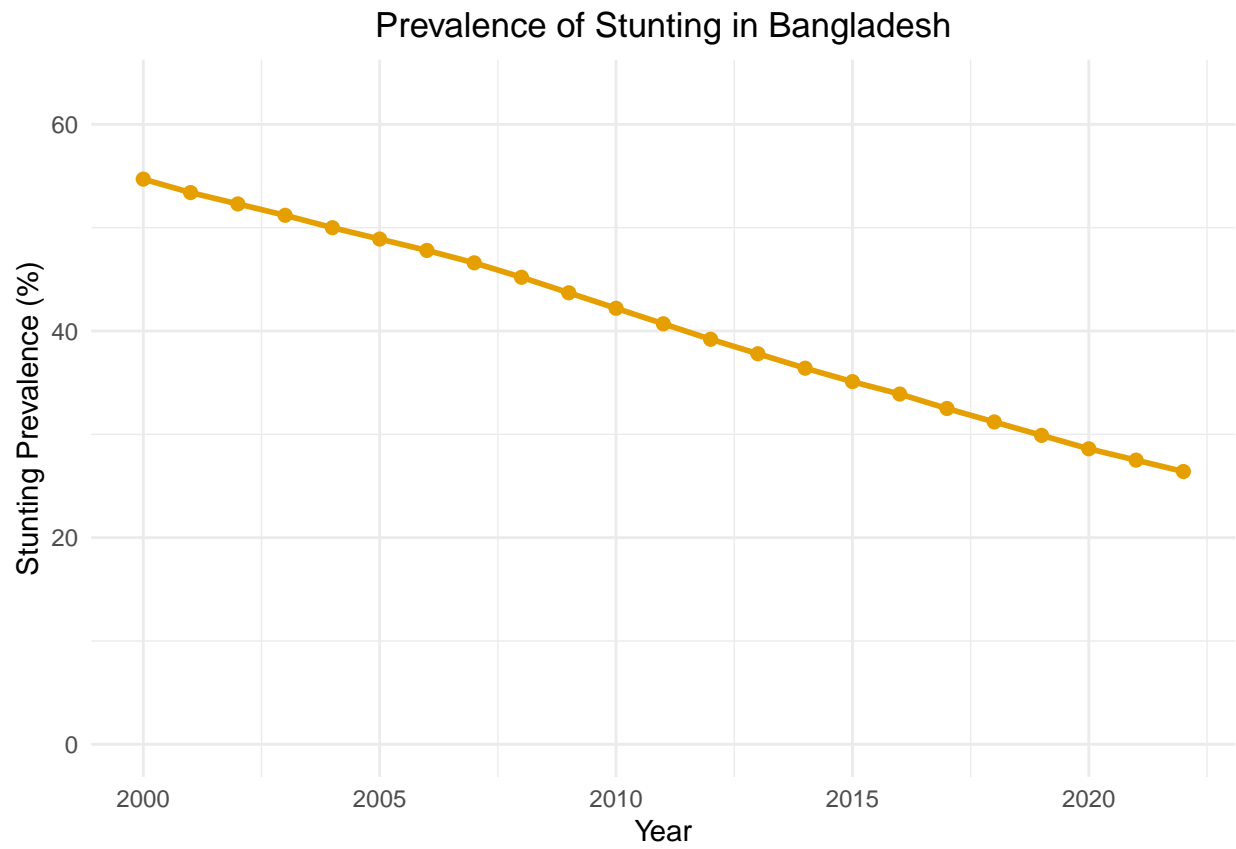
geom_line(color = "#E69F00", size = 1) +
geom_point(color = "#E69F00", size = 2) +
labs(title = paste("Prevalence of Stunting in", country),
      x = "Year",
      y = "Stunting Prevalence (%)") +
theme_minimal() +
theme(plot.title = element_text(hjust = 0.5)) +
coord_cartesian(ylim = c(0, y_max))

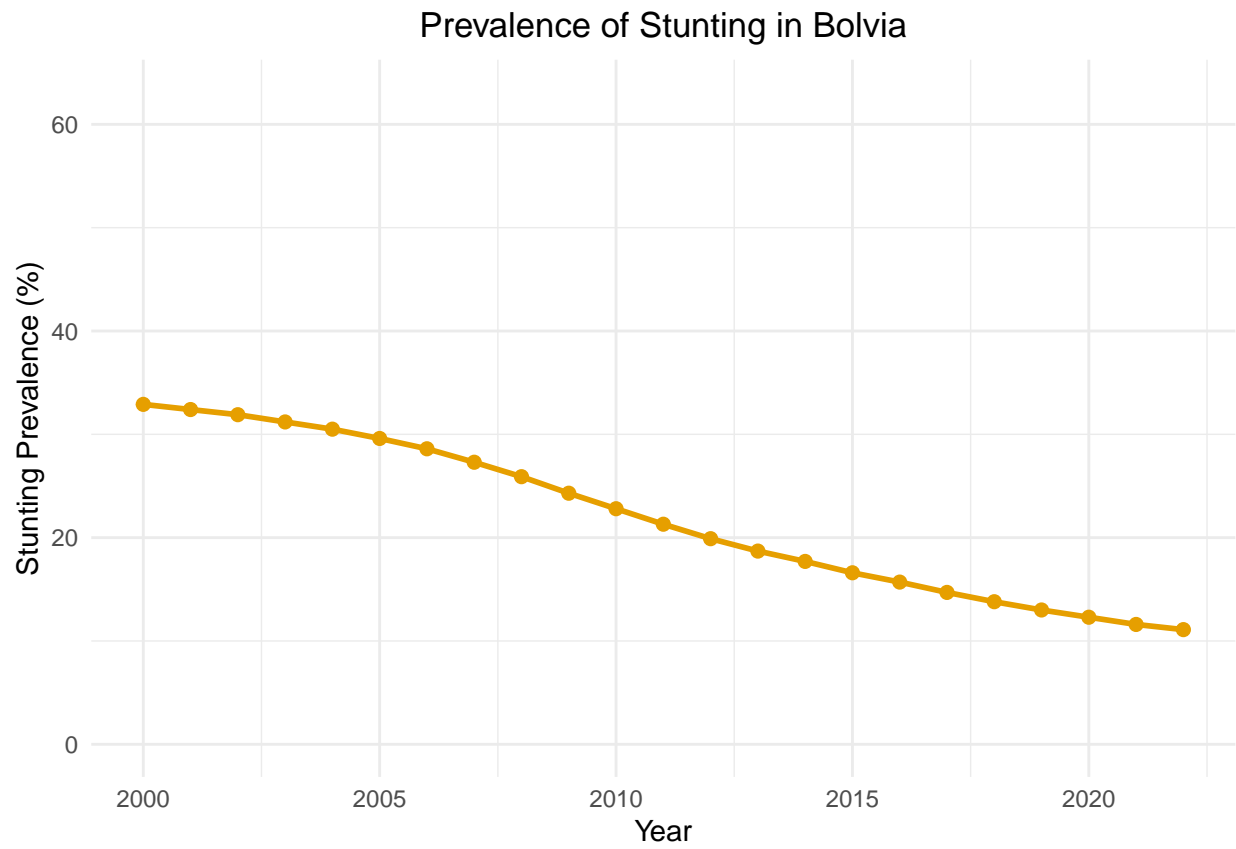
print(p)

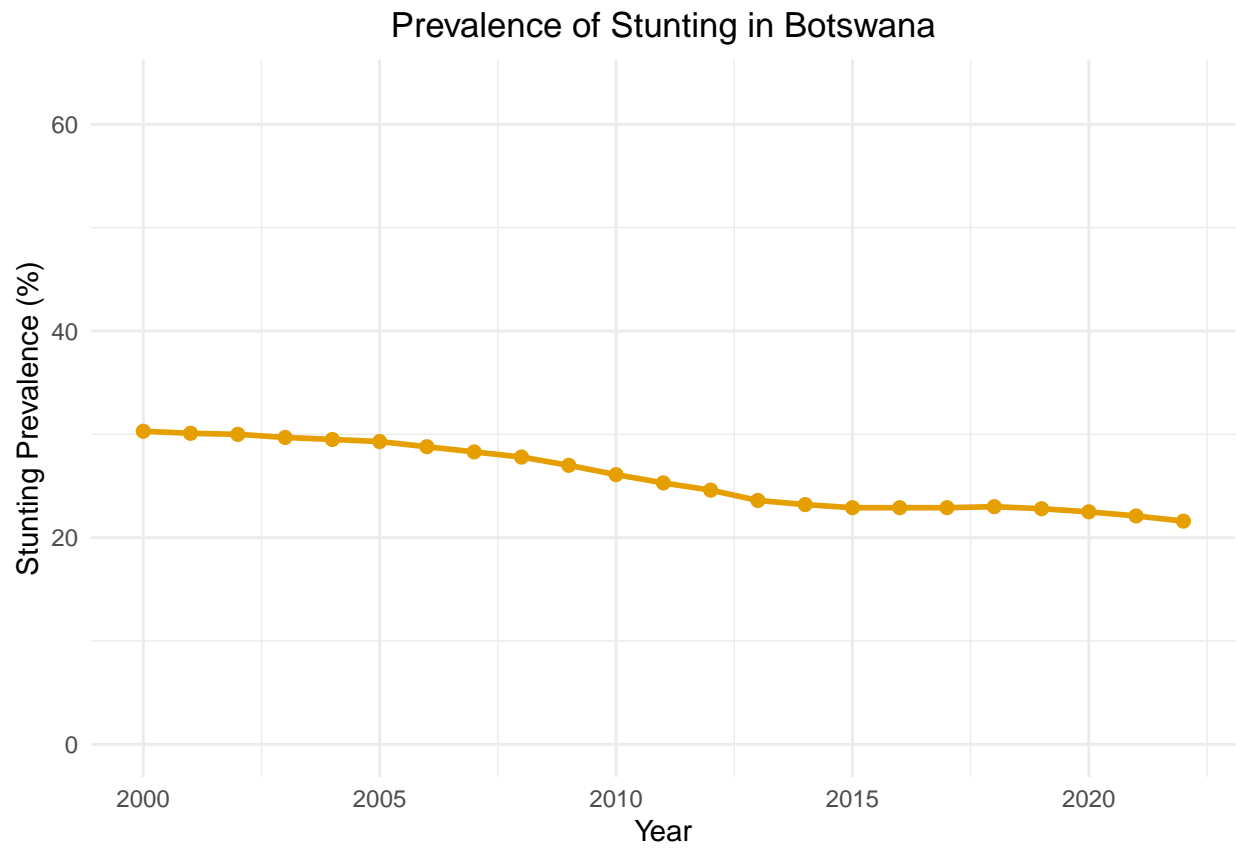
safe_country <- gsub("[^A-Za-z0-9_]", "_", country)
ggsave(filename = paste0("stunting_", safe_country, ".png"),
        plot = p, width = 8, height = 6, dpi = 300)
}

```

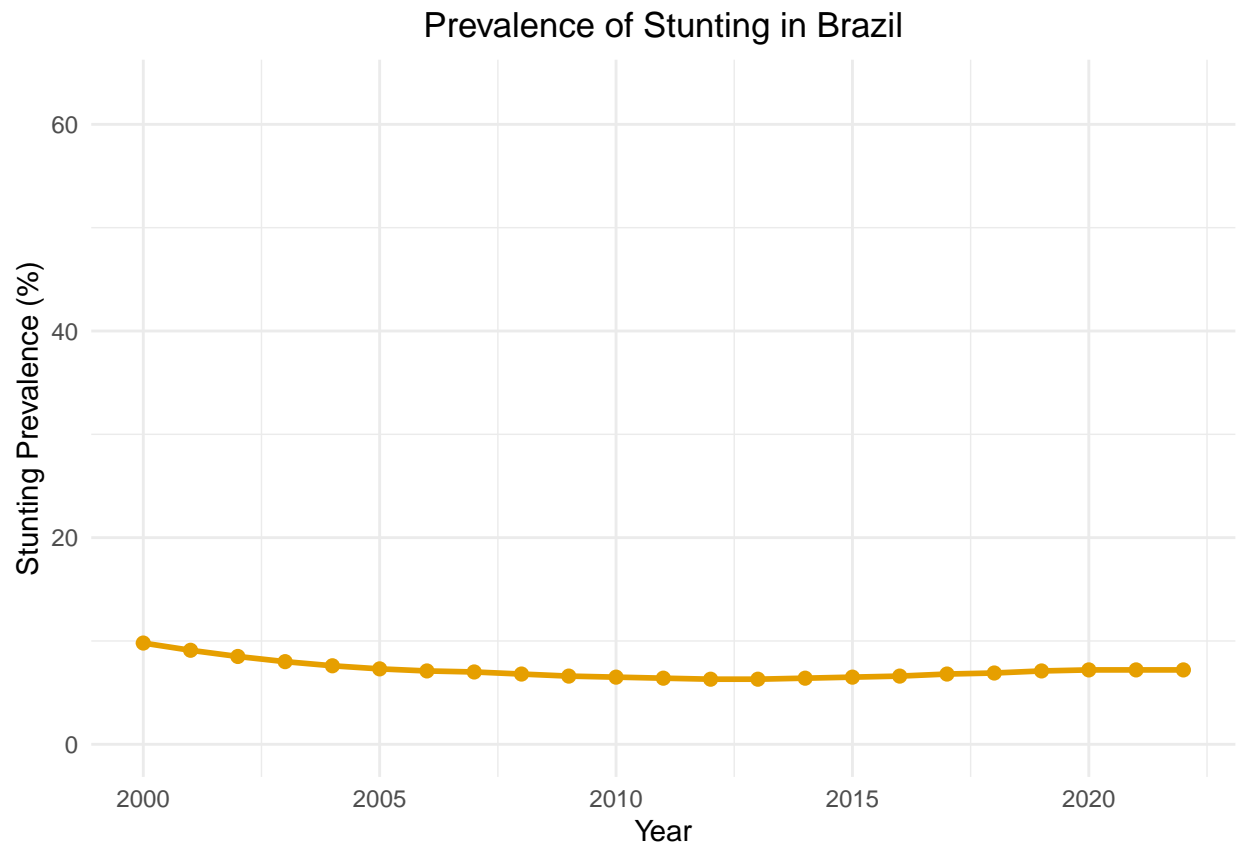


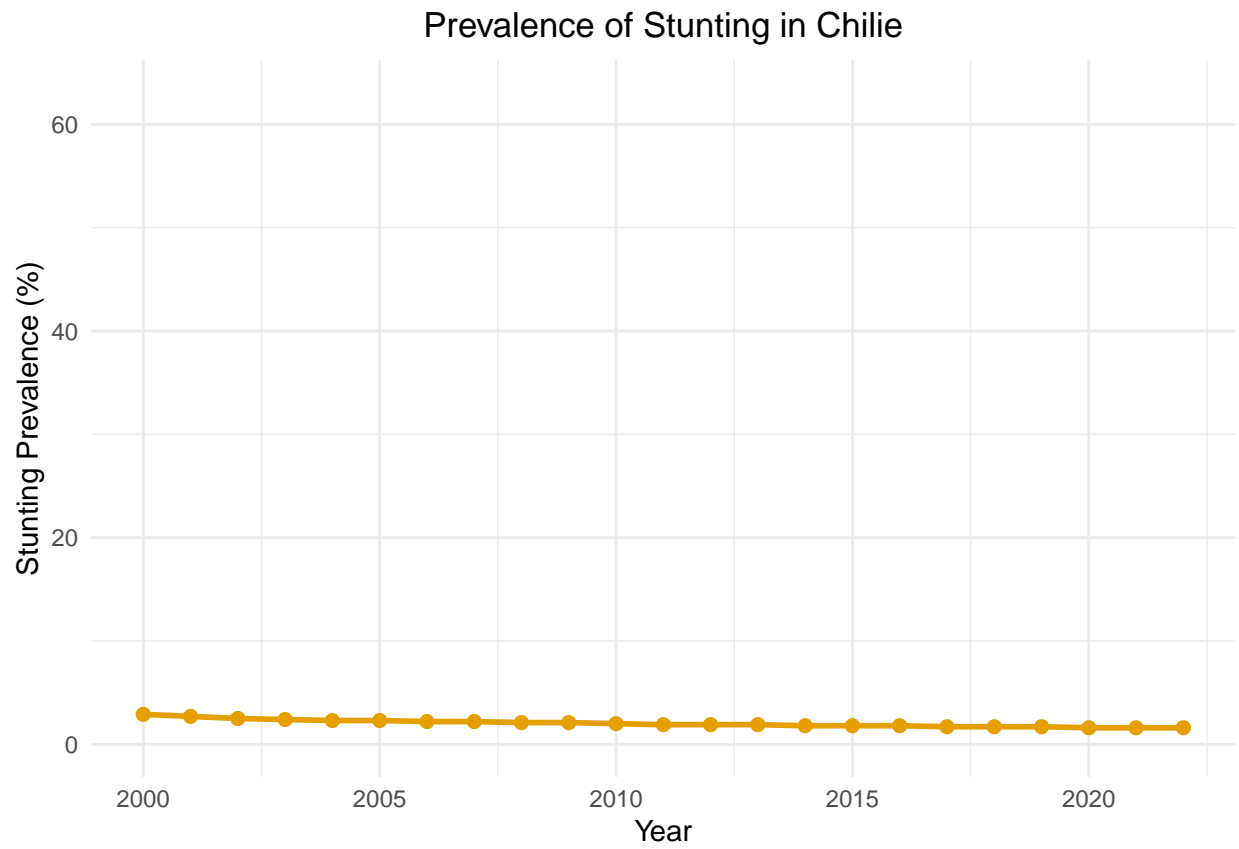


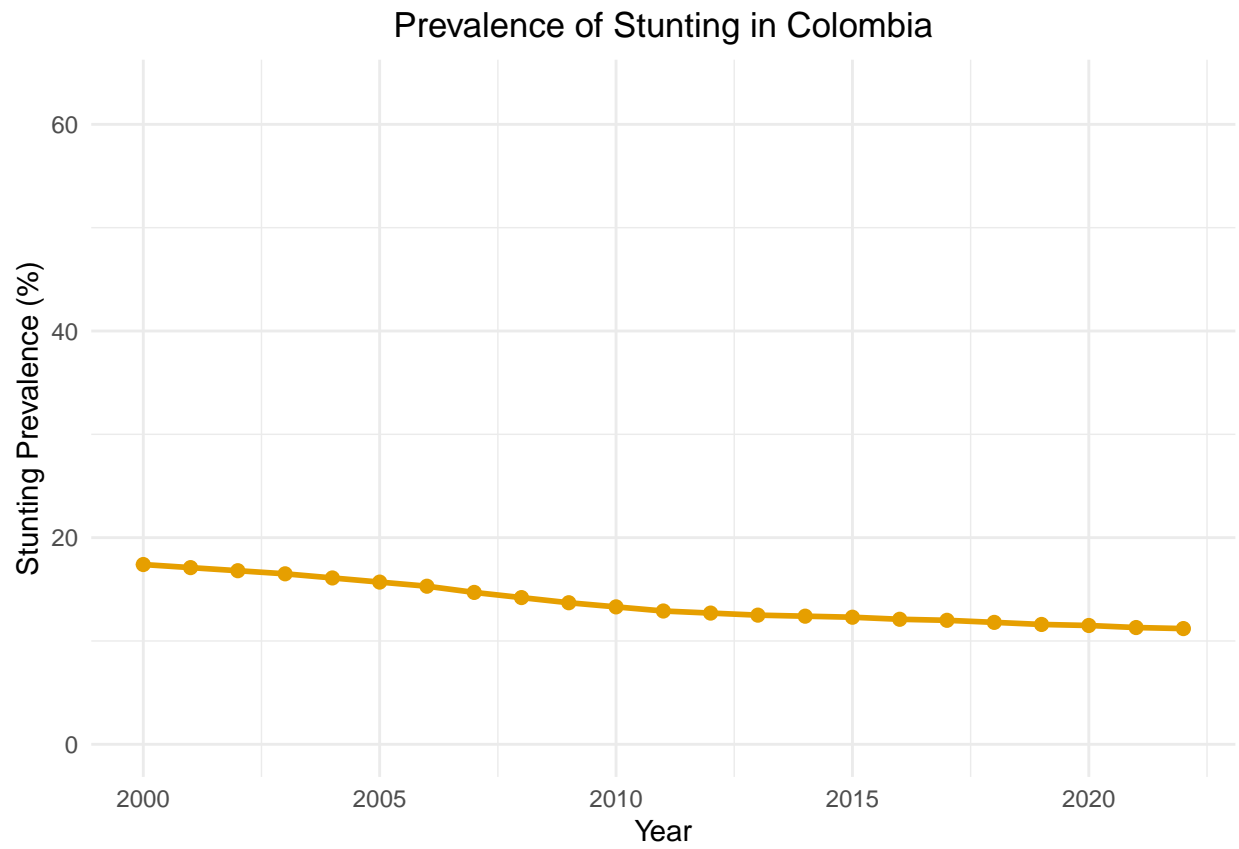


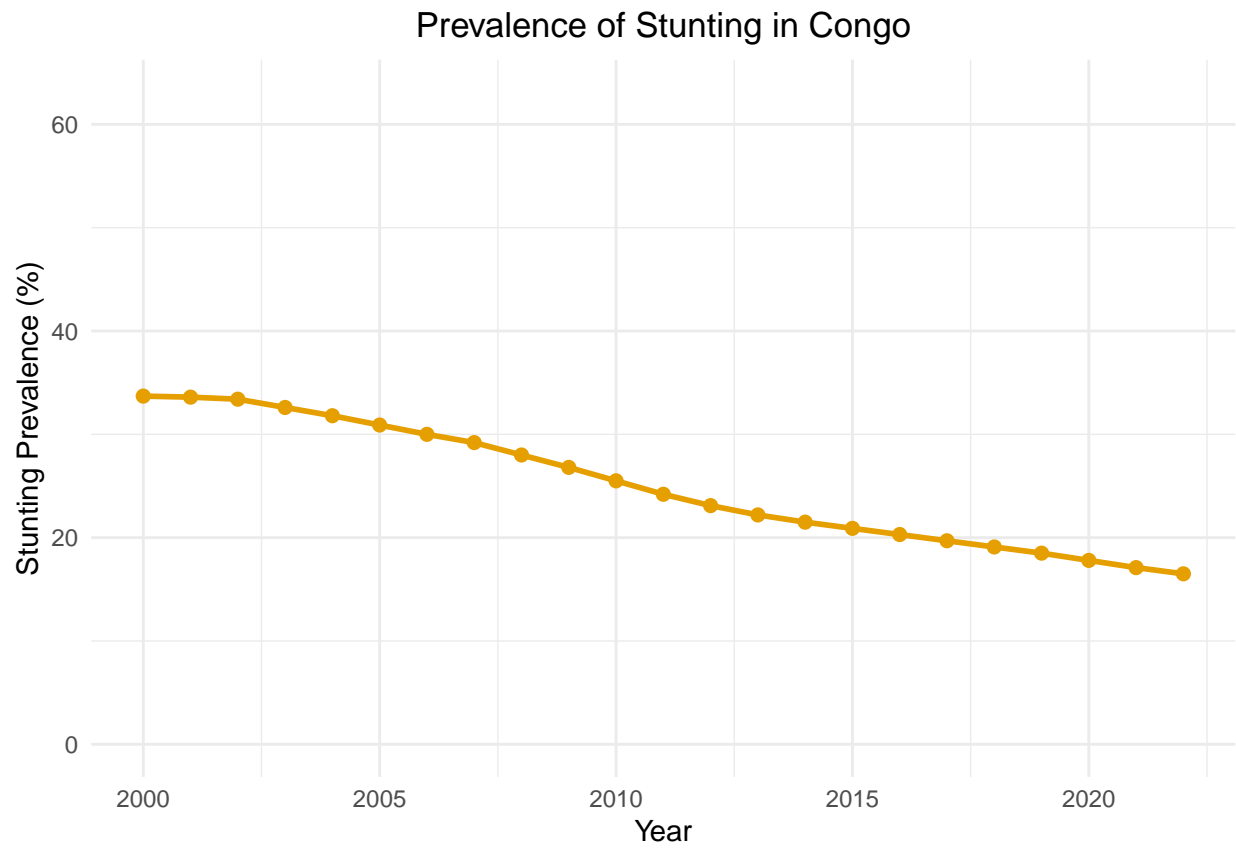


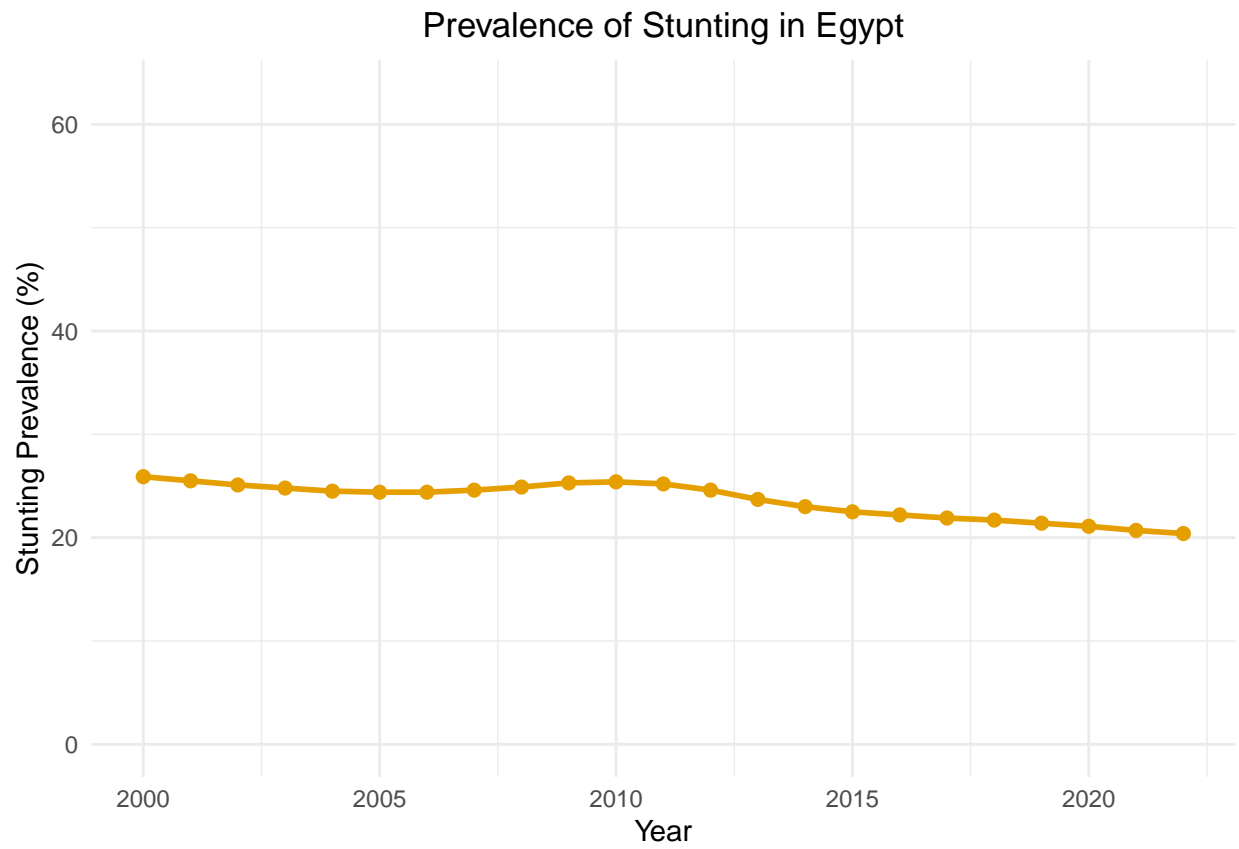


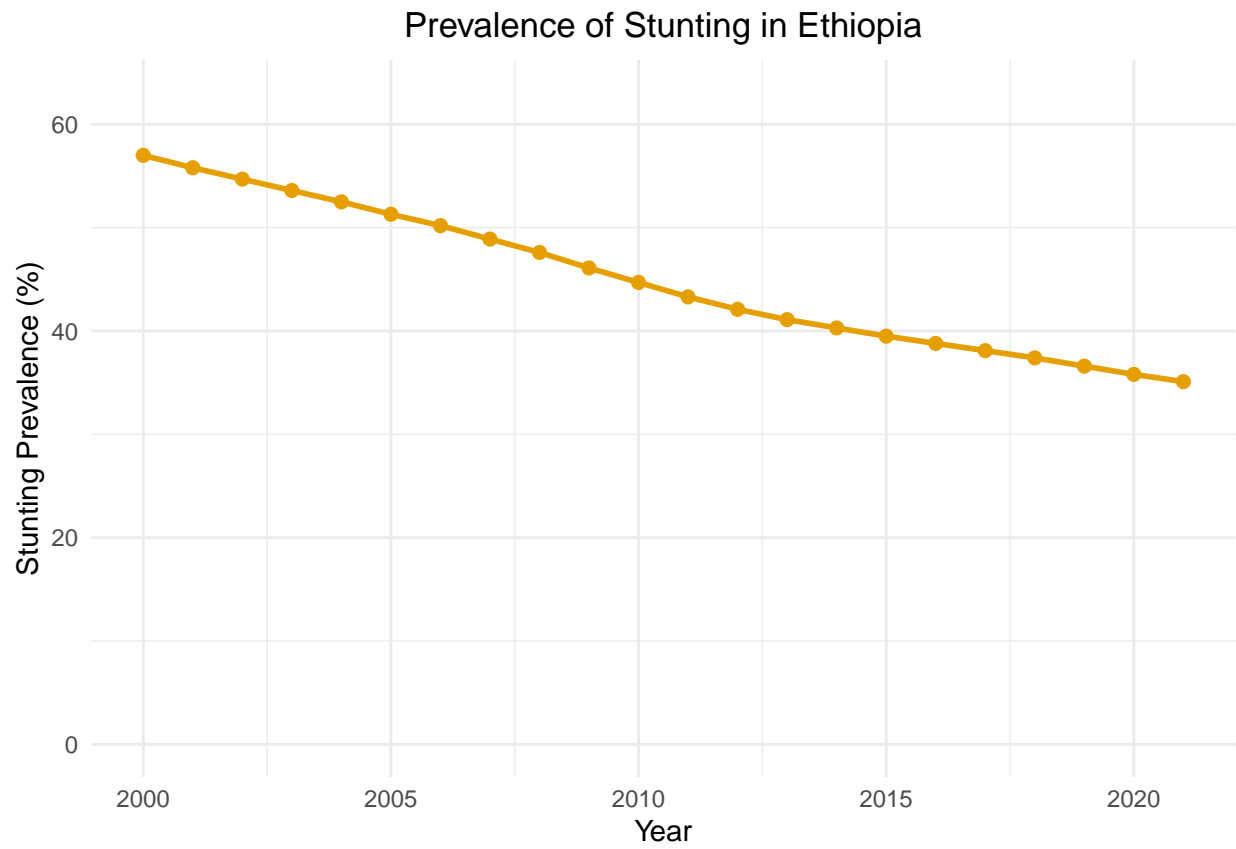


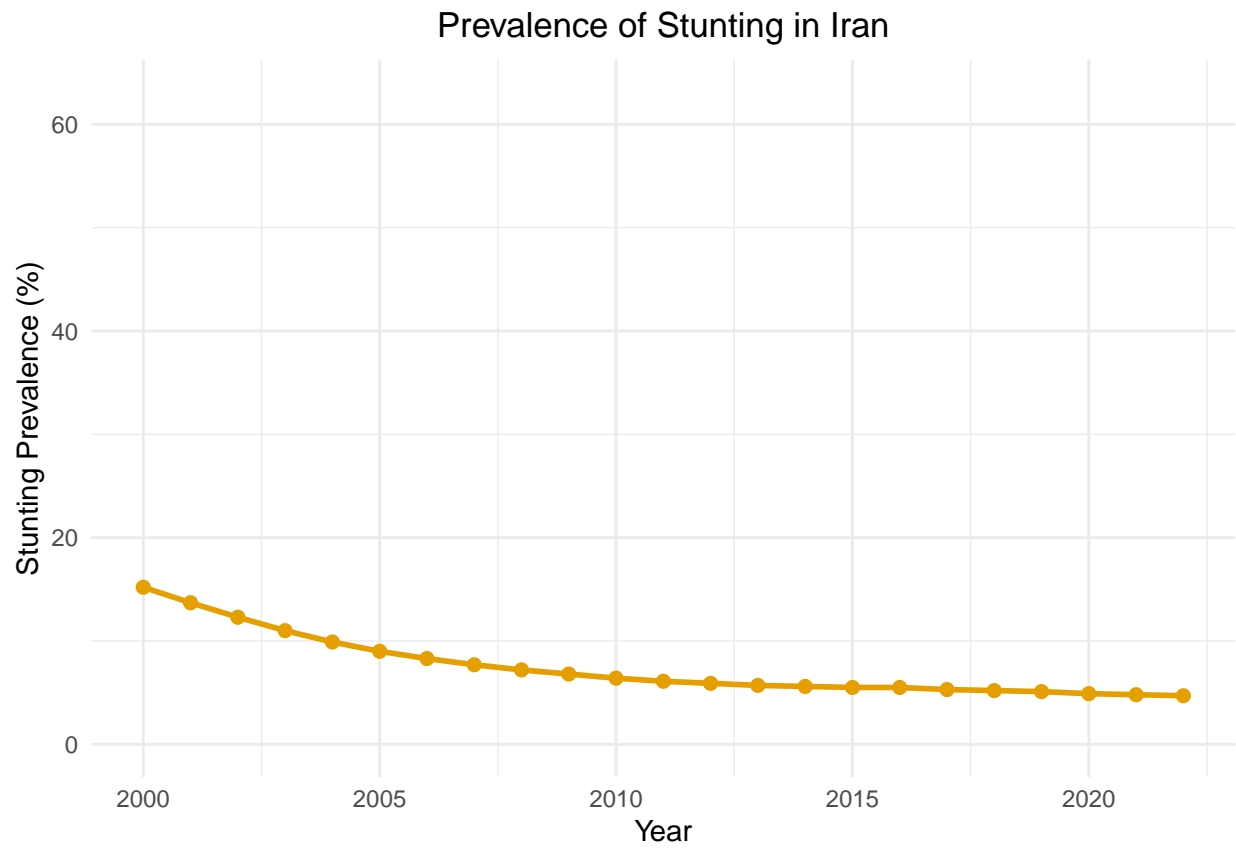


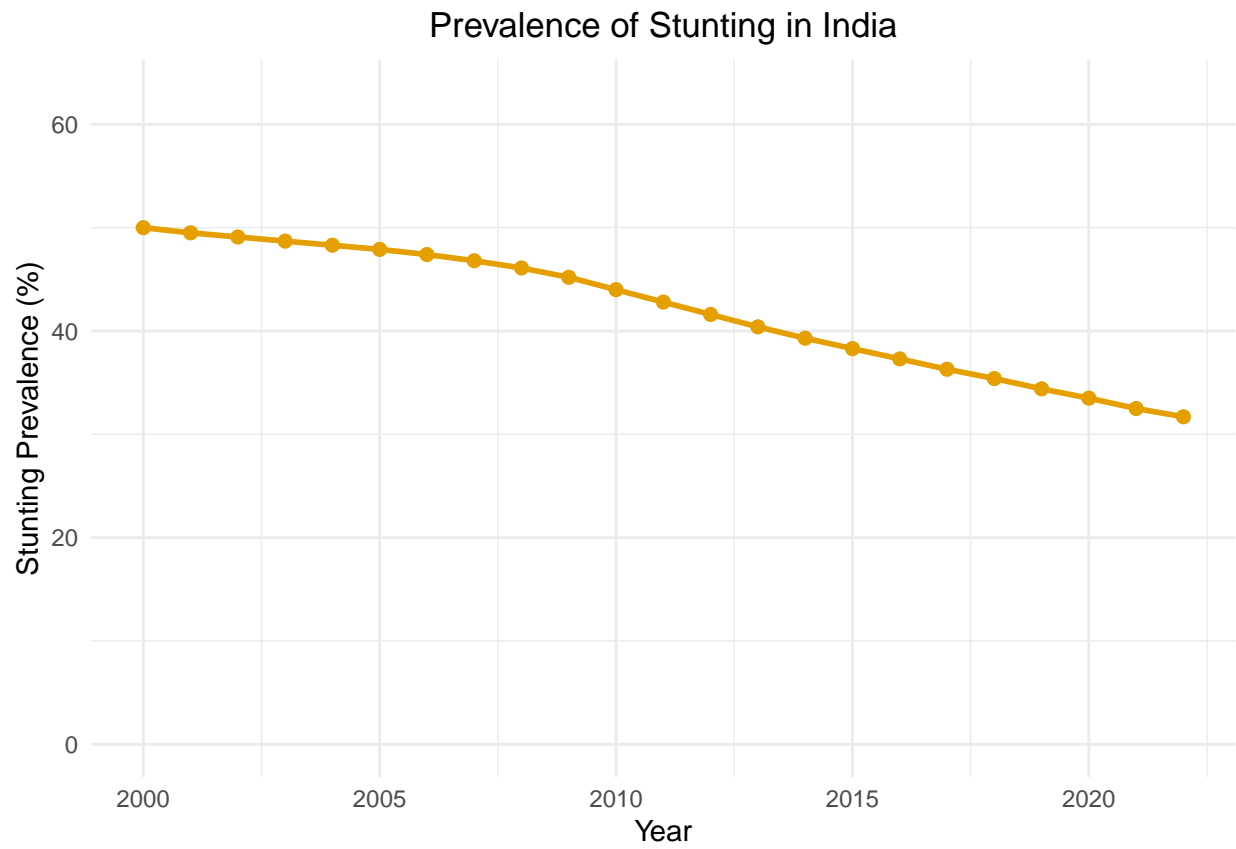




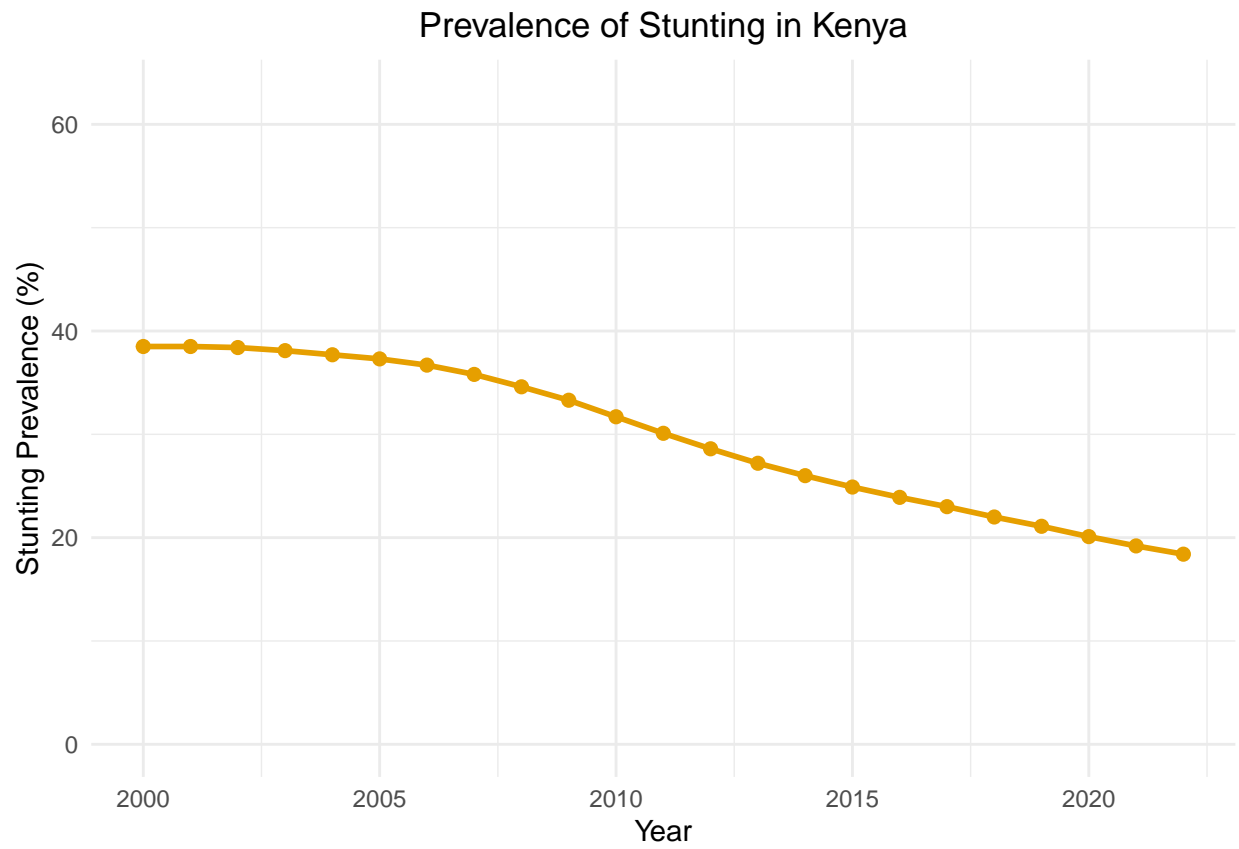


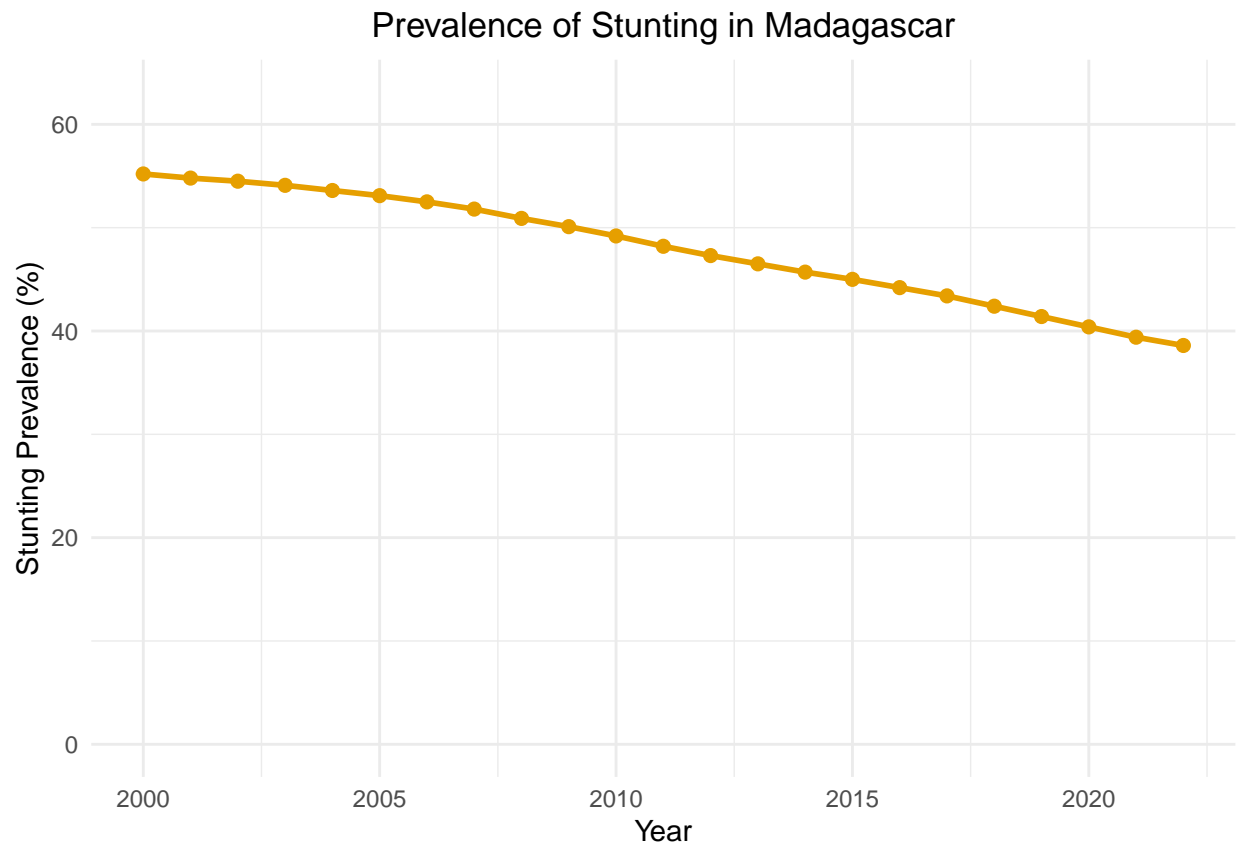


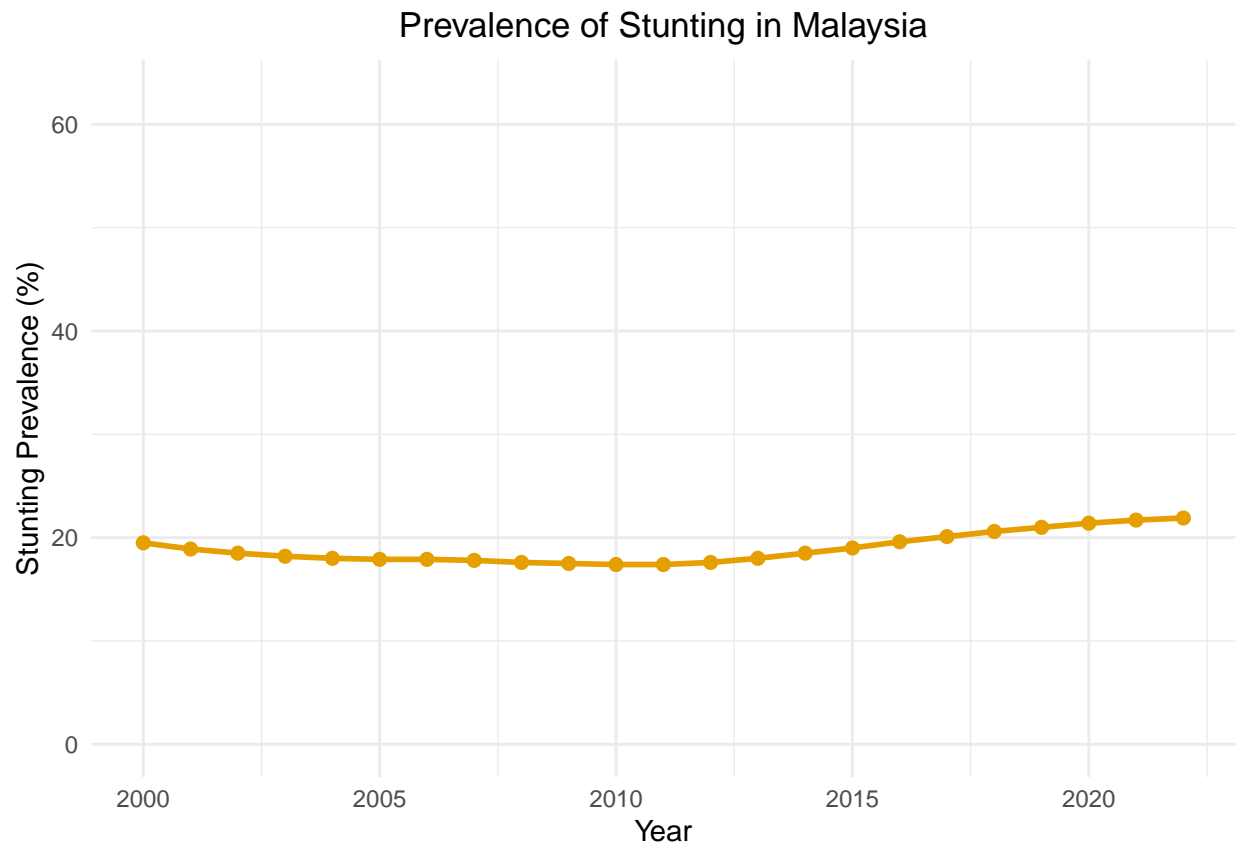


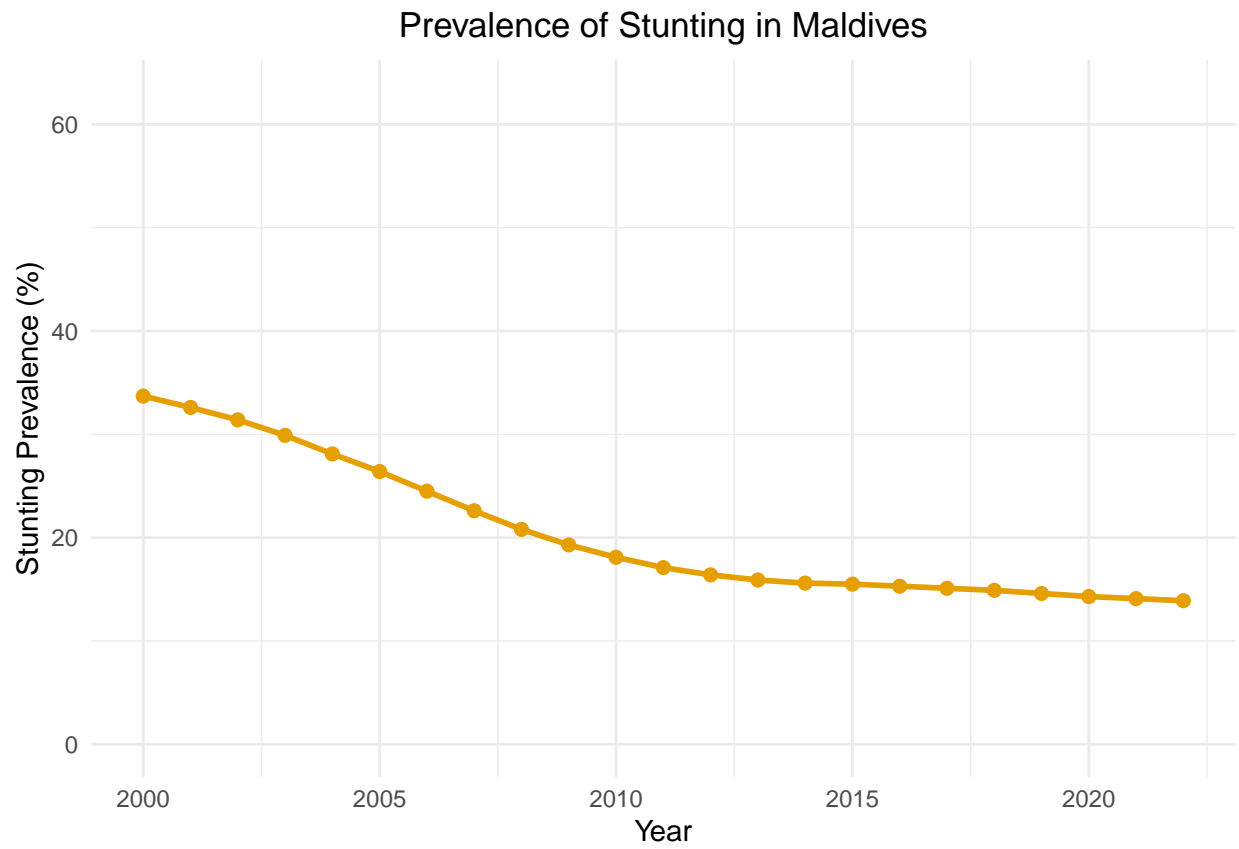


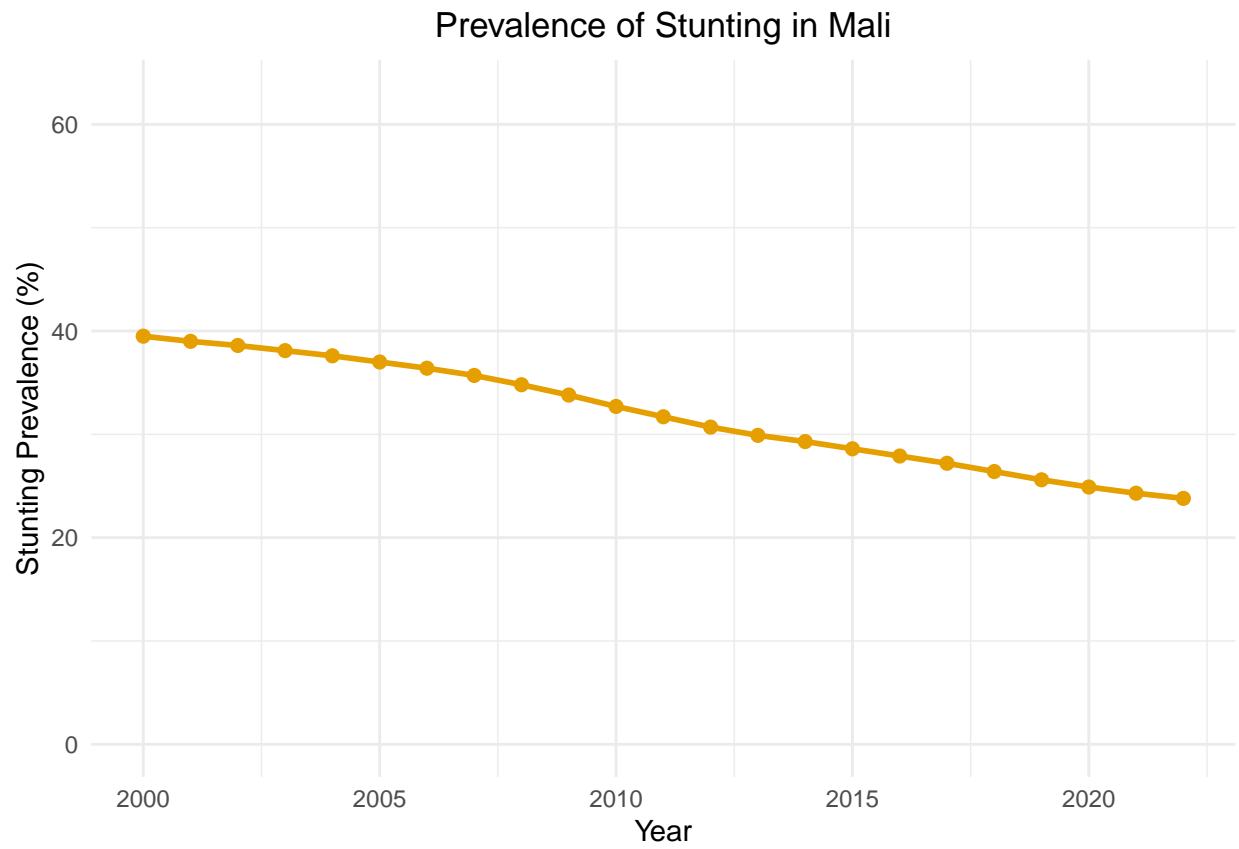


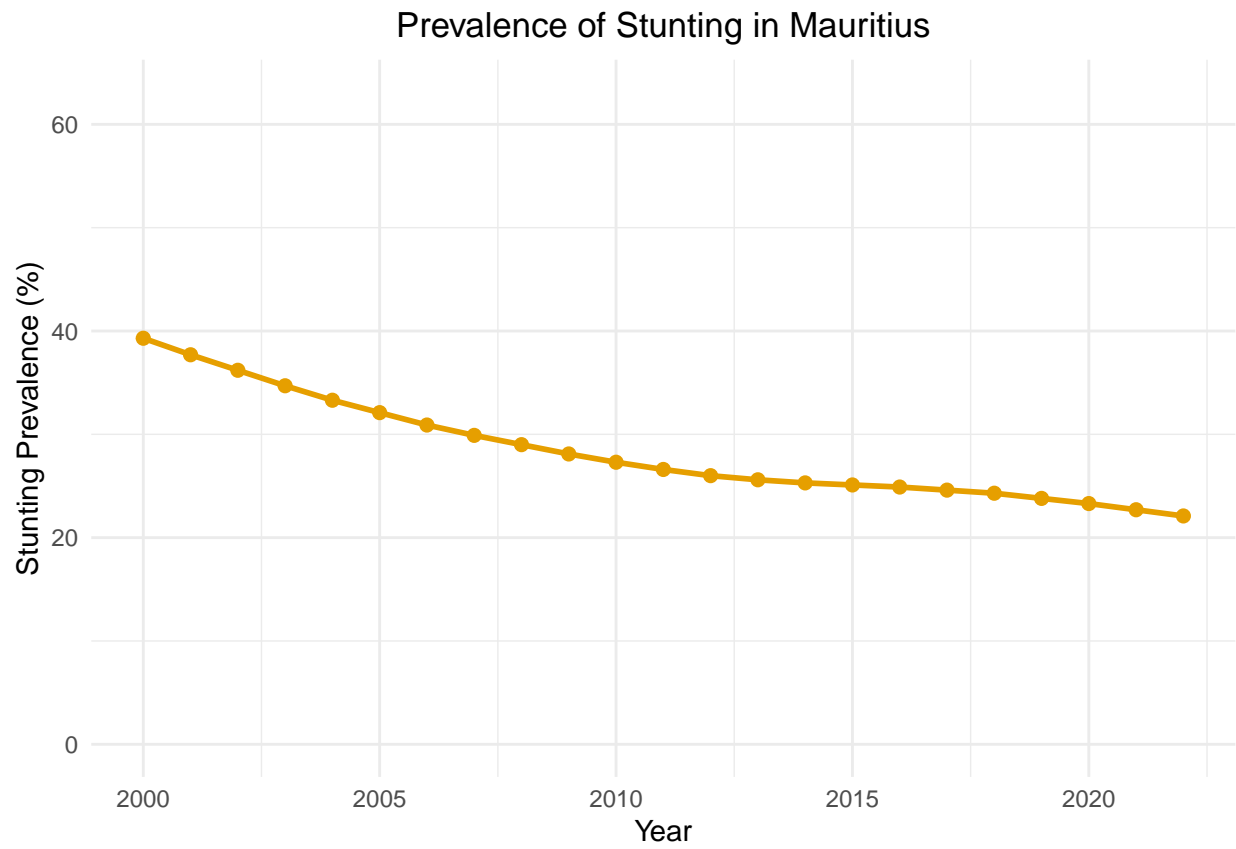


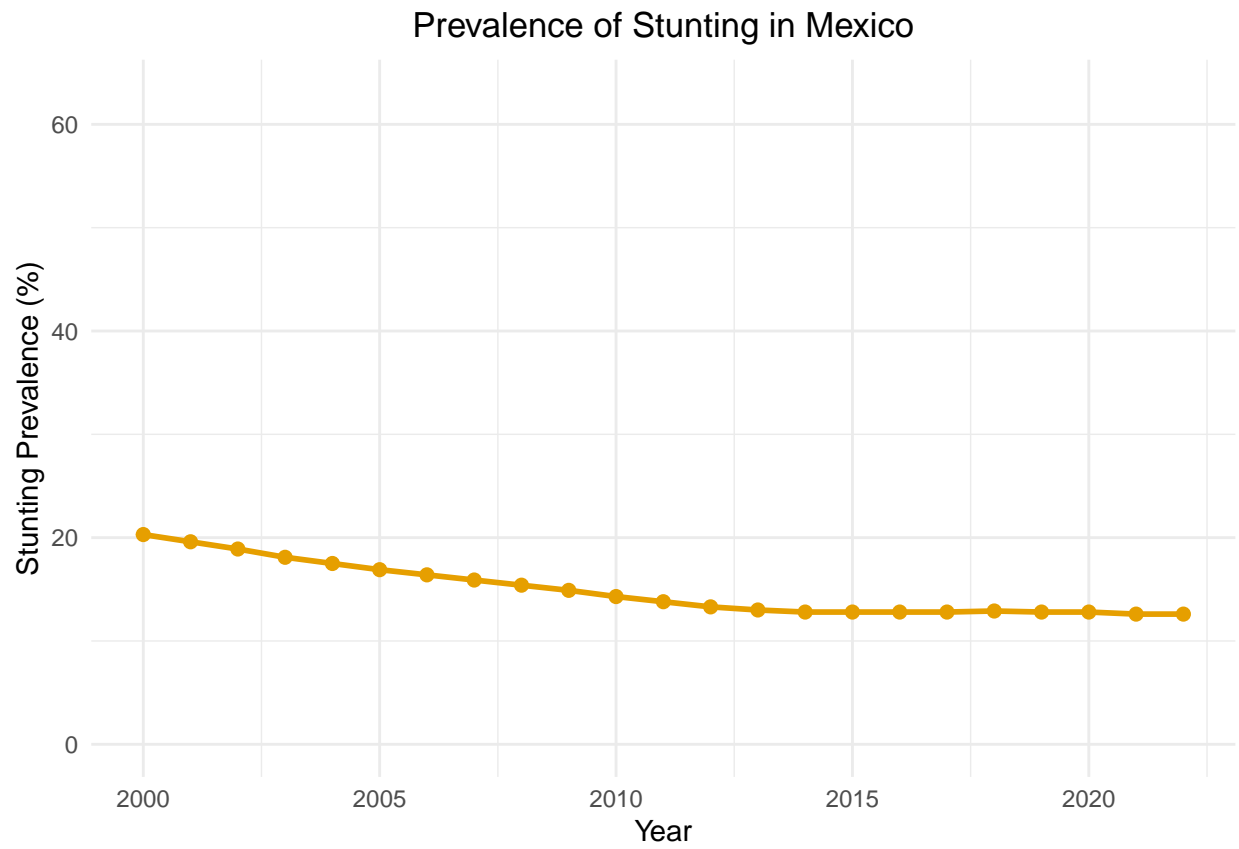


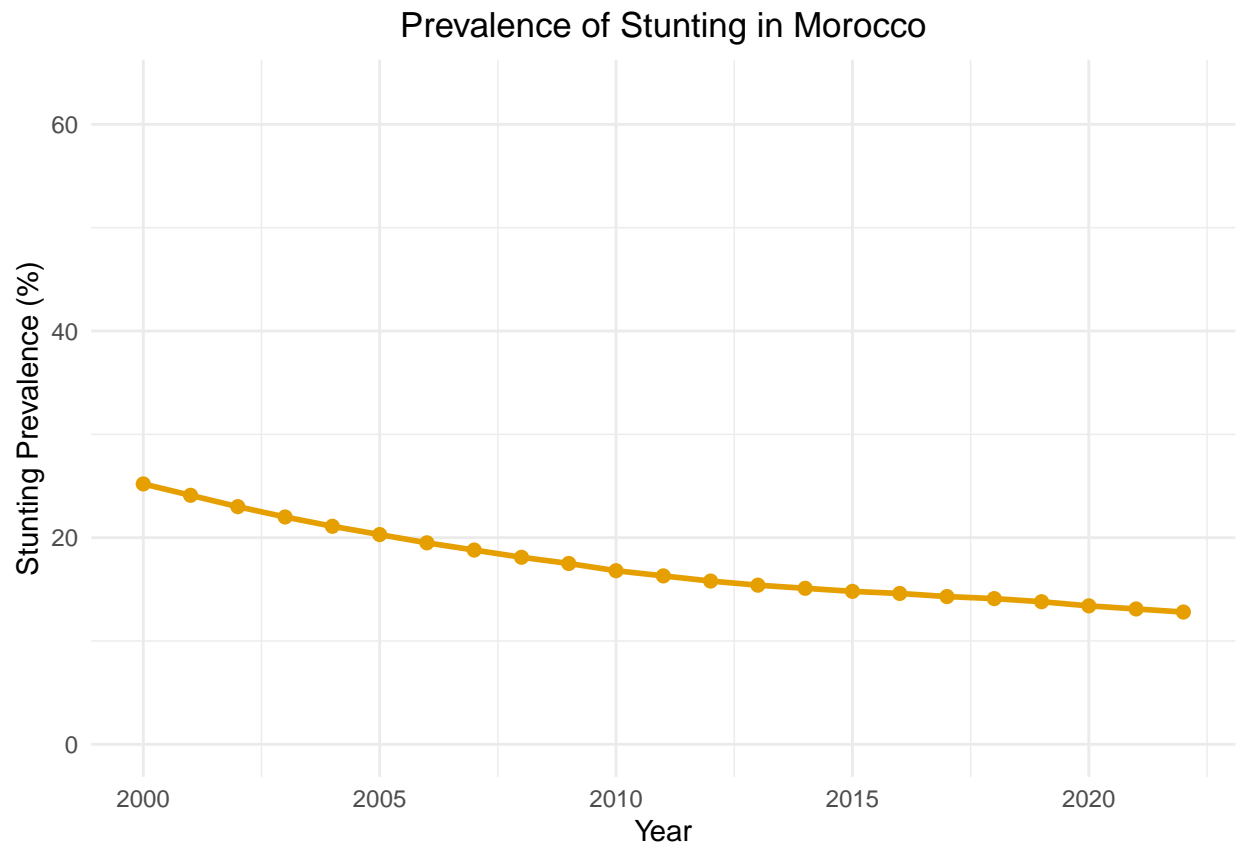




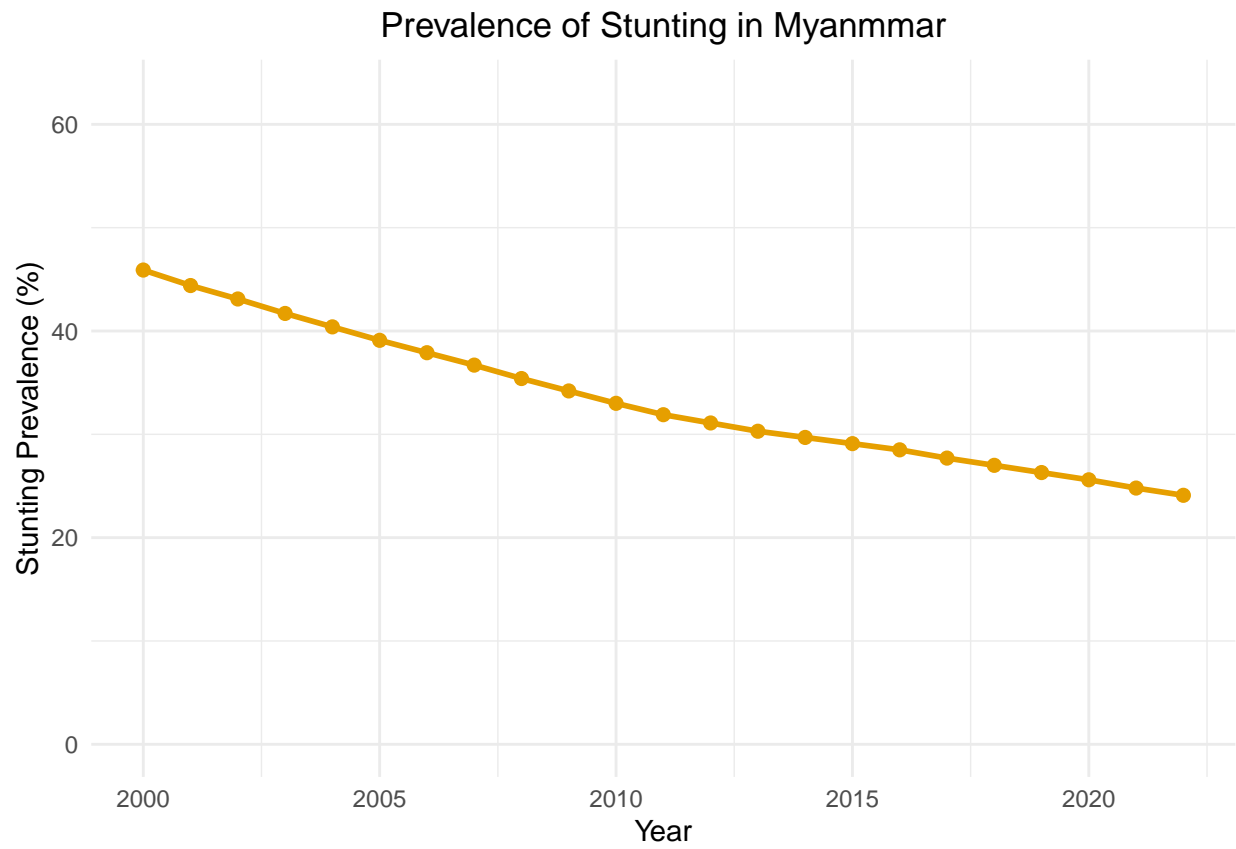


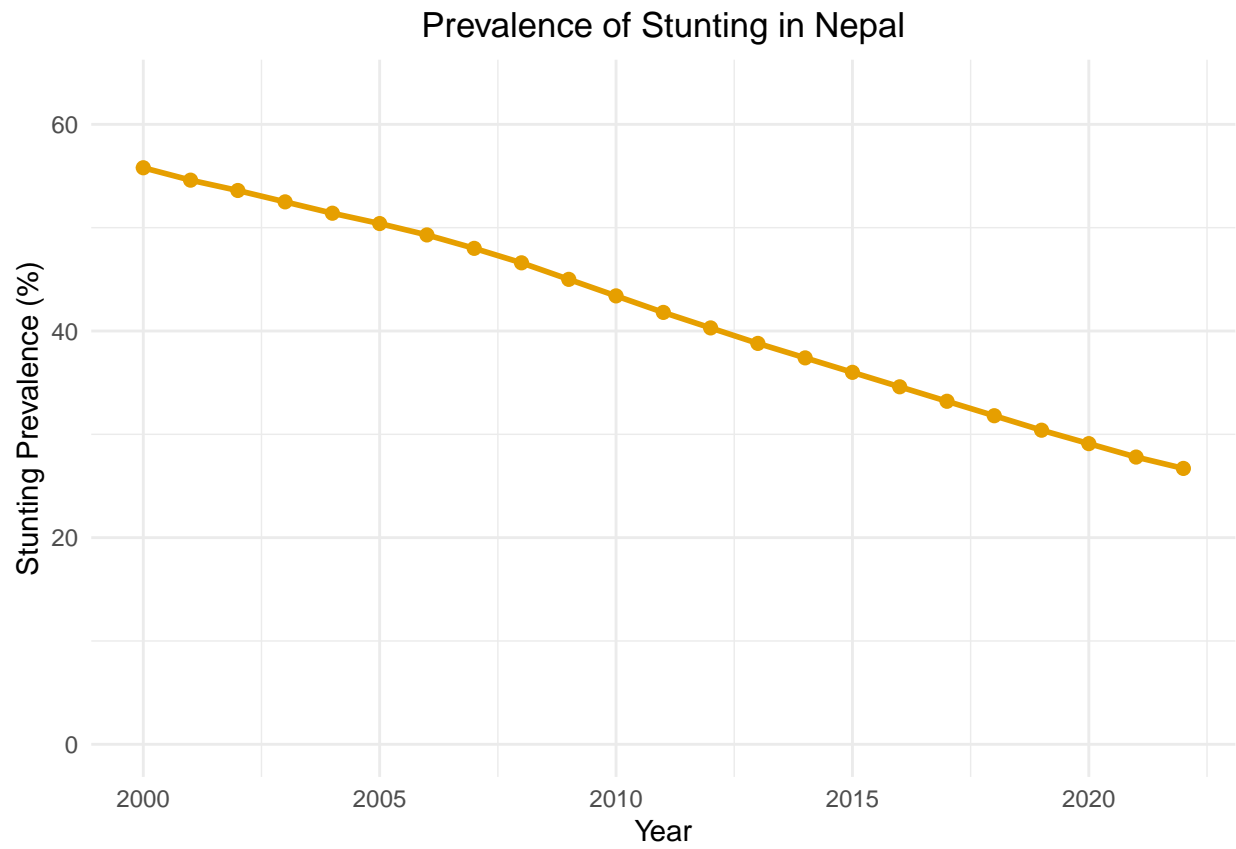


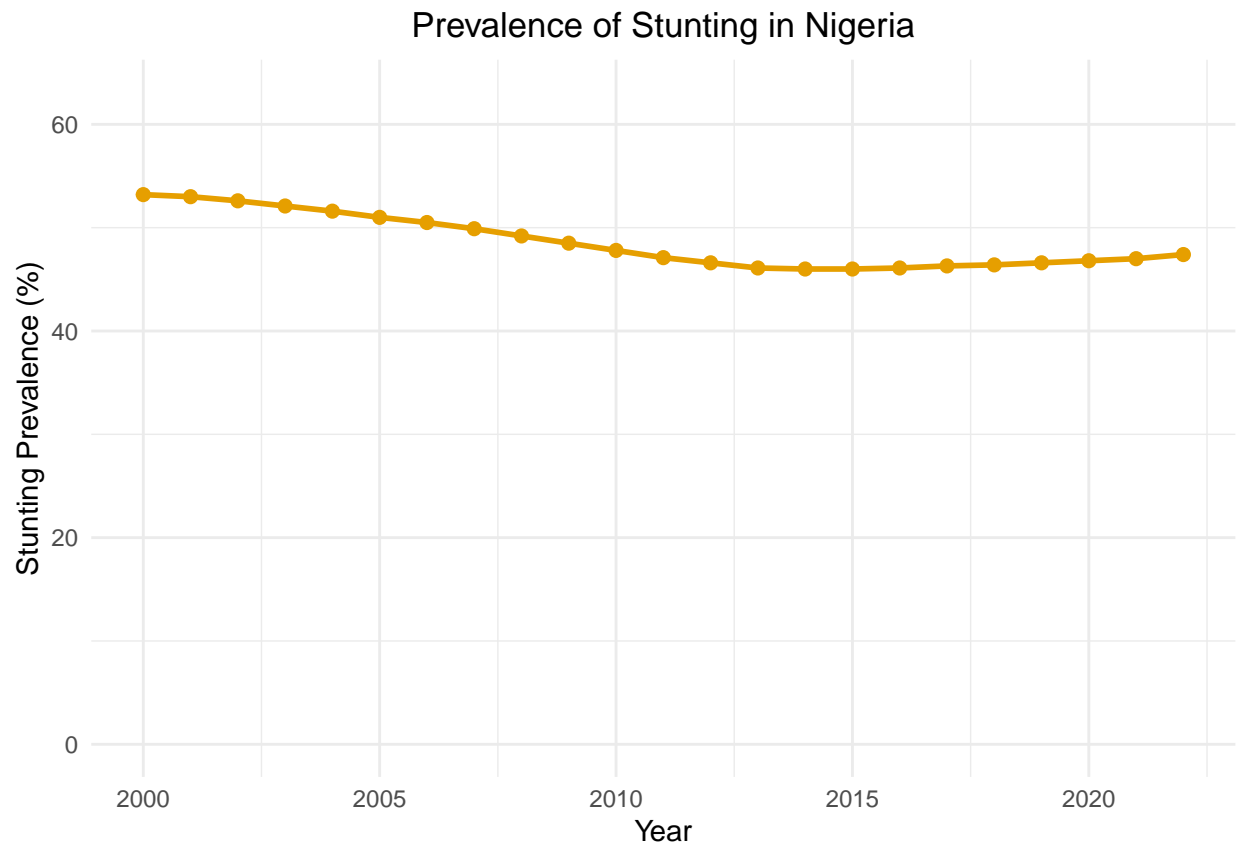


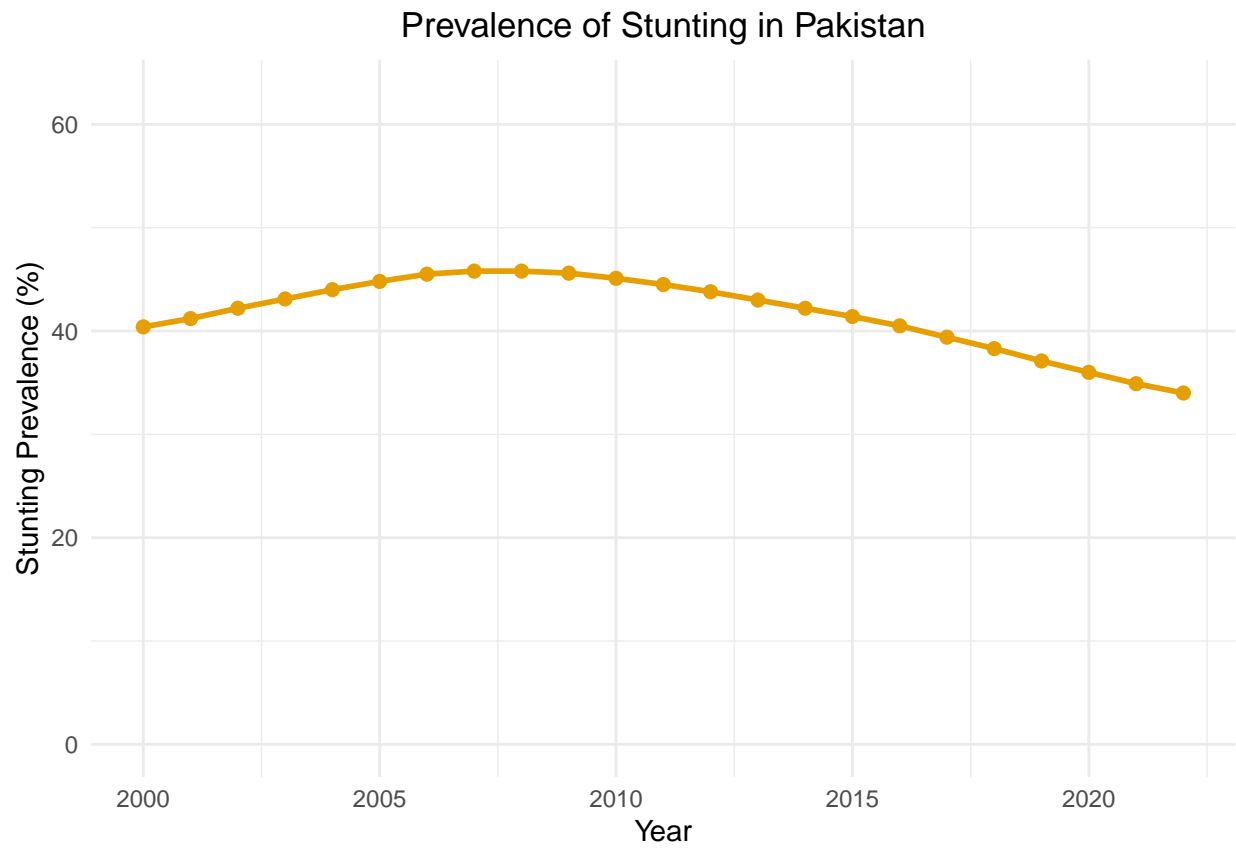


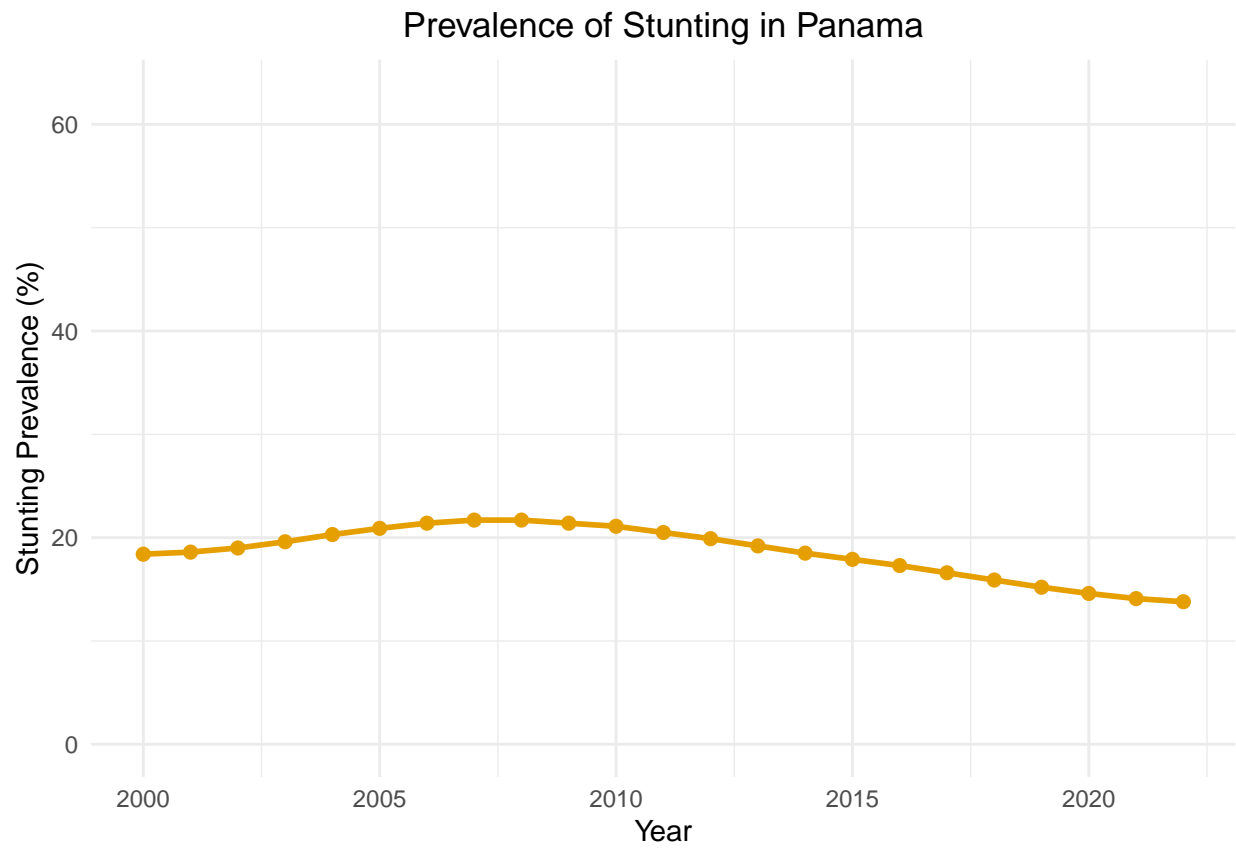


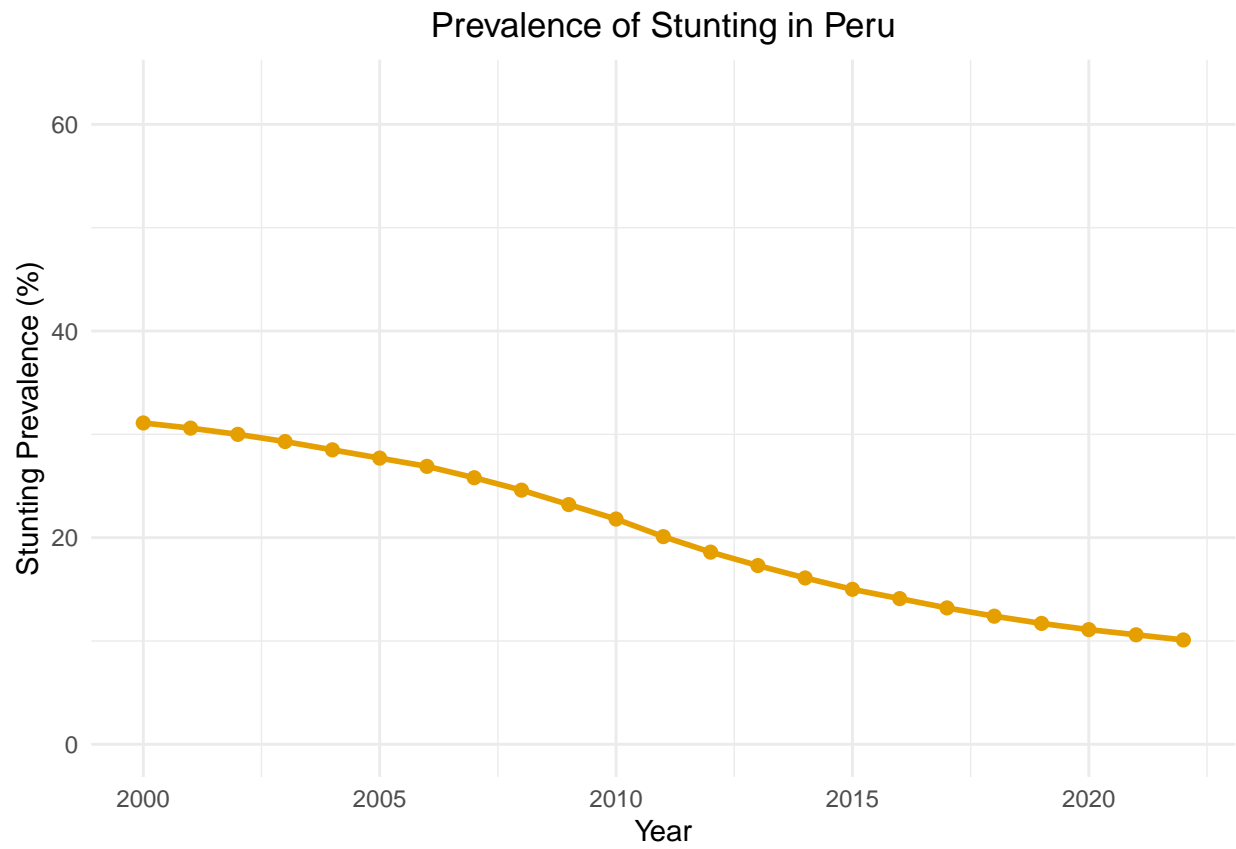


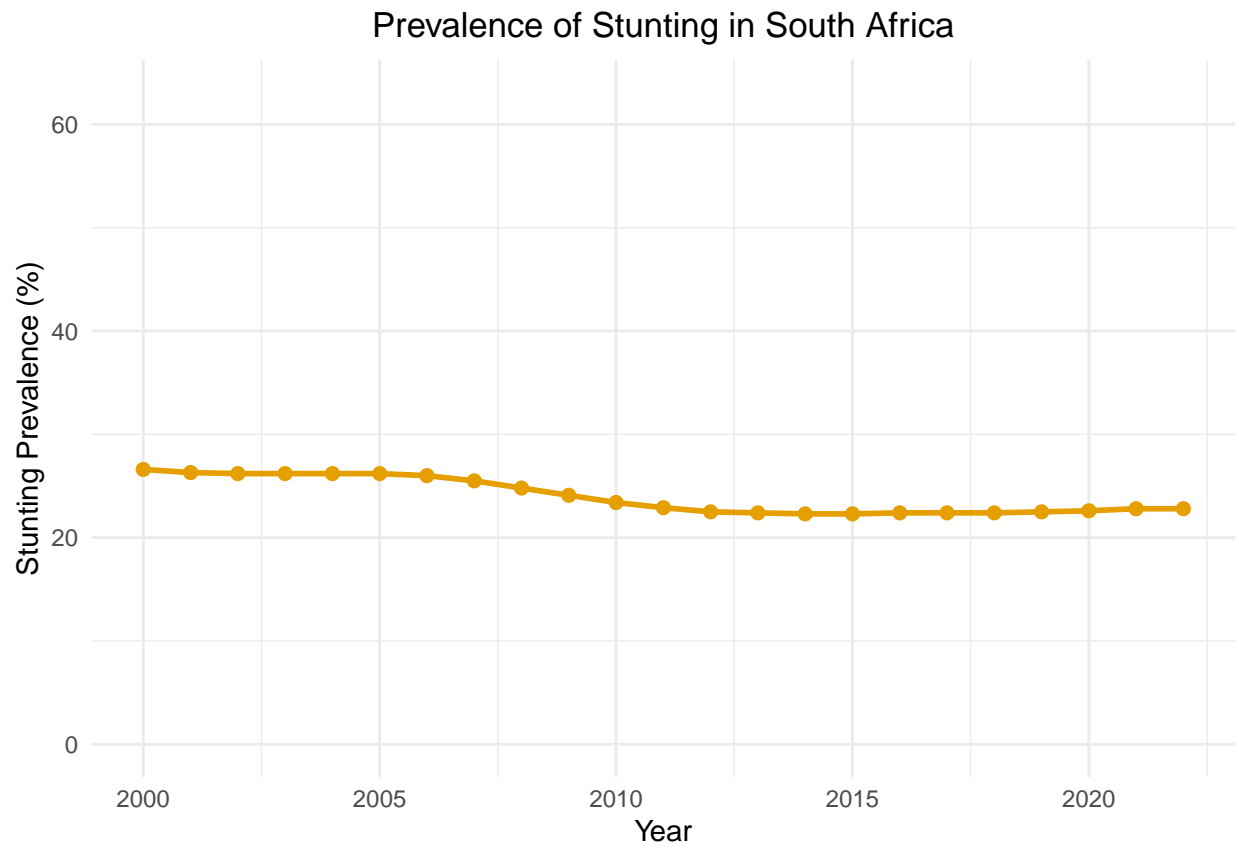


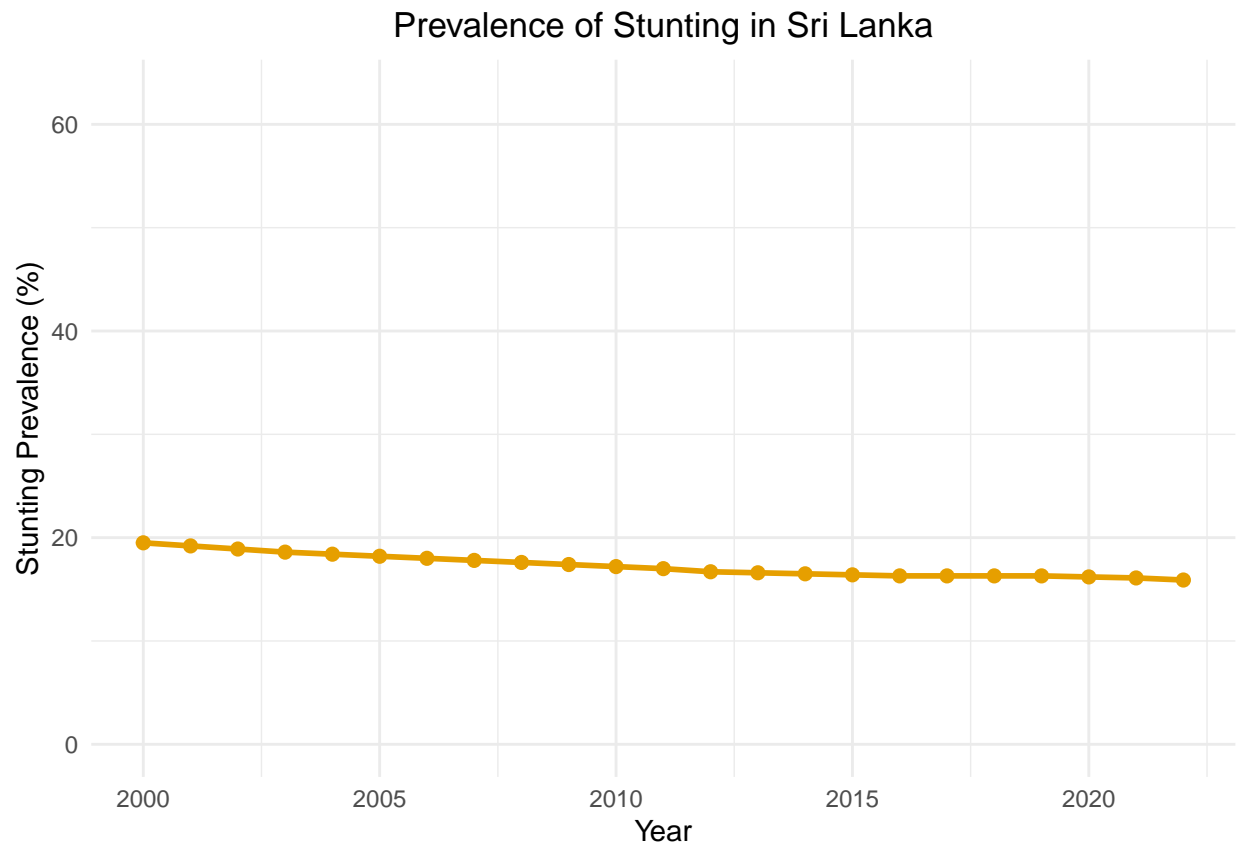




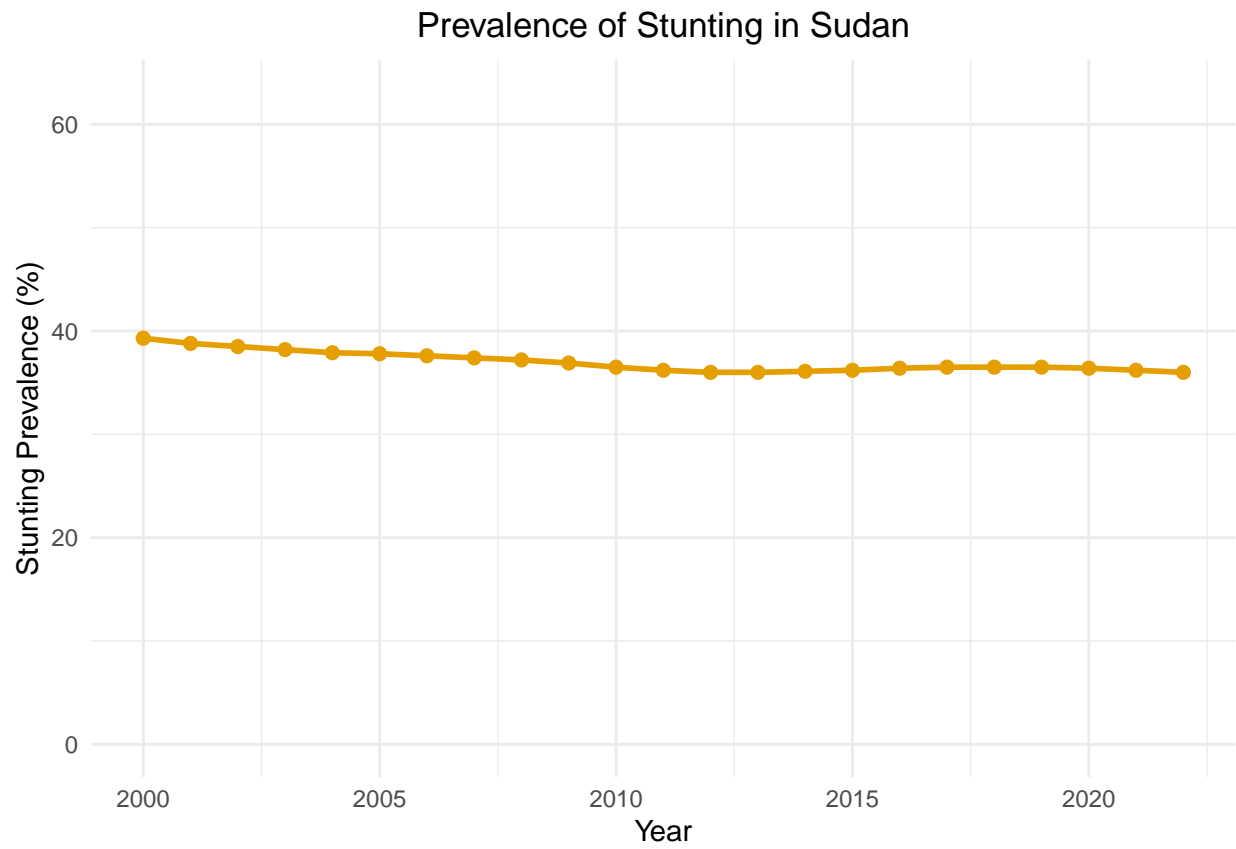


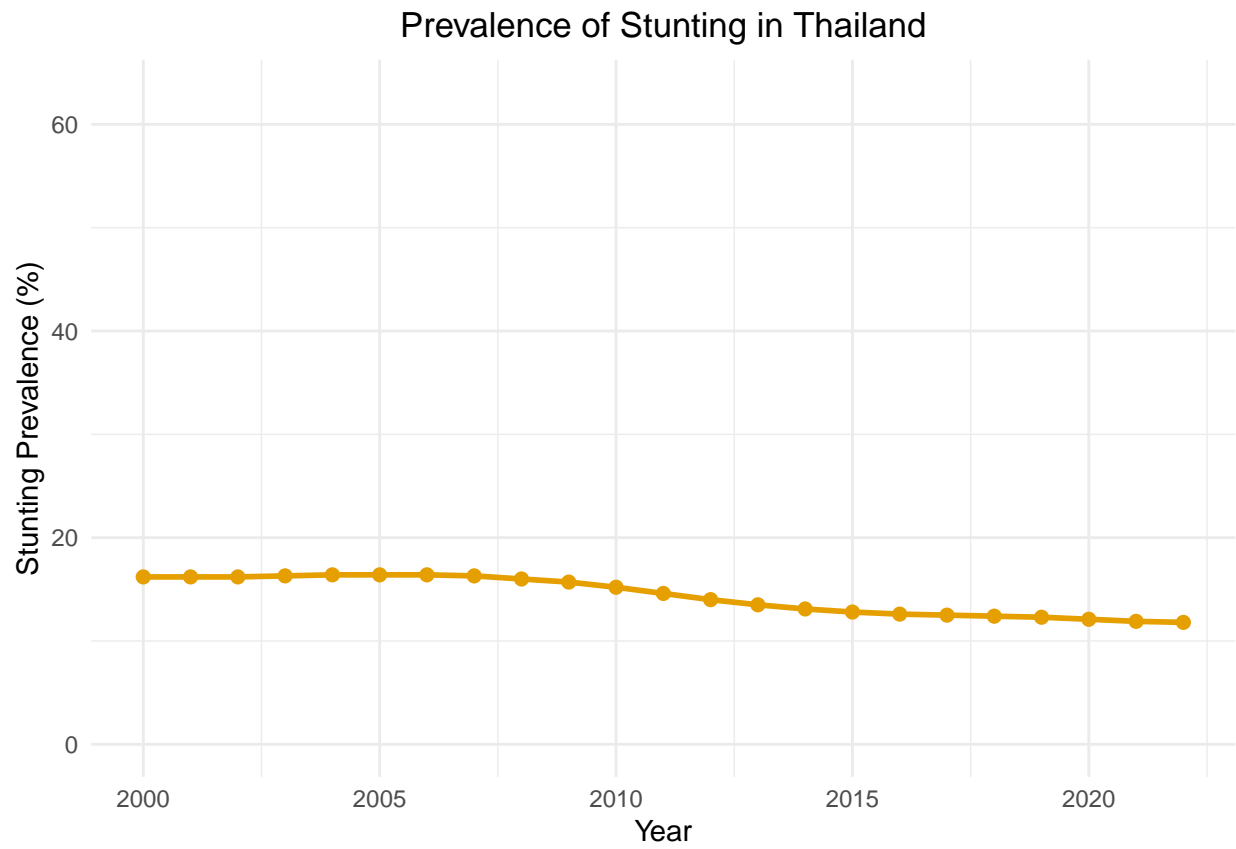


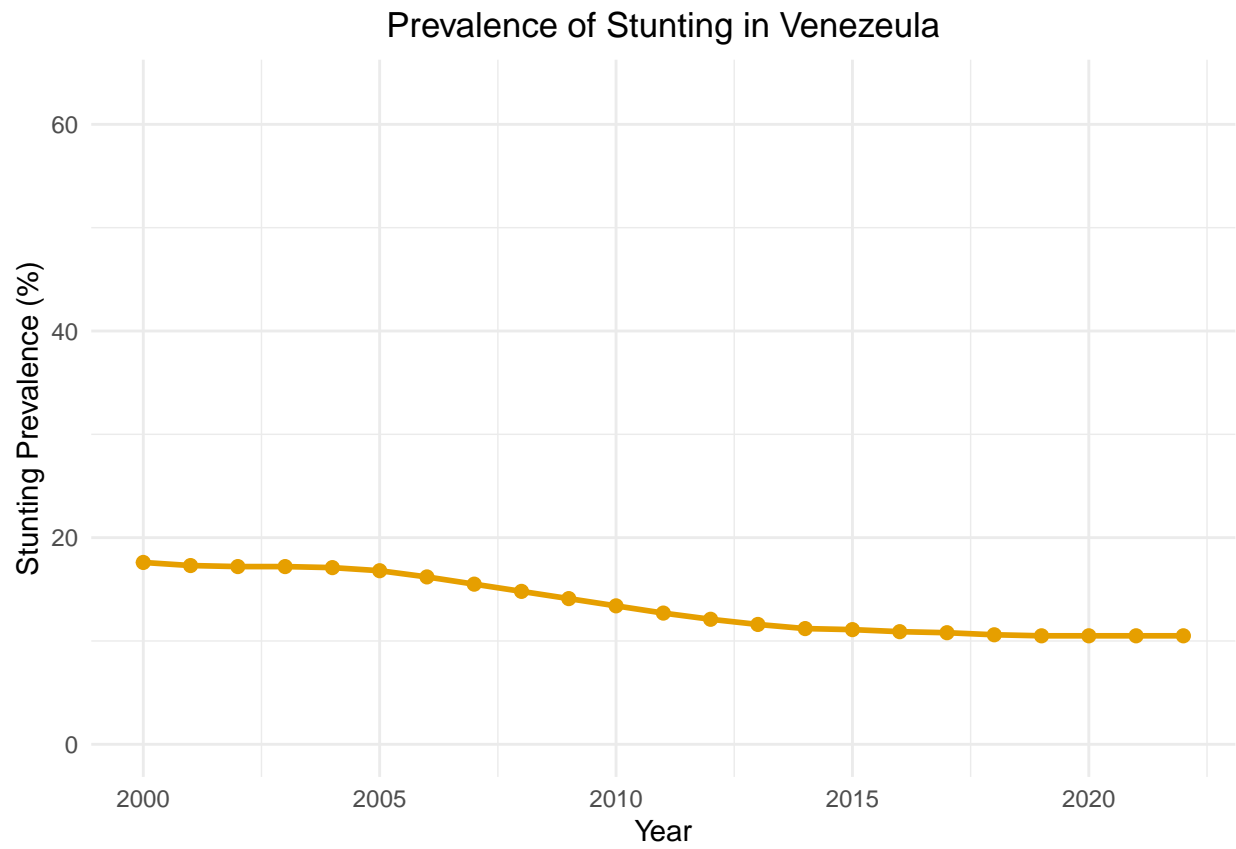


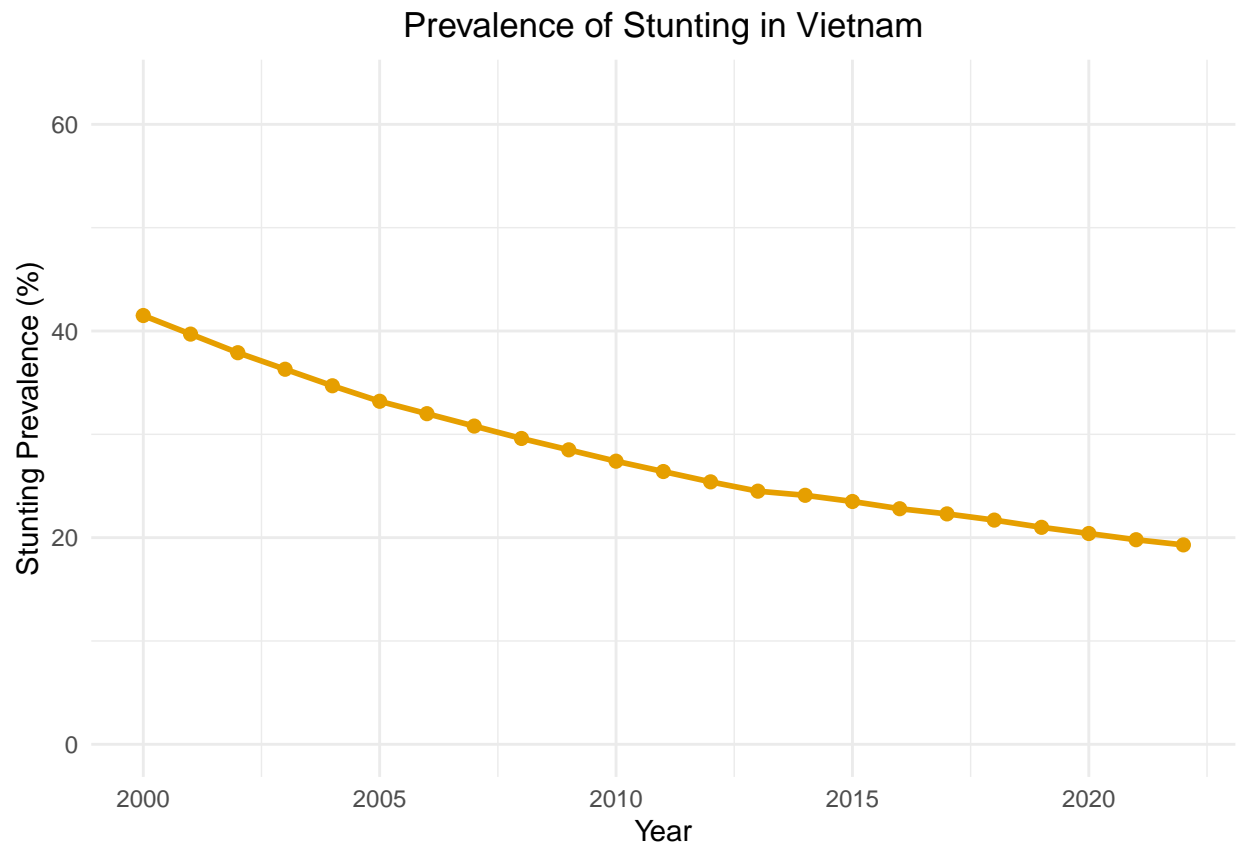


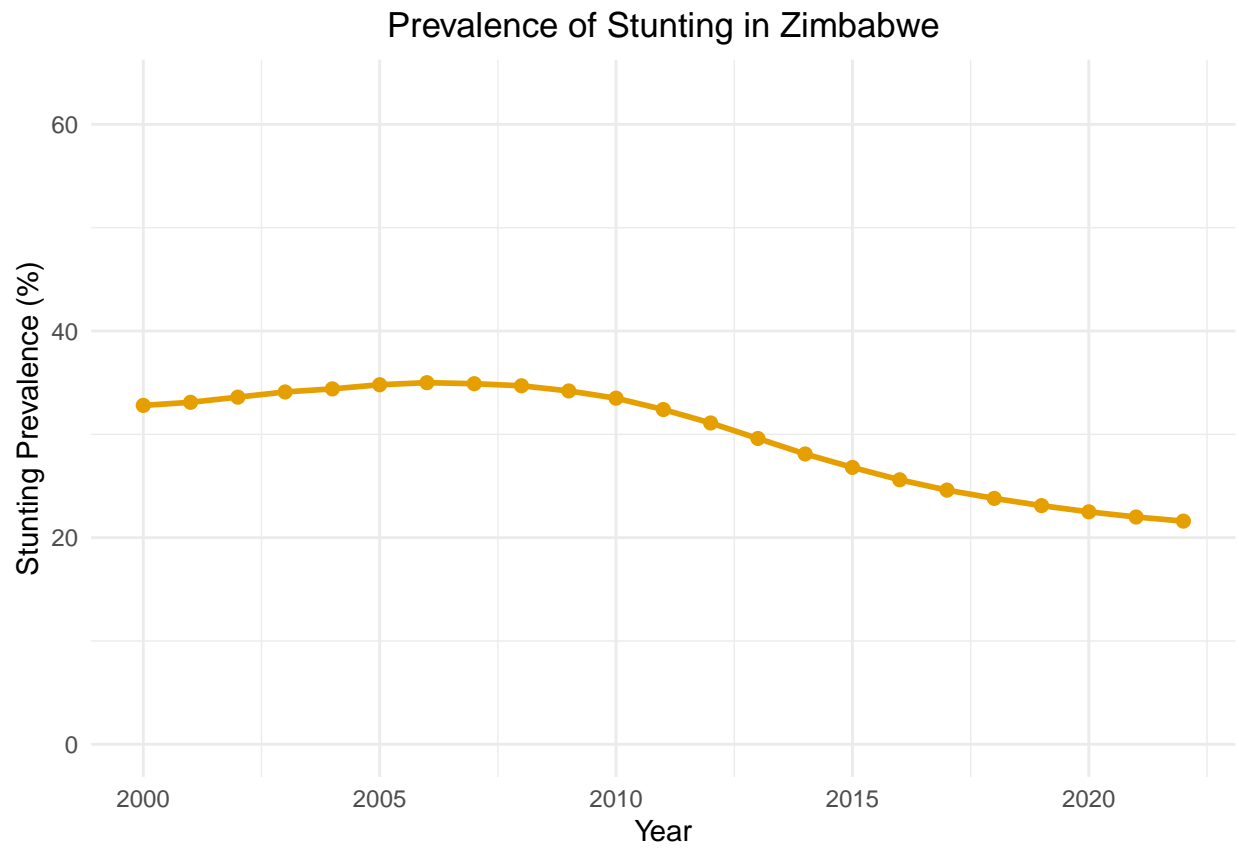


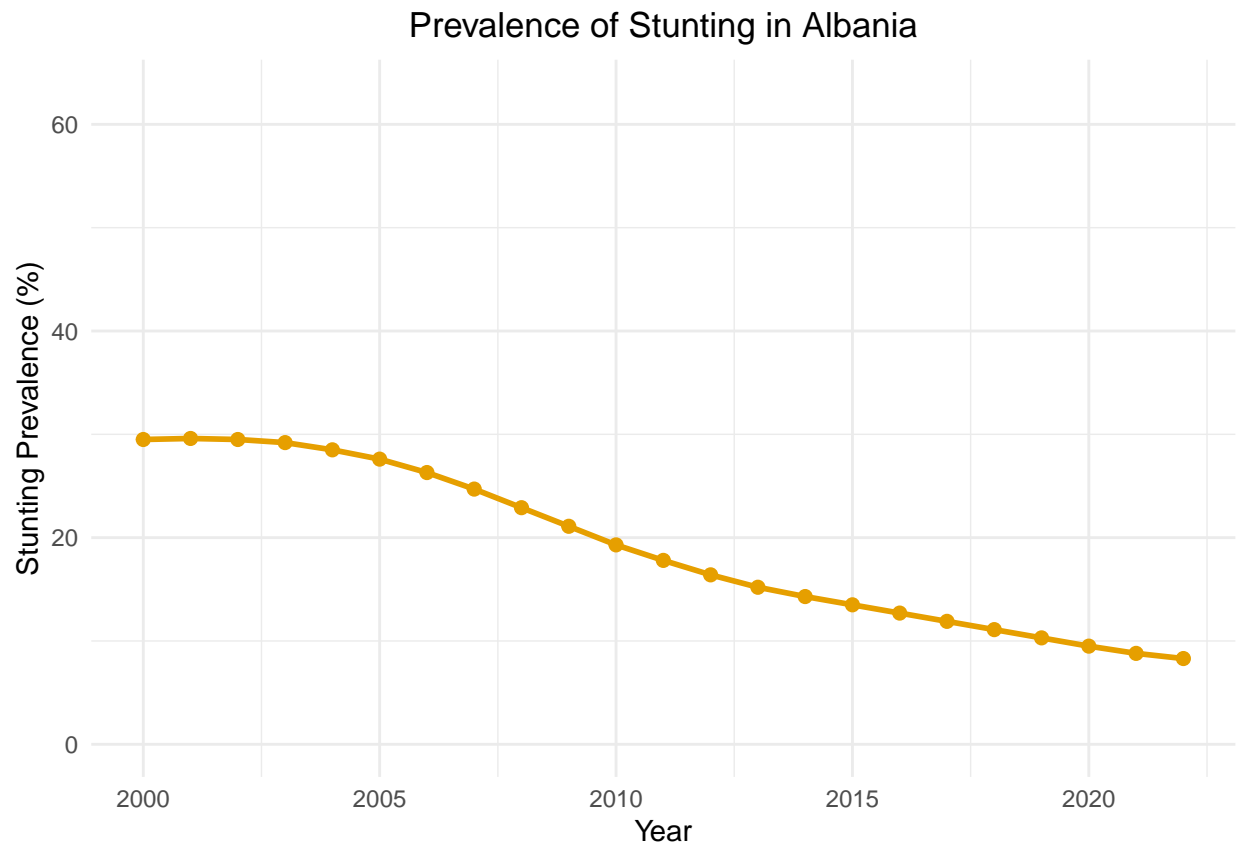


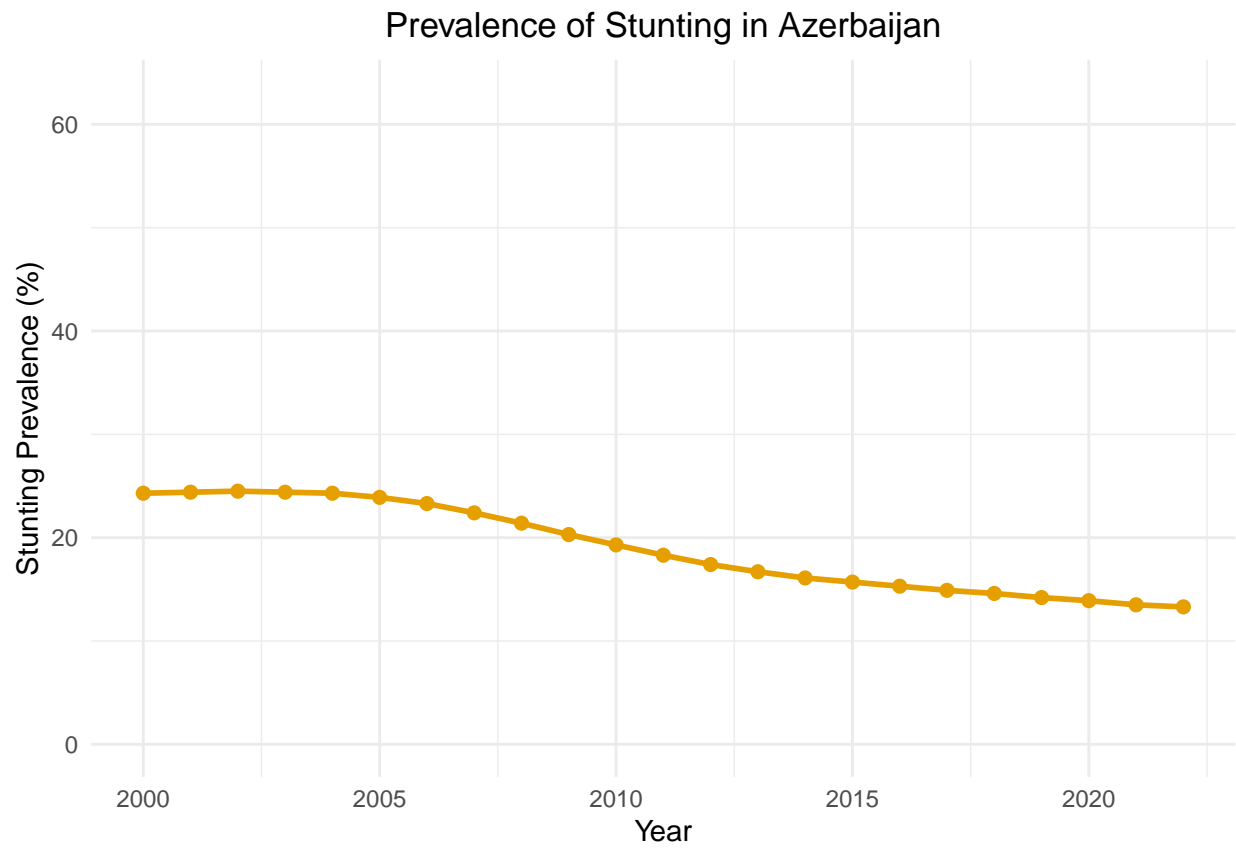


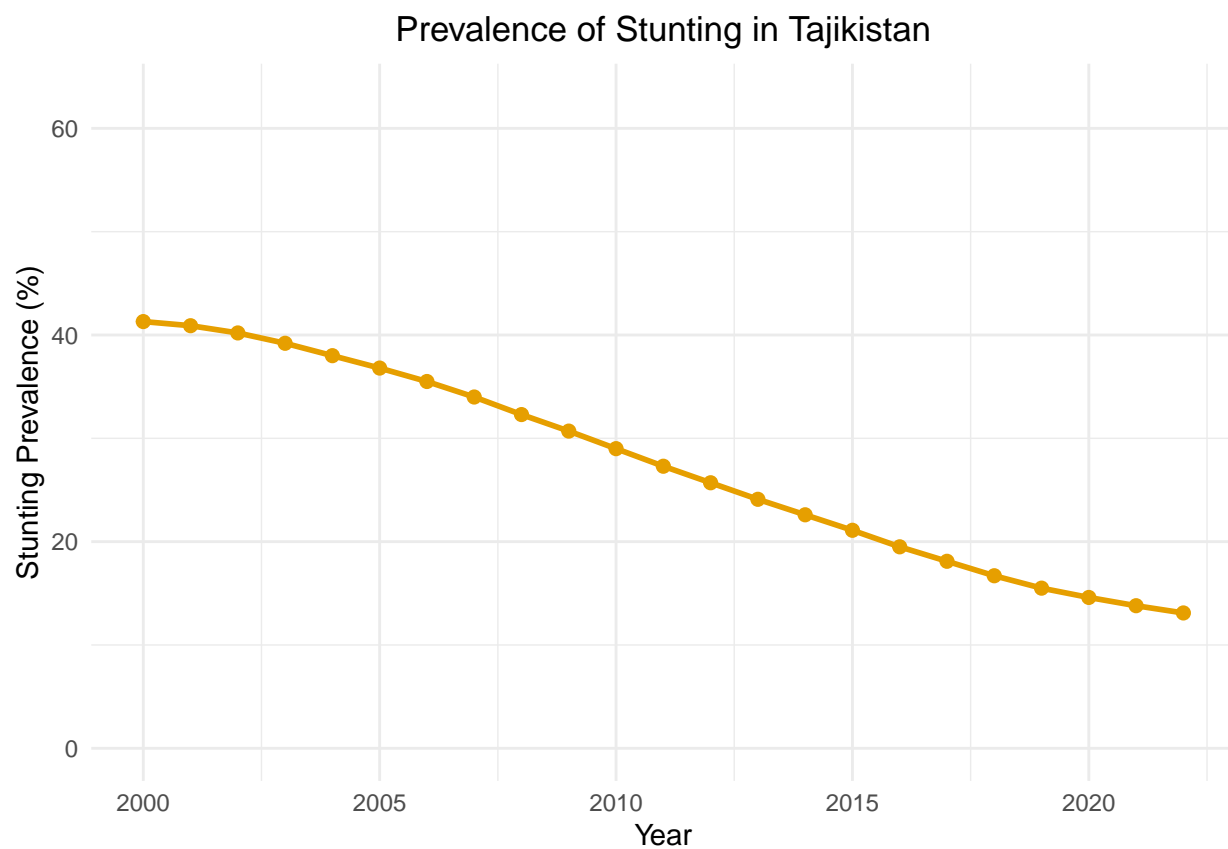




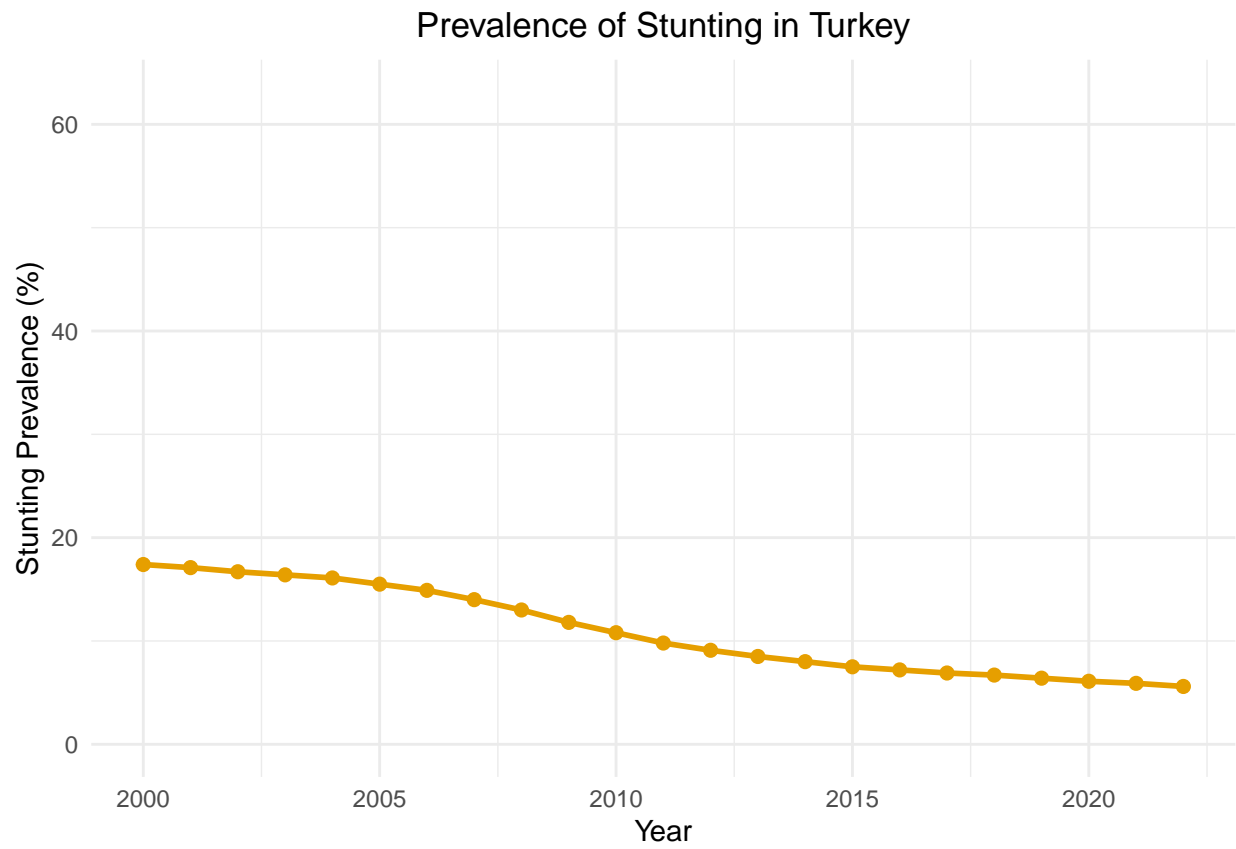


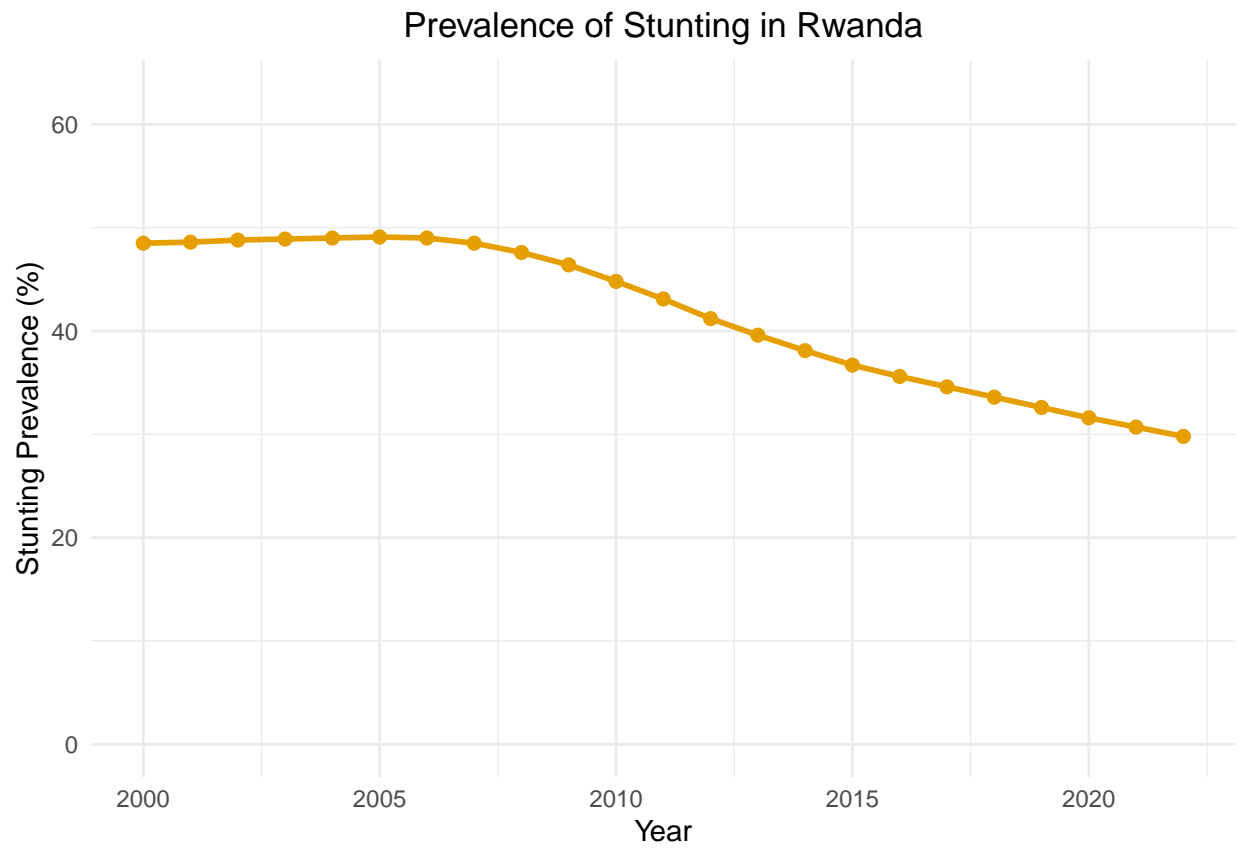


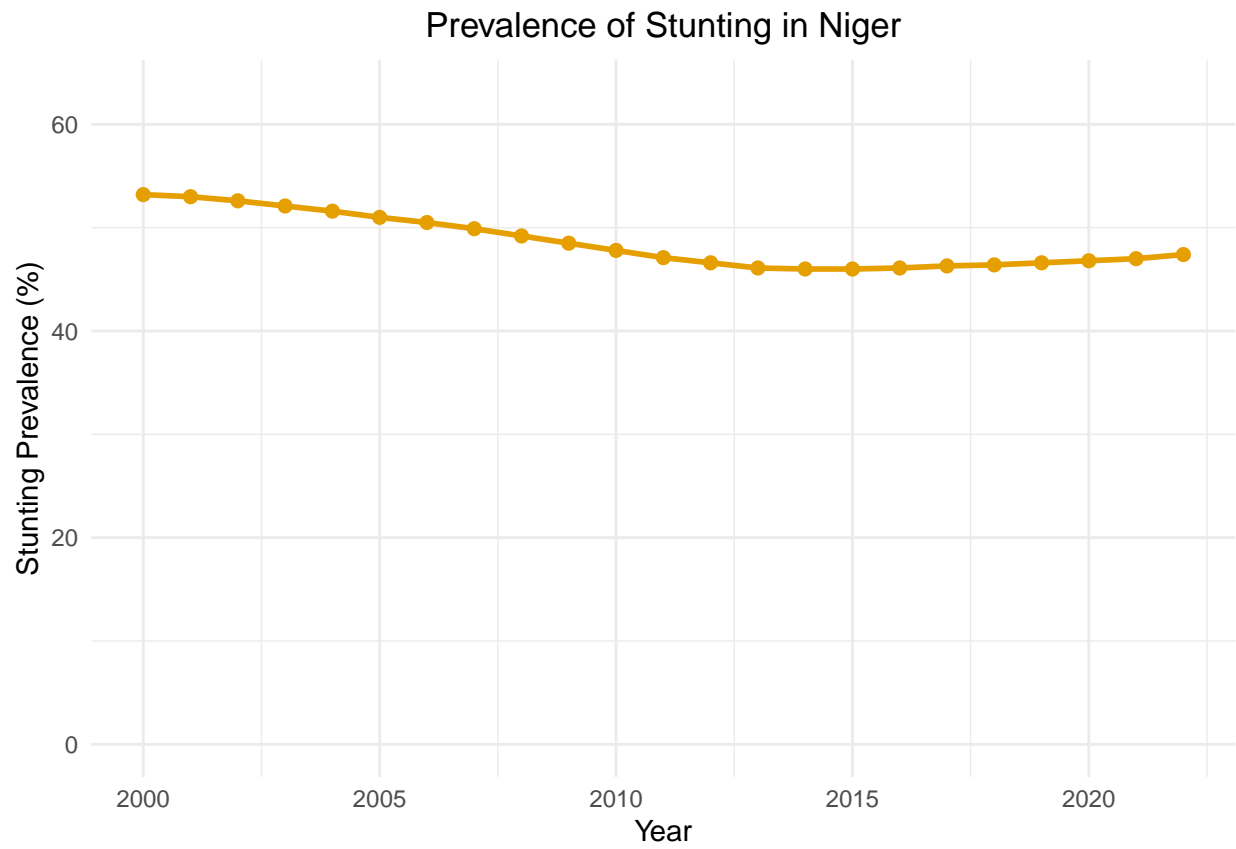


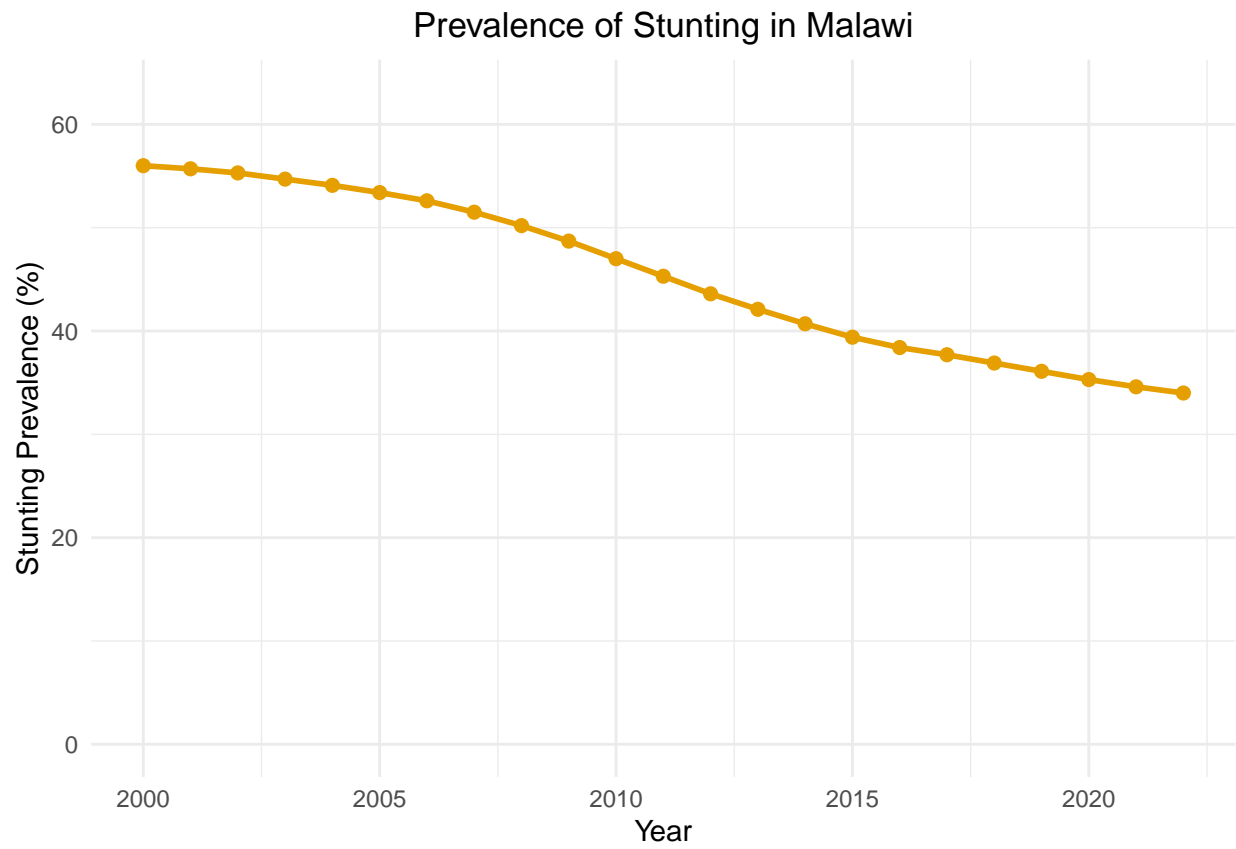


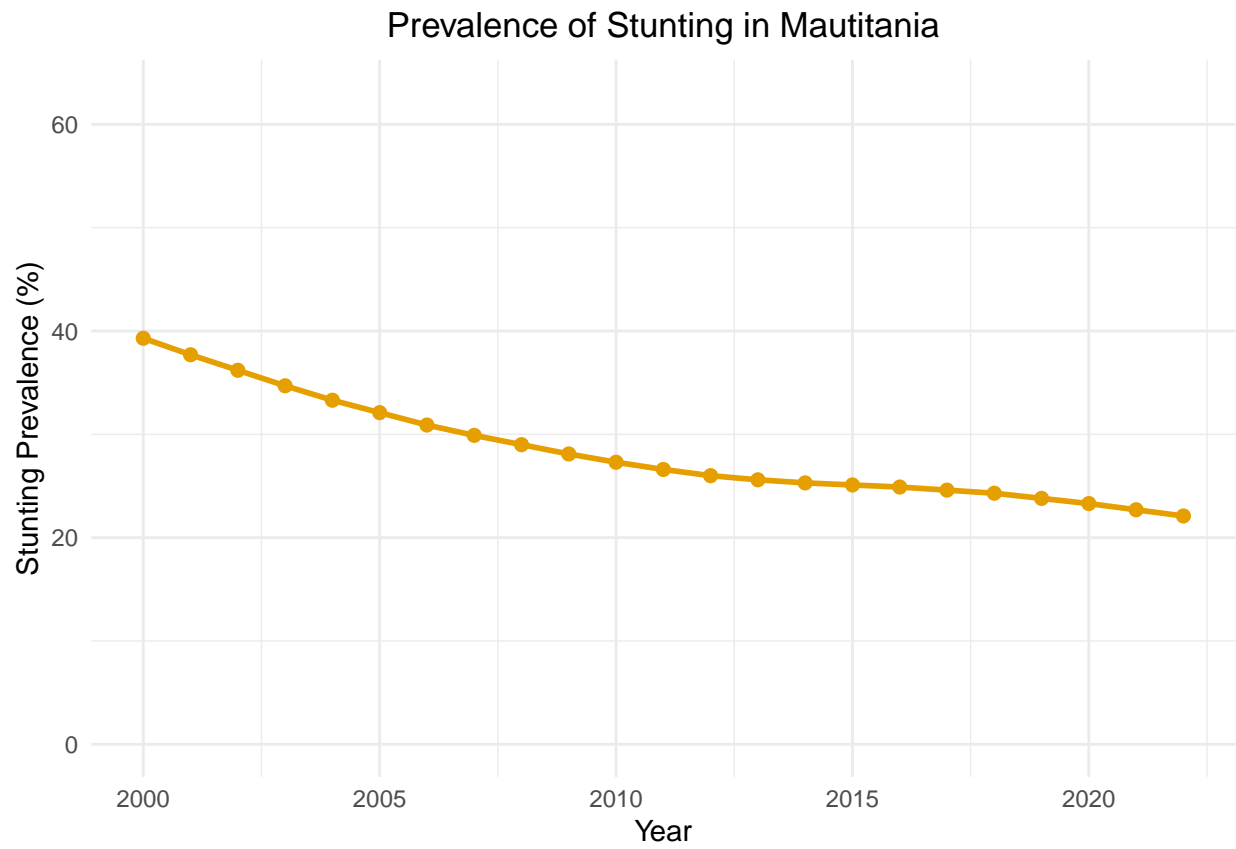


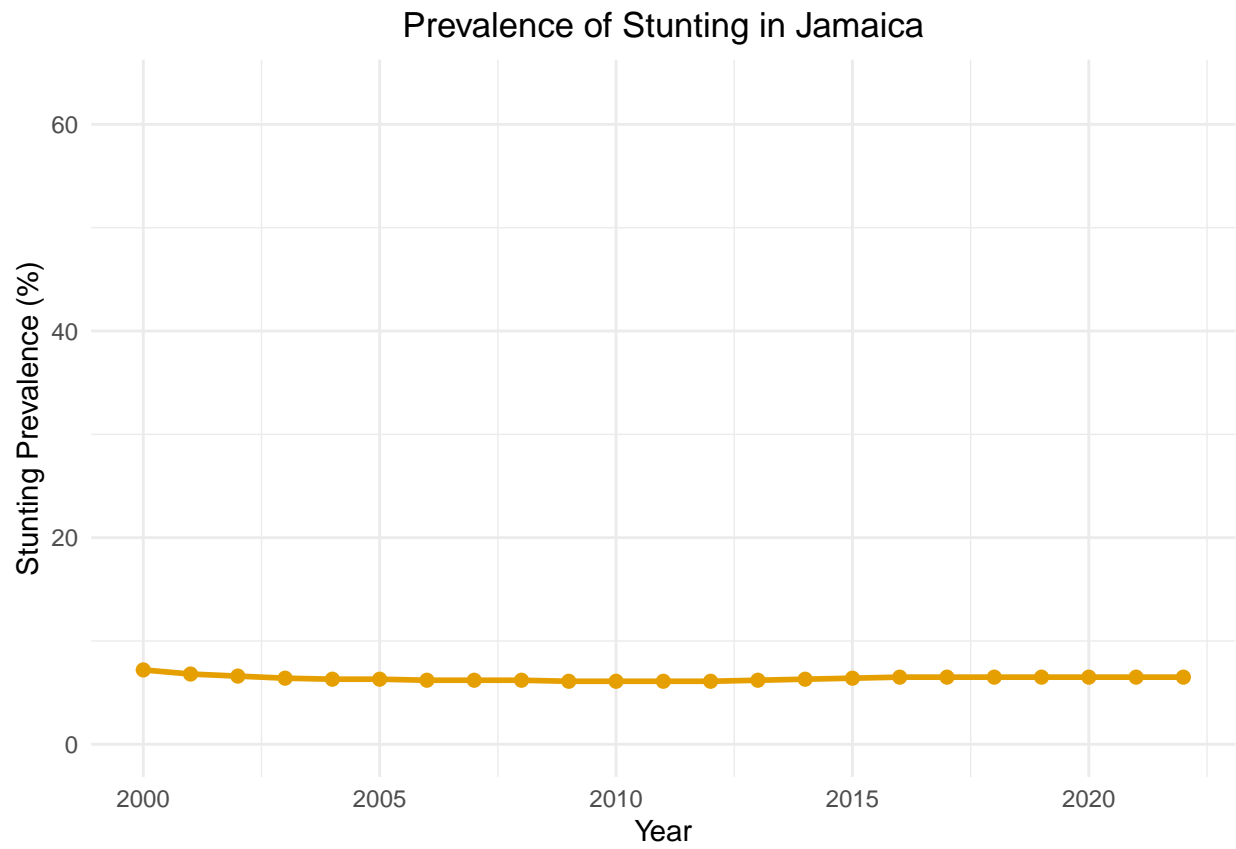


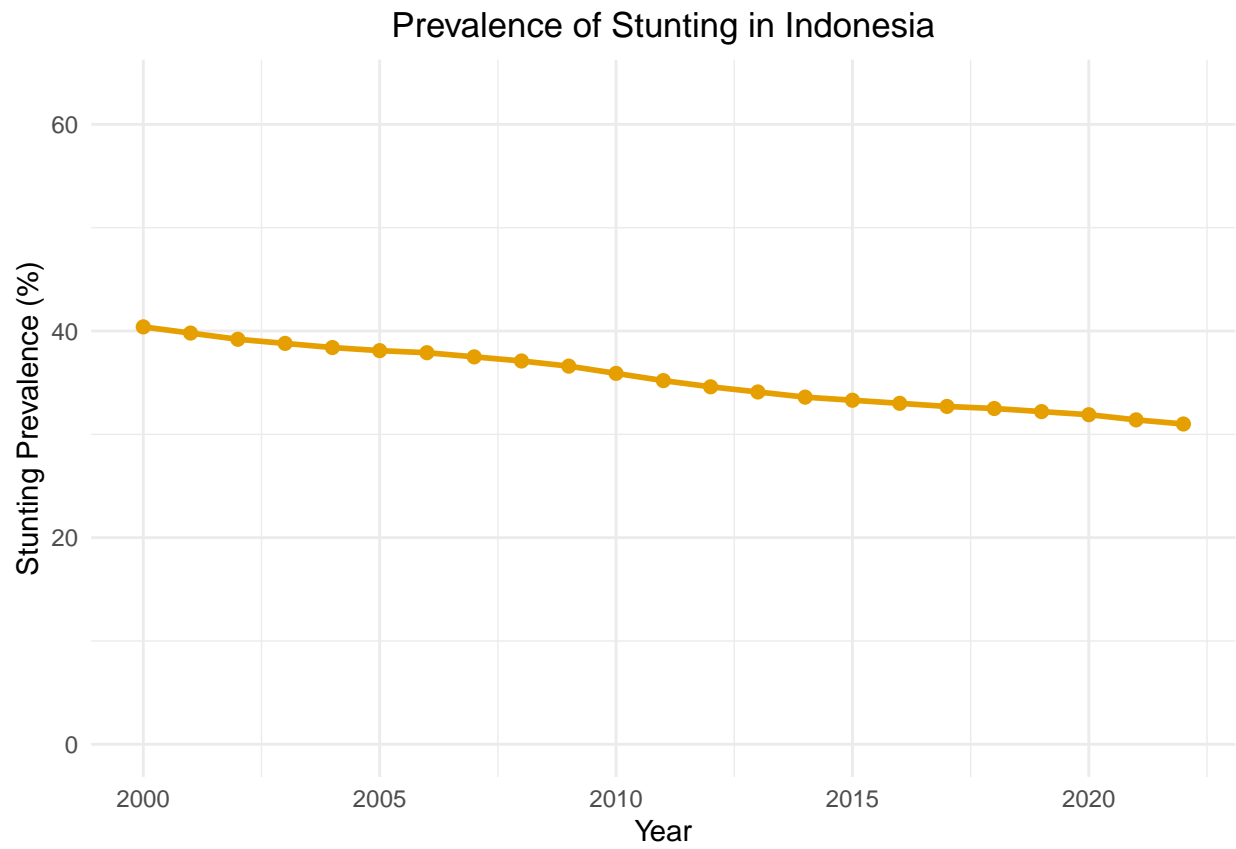


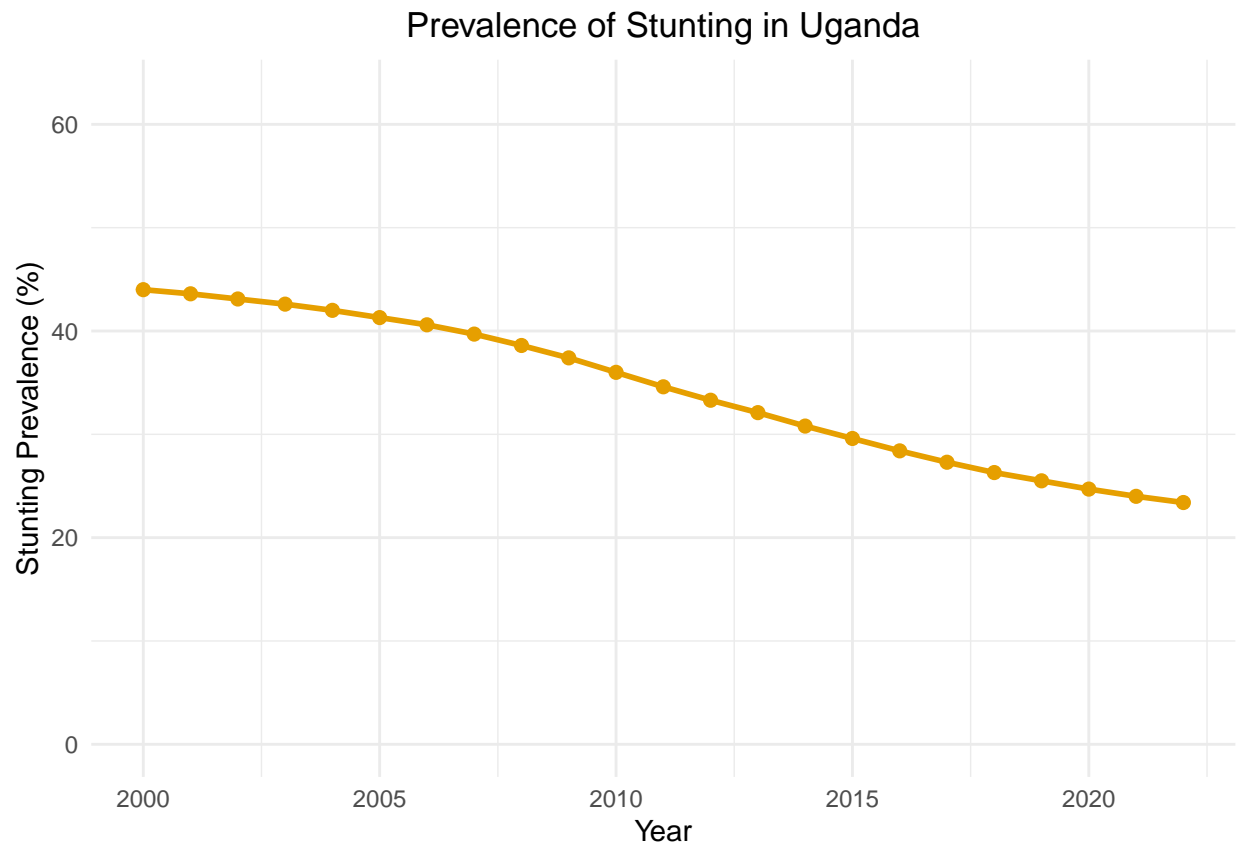




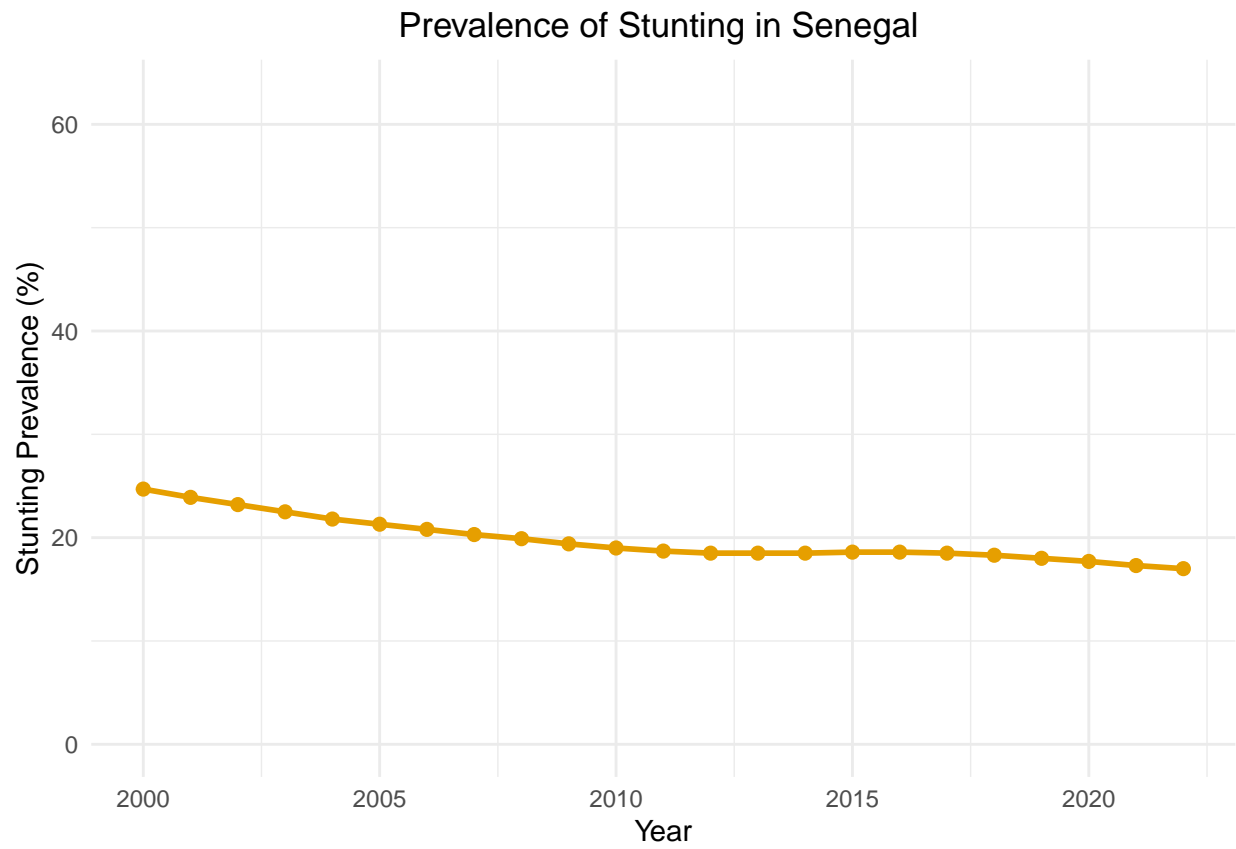


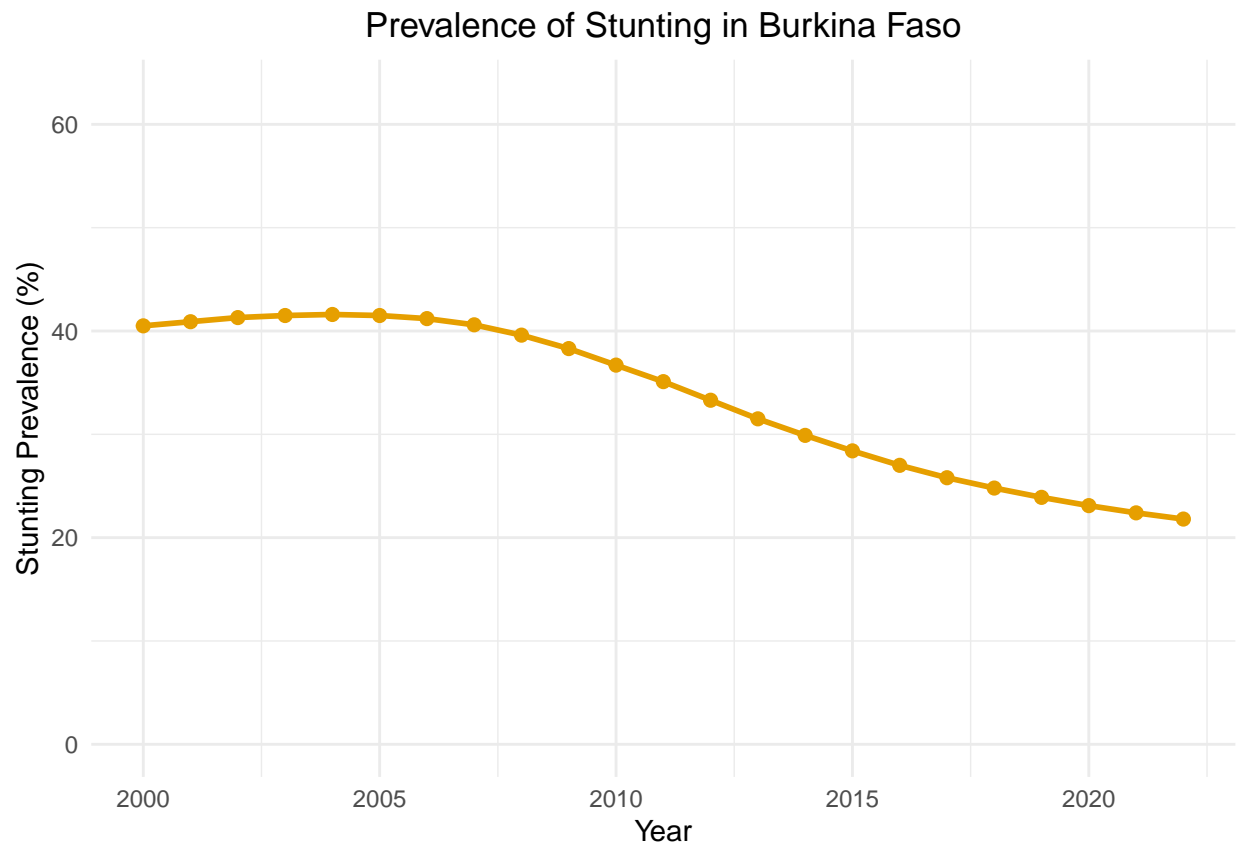


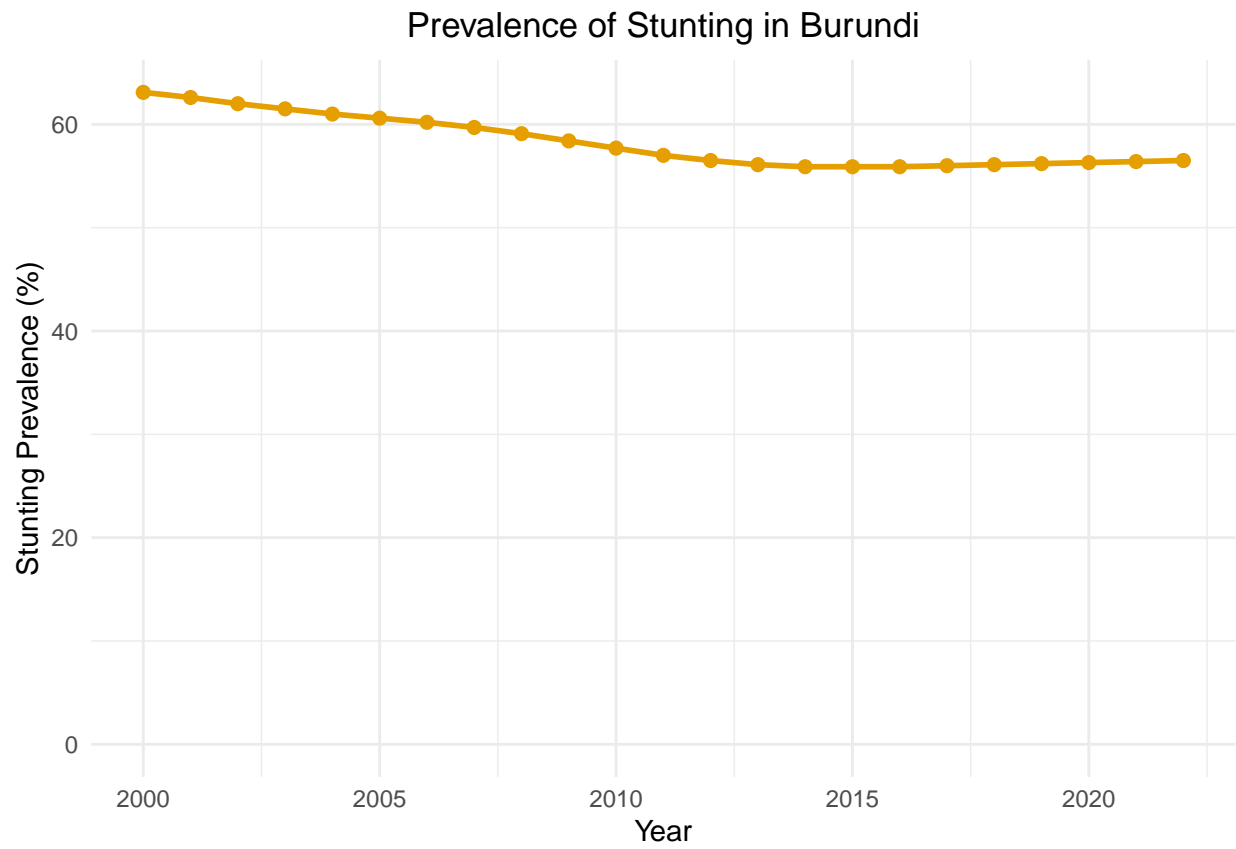


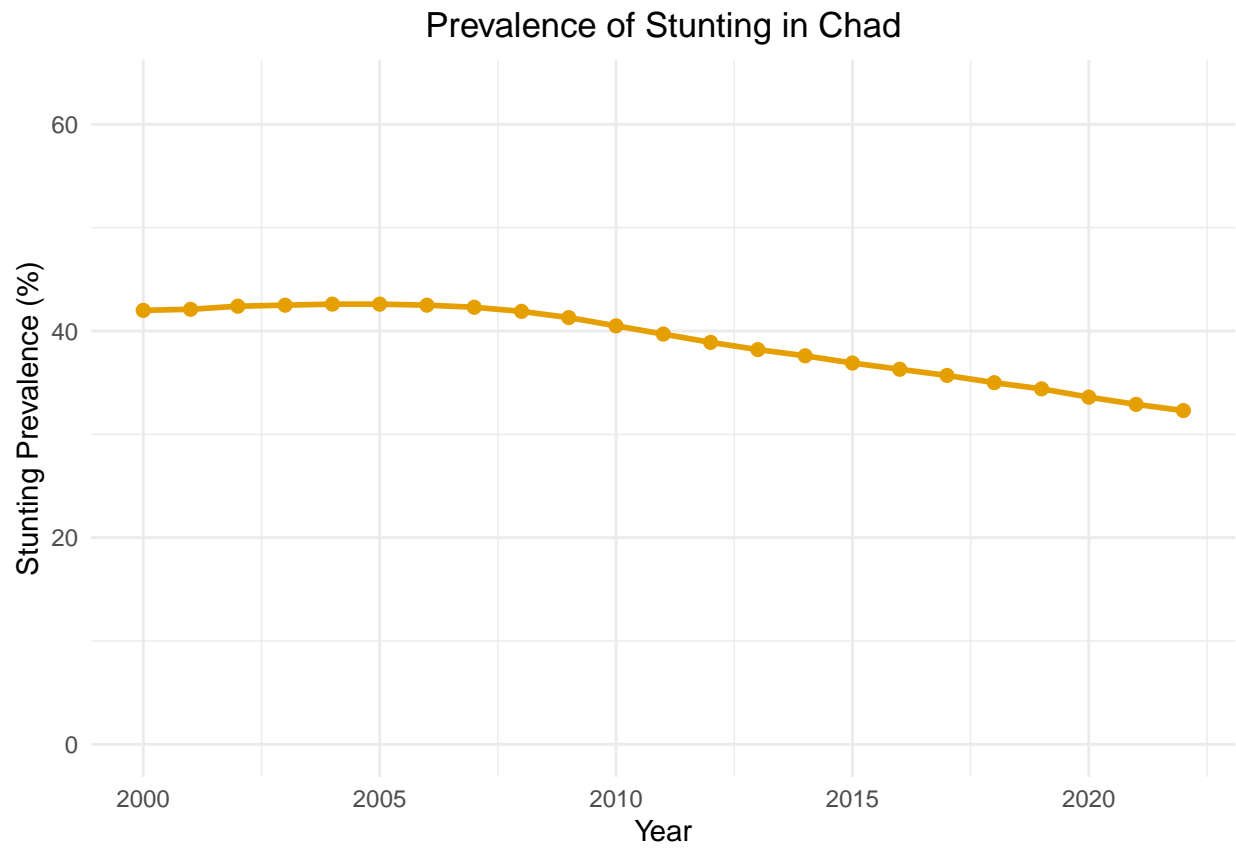


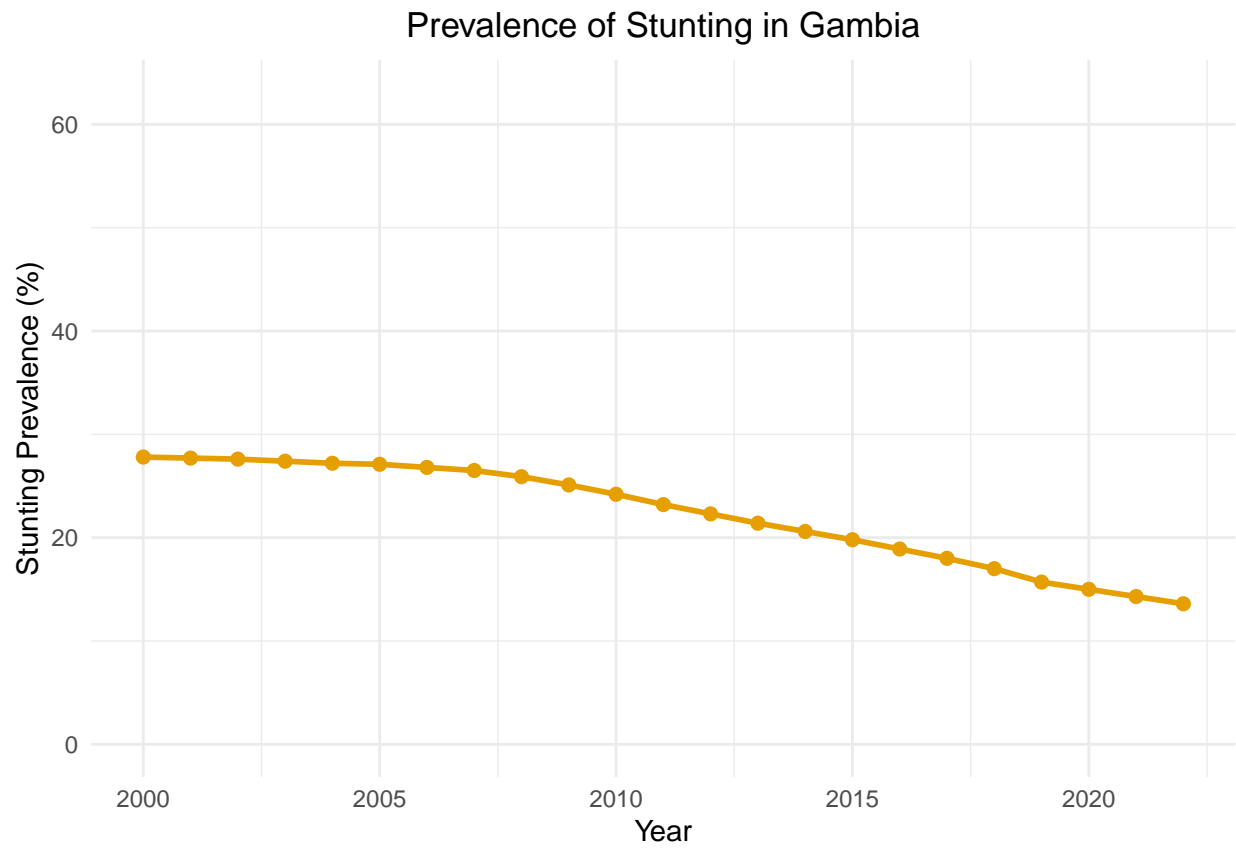


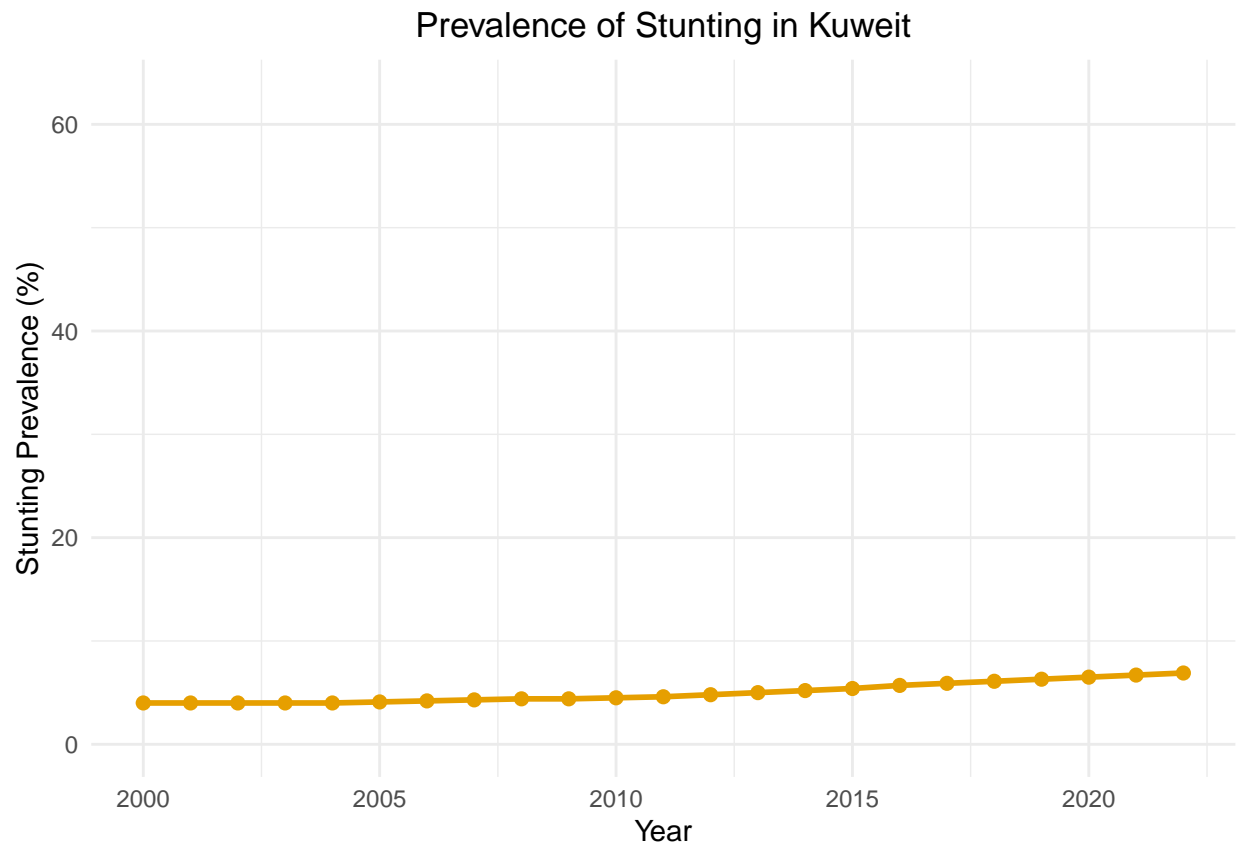


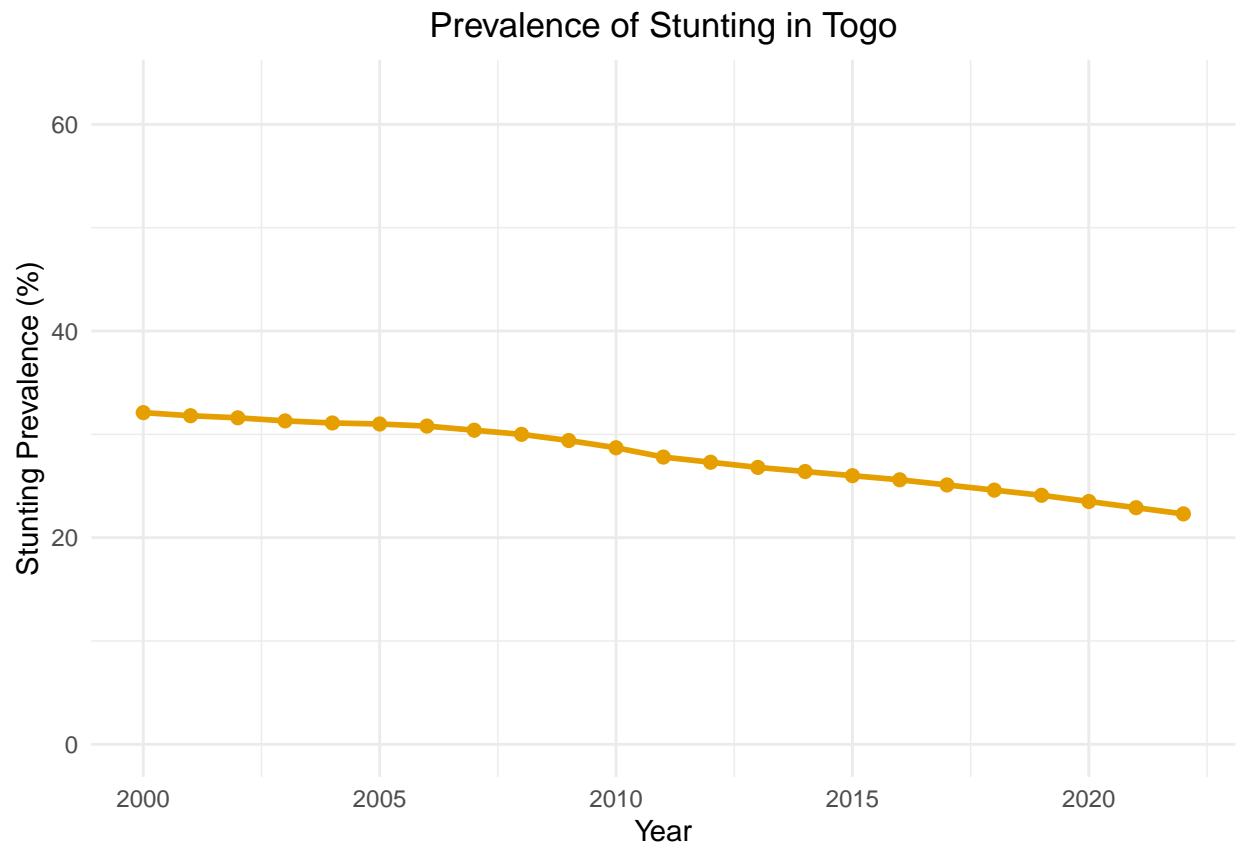


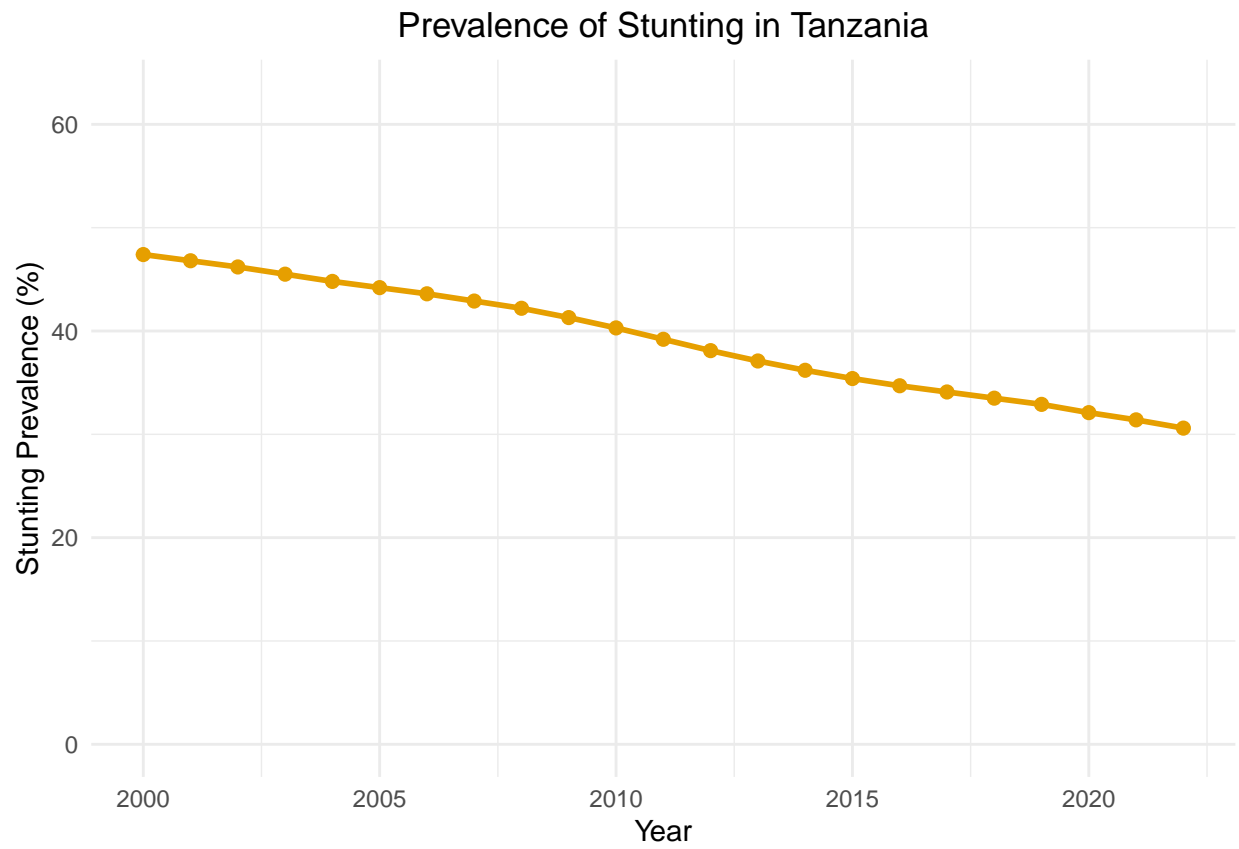






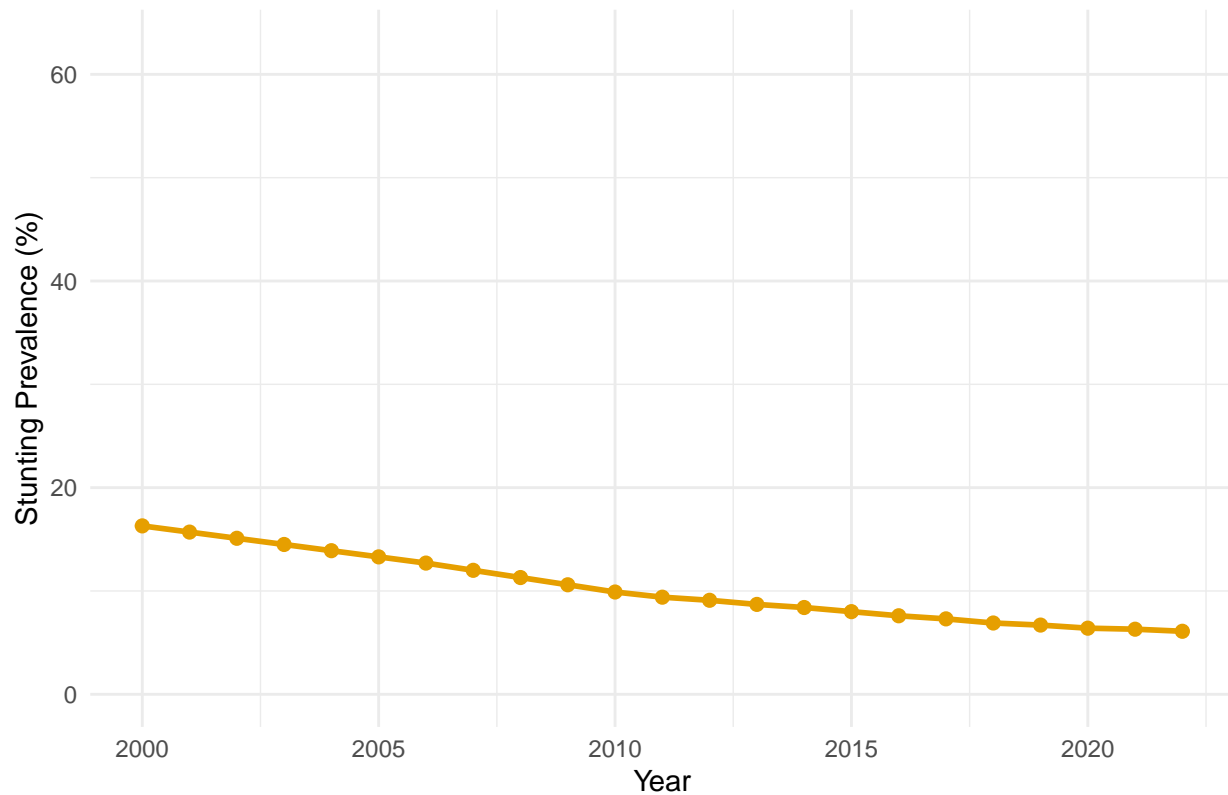








## Prevalence of Stunting in Uruguay



almost all countries have similar prevalence of stunting over time

```
data_filtered_st <- data %>%
  filter(Country %in% selected_countries & !is.na(Stunting_model)) %>%
  group_by(Country) %>%
  filter(Year == max(Year, na.rm = TRUE)) %>%
  select(Country, Year, Stunting_model)
```

```
world_map1 <- map_data("world")
```

```
highlighted_map1 <- world_map1 %>%
  filter(region %in% selected_countries) %>%
  left_join(data_filtered_st, by = c("region" = "Country"))
```

```
highlighted_map <- world_map %>%
  left_join(stunting_data, by = c("region" = "Country"))
```

```
## Warning in left_join(., stunting_data, by = c(region = "Country")): Detected an unexpected many-to-many
## i Row 765 of `x` matches multiple rows in `y`.
## i Row 759 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

```
highlighted_map$Stunting <- as.numeric(highlighted_map$Stunting)
```

```
ggplot() +
  geom_polygon(
```

```

data = world_map,
aes(x = long, y = lat, group = group),
fill = "lightgray", color = "white", linewidth = 0.1
) +
geom_polygon(
  data = highlighted_map,
  aes(x = long, y = lat, group = group, fill = Stunting),
  color = "black", linewidth = 0.2
) +
scale_fill_viridis(
  name = "Stunting (%)",
  option = "plasma", direction = -1, na.value = "grey90",
  limits = c(0, max(highlighted_map$Stunting, na.rm = TRUE))
) +
coord_fixed(ratio = 1.3) +
theme_minimal() +
labs(title = "Stunting Percentage by Country", x = "Longitude", y = "Latitude") +
theme(
  legend.position = "bottom",
  legend.title = element_text(size = 10),
  legend.text = element_text(size = 8),
  axis.text = element_blank(),
  axis.ticks = element_blank(),
  panel.grid = element_blank()
)

```

Stunting Percentage by Country

