PPHA 37040 Problem Set 1: Structural Transformation Training

Nat Larsen, nclarsen

2025-01-17

Contents

Analysis of the Philippines

```
# Install Necessary Packages

# Install dplyr
if (!requireNamespace("dplyr", quietly = TRUE)) {
  install.packages("dplyr")
}

# Install ggplot2
if (!requireNamespace("ggplot2", quietly = TRUE)) {
  install.packages("ggplot2")
}

# Install dendextend
if (!requireNamespace("dendextend", quietly = TRUE)) {
  install.packages("dendextend")
}
```

```r
# Install strucchange
if (!requireNamespace("strucchange", quietly = TRUE)) {
  install.packages("strucchange")
}

# Install haven package to read the .dta file
if (!requireNamespace("haven", quietly = TRUE)) {
  install.packages("haven")
}

# Install tidyr
if (!requireNamespace("tidyr", quietly = TRUE)) {
  install.packages("tidyr")
}

# Install scales
if (!requireNamespace("scales", quietly = TRUE)) {
  install.packages("scales")
}

# Install ggrepel
if (!requireNamespace("ggrepel", quietly = TRUE)) {
  install.packages("ggrepel")
}
```

Data Preparation

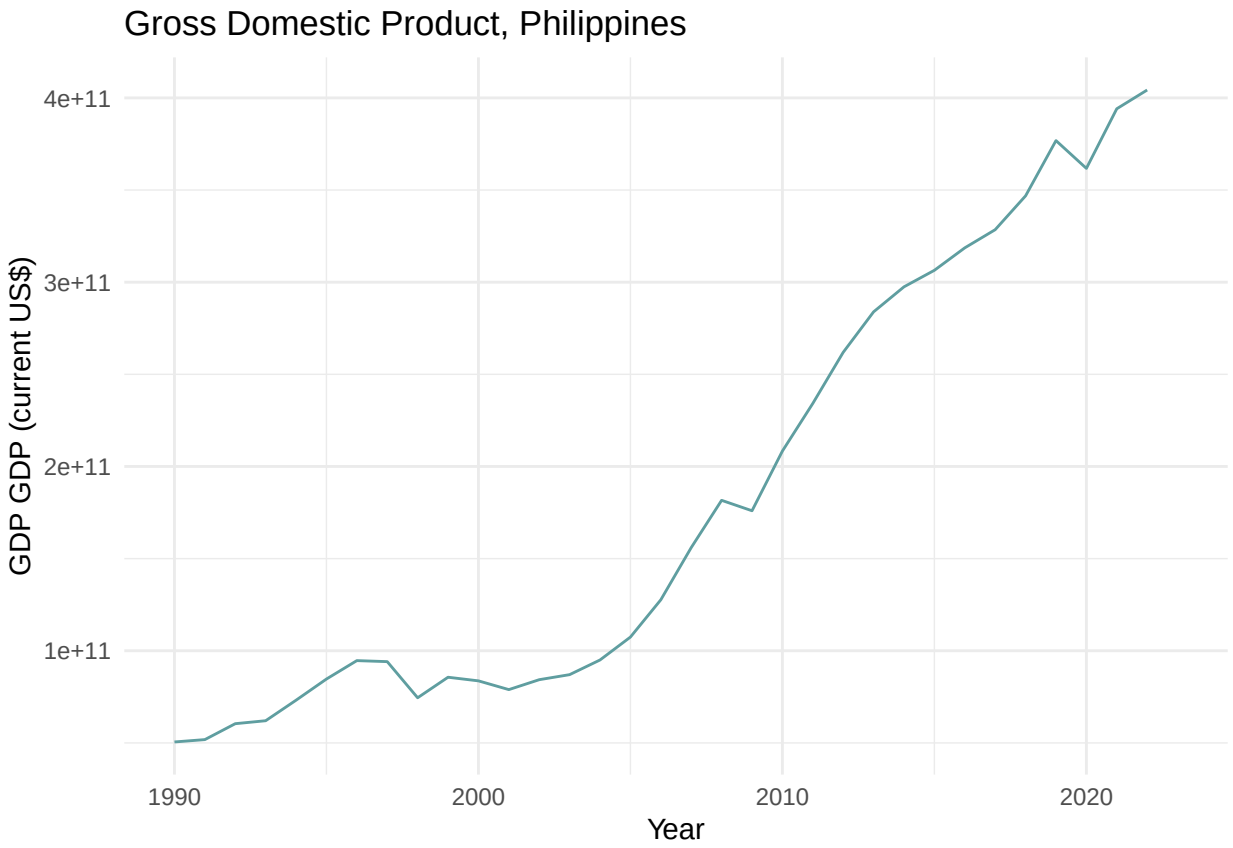
Load and Prepare Data

```r
# Read the .dta file
# .dta file must be in the same working directory as this .Rmd
data <- read_dta("/Users/natlarsen/Downloads/Master Data/glmacro_master_alldata.dta")
```

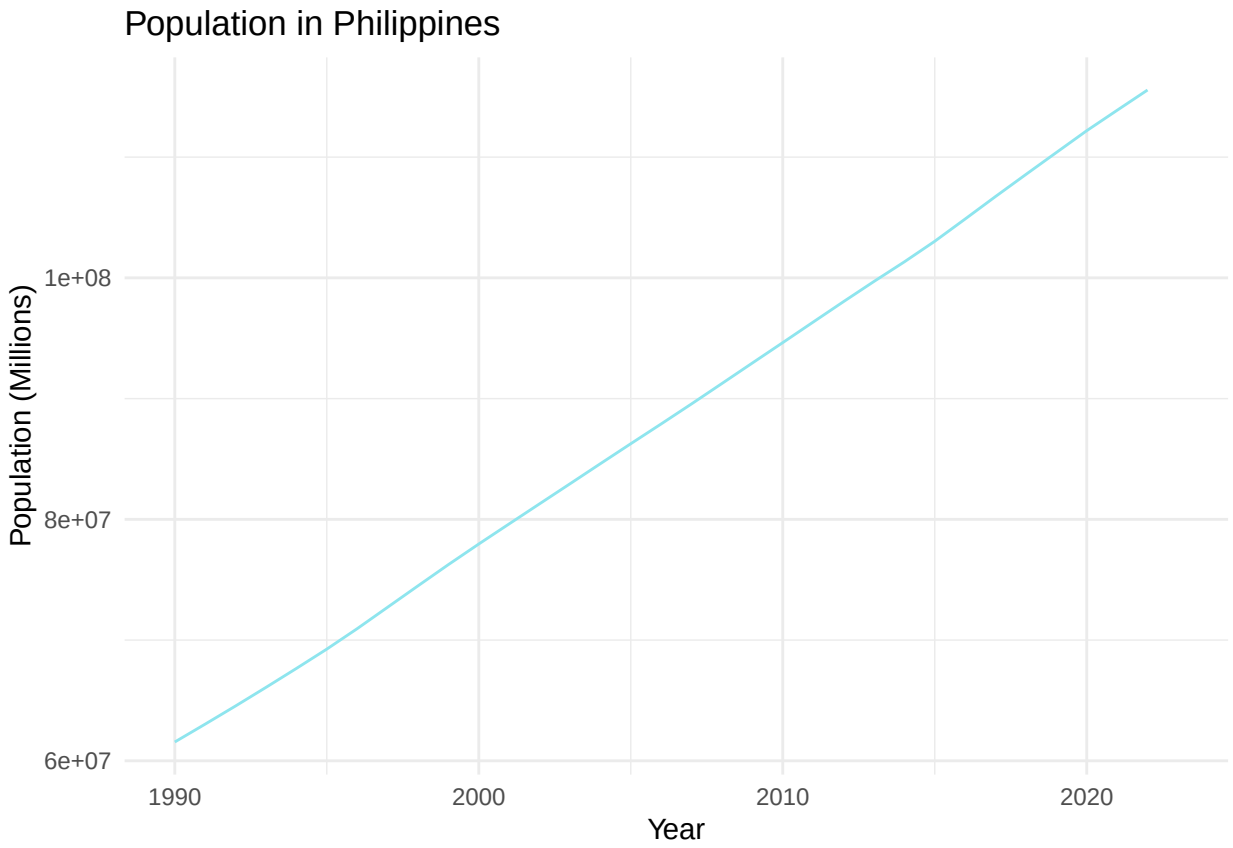
```r
# Select the Philippines as the country
country <- "Philippines"

# Add a dummy column for a chosen country
data <- data %>%
  mutate(dummy_chosen_country = ifelse(wb_countryname == country, 1, 0))
```

Question 1: History of Population and GDP (10 points)

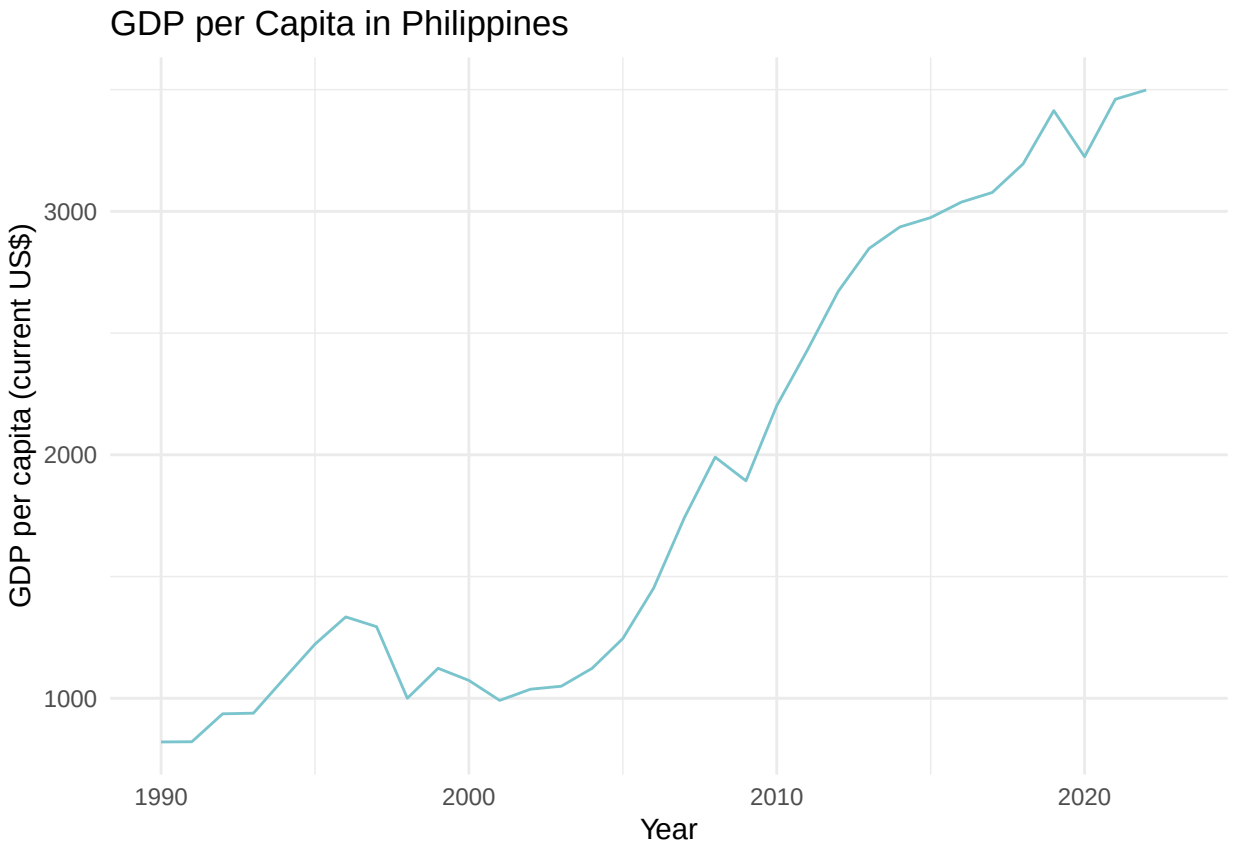
```r
# Filter the dataset and plot the GDP
ggplot(data %>% filter(wb_countryname == country & year >= 1990 & year <= 2023),
       aes(x = year, y = wdi_ny_gdp_mktp_cd)) +
  geom_line(color = "cadetblue") +
  labs(title = paste("Gross Domestic Product,", country),
       y = "GDP GDP (current US$)",
       x = "Year") +
  theme_minimal()
```




```
# Plot Population Time Series
ggplot(data %>% filter(wb_countryname == country & year >= 1990 & year <= 2023),
       aes(x = year, y = wdi_sp_pop_totl)) +
  geom_line(color = "cadetblue2") +
  labs(title = paste("Population in", country),
       y = "Population (Millions)",
       x = "Year") +
  theme_minimal()
```



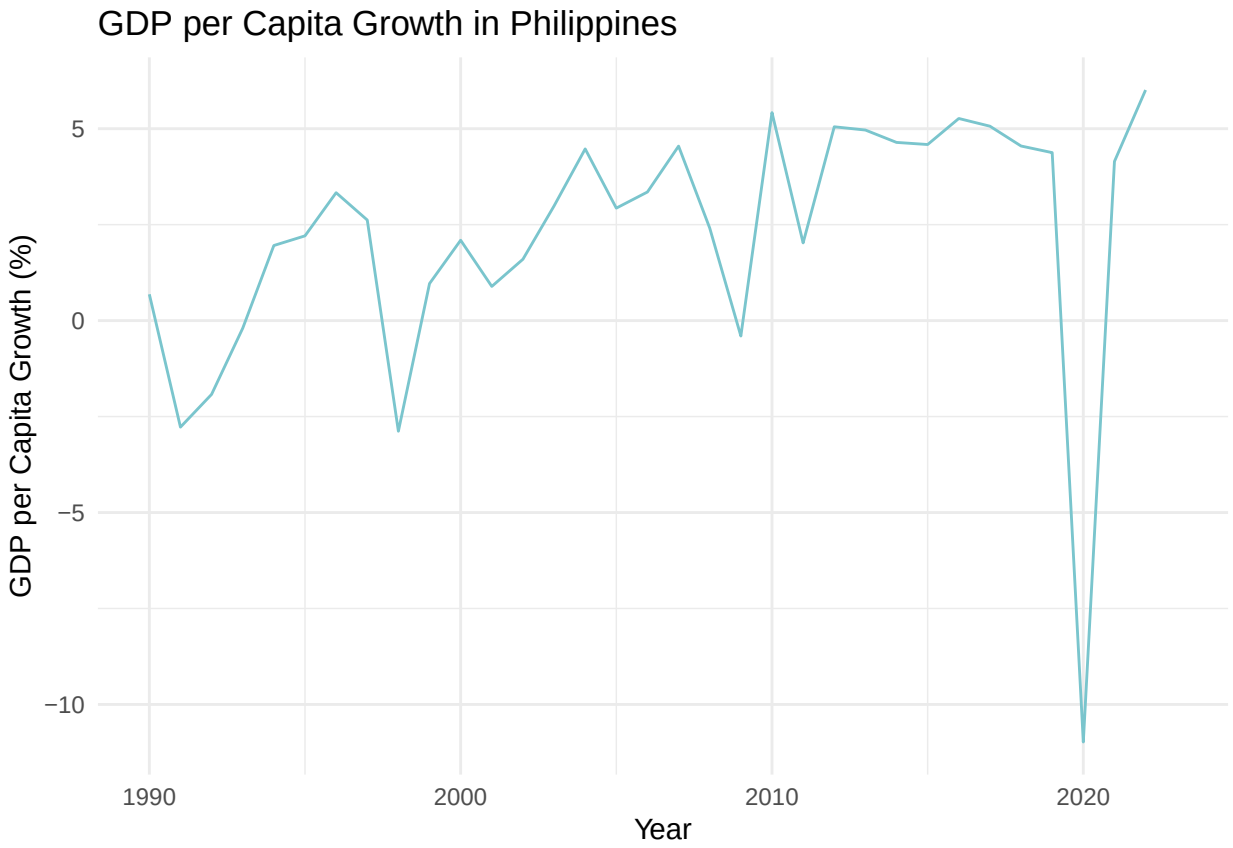

```
# Plot GDP per capita Time Series
ggplot(data %>% filter(wb_countryname == country & year >= 1990 & year <= 2023),
       aes(x = year, y = wdi_ny_gdp_pcap_cd)) +
  geom_line(color = "cadetblue3") +
  labs(title = paste("GDP per Capita in", country),
       y = "GDP per capita (current US$)",
       x = "Year") +
  theme_minimal()
```




```
# Plot Real GDP Growth
ggplot(data %>% filter(wb_countryname == country & year >= 1990 & year <= 2023),
       aes(x = year, y = wdi_ny_gdp_mktp_kd_zg)) +
  geom_line(color = "cadetblue") +
  labs(title = paste("GDP Growth in", country),
       y = "GDP Growth, annual (%)",
       x = "Year") +
  theme_minimal()
```


```
# Plot Population Growth
ggplot(data %>% filter(wb_countryname == country & year >= 1990 & year <= 2023),
       aes(x = year, y = wdi_sp_pop_grow)) +
  geom_line(color = "cadetblue2") +
  labs(title = paste("Population Growth in", country),
       y = "Population Growth, annual (%)",
       x = "Year") +
  theme_minimal()
```

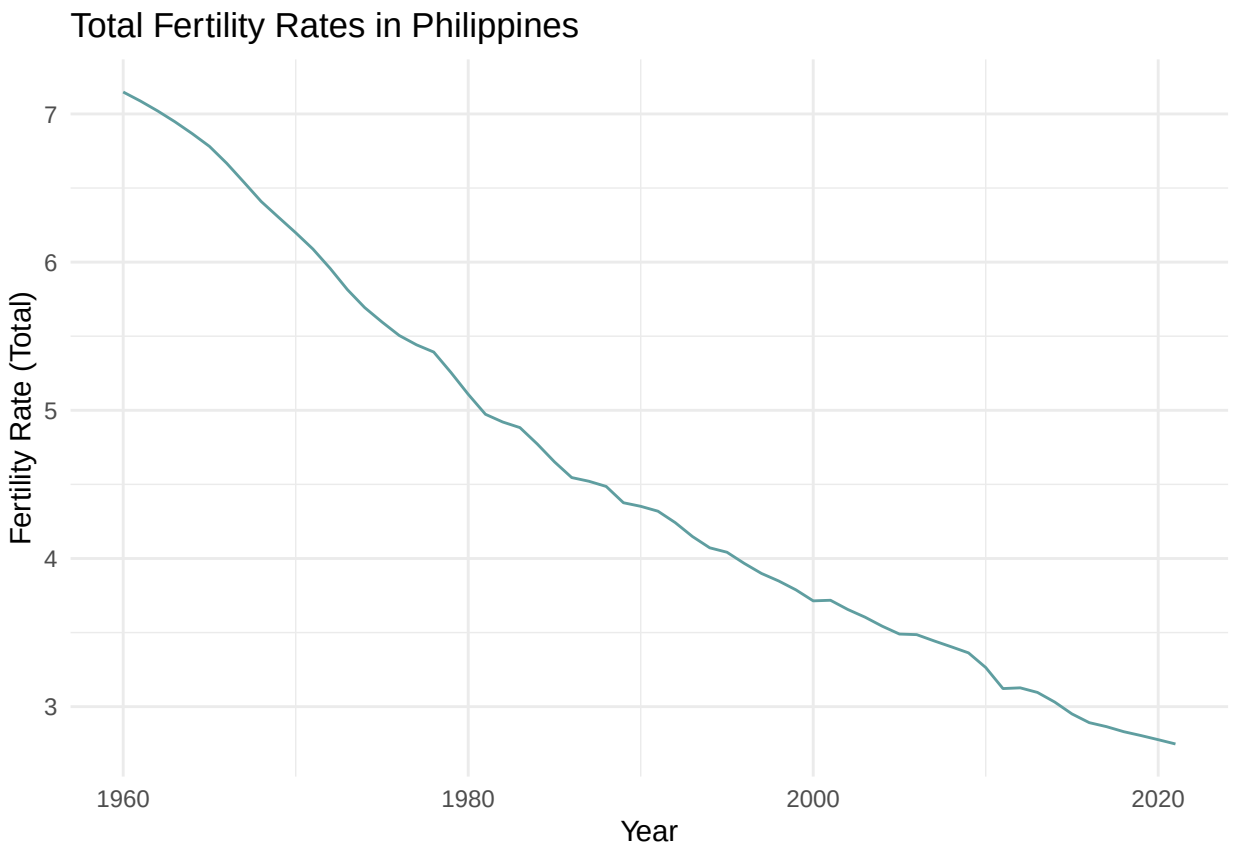

```
# Plot GDP Per Capita Growth
ggplot(data %>% filter(wb_countryname == country & year >= 1990 & year <= 2023),
       aes(x = year, y = wdi_ny_gdp_pcap_kd_zg)) +
  geom_line(color = "cadetblue3") +
  labs(title = paste("GDP per Capita Growth in", country),
       y = "GDP per Capita Growth (%)",
       x = "Year") +
  theme_minimal()
```



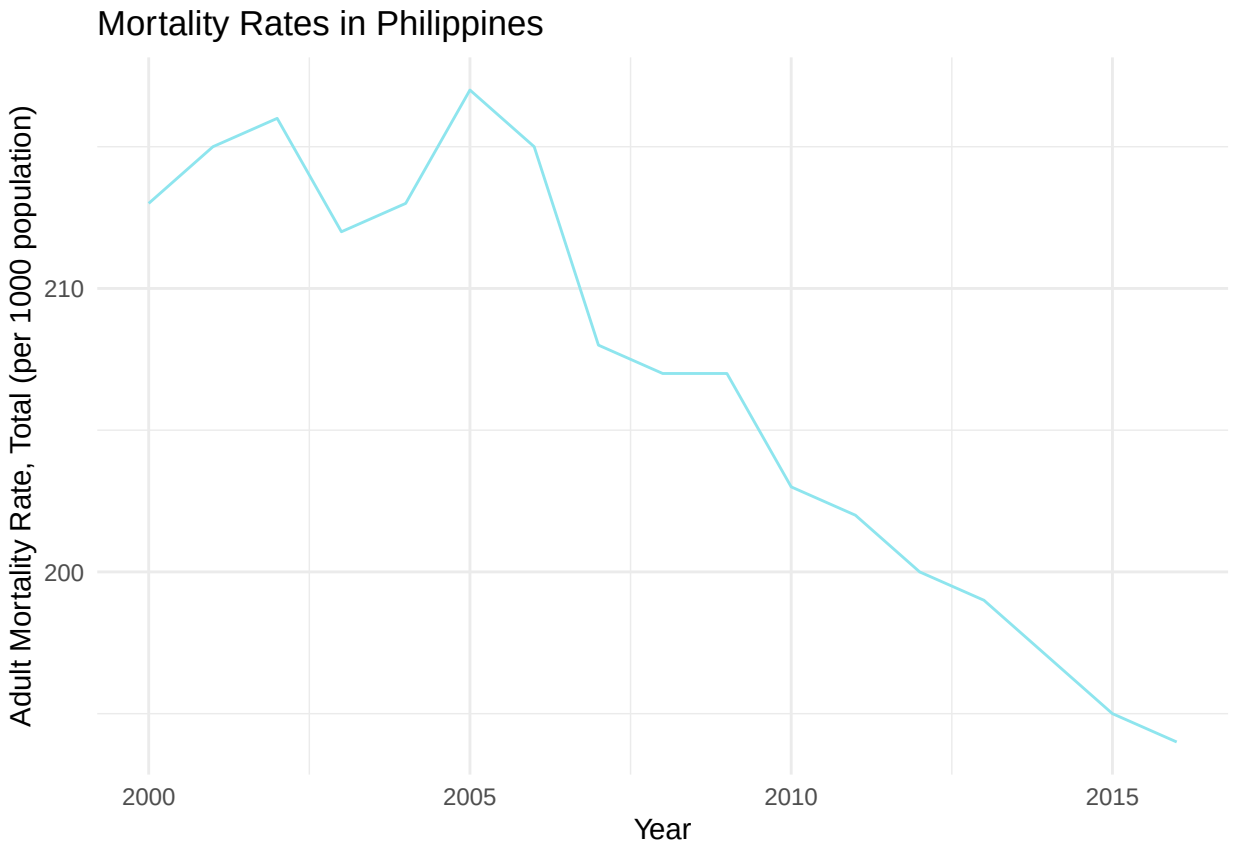
Here I write the interpretation of my results (maximum 200 words).

Question 2: Demographic Transition (10 points)

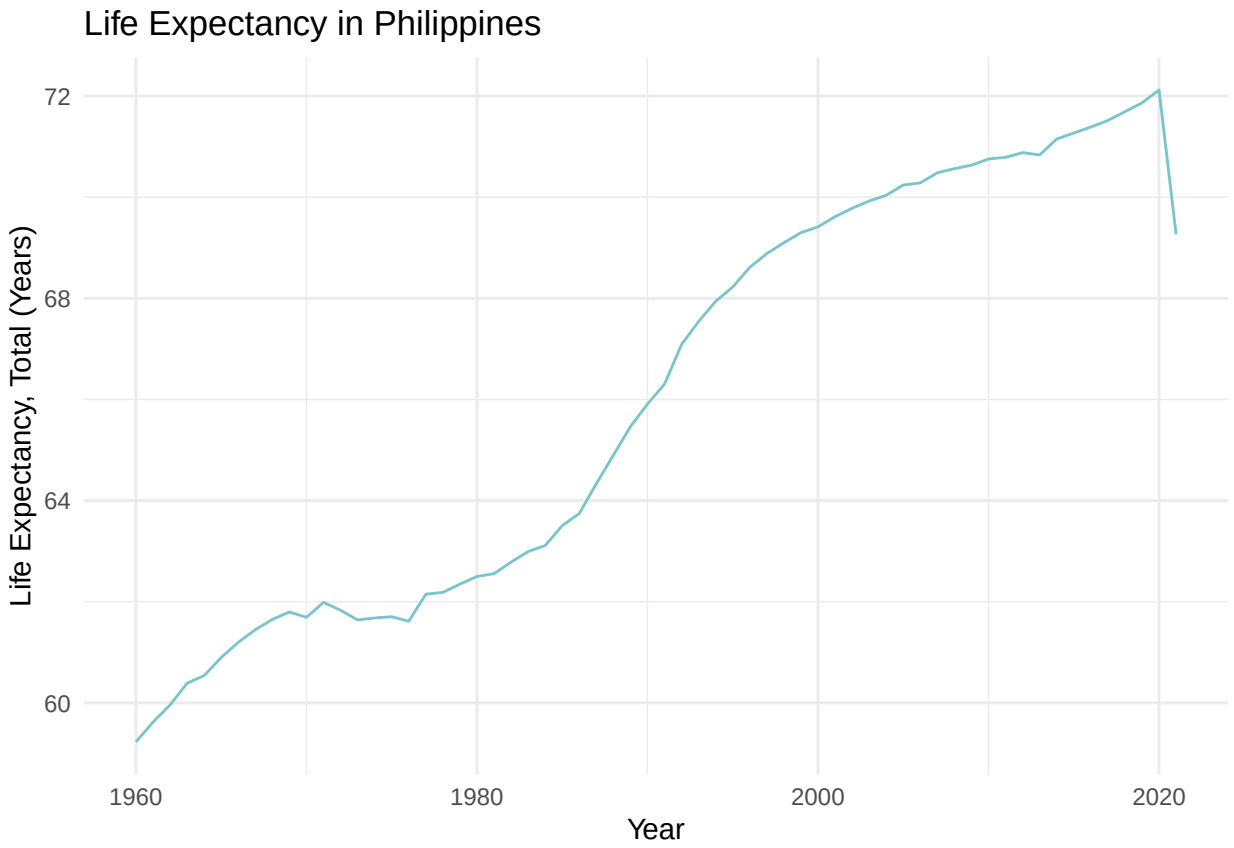
```
# Fertility Rates Time Series
ggplot(data %>% filter(wb_countryname == country & !is.na(wdi_sp_dyn_tfrt_in)),
       aes(x = year, y = wdi_sp_dyn_tfrt_in)) +
  geom_line(color = "cadetblue") +
  labs(title = paste("Total Fertility Rates in", country),
       y = "Fertility Rate (Total)",
       x = "Year") +
  theme_minimal()
```



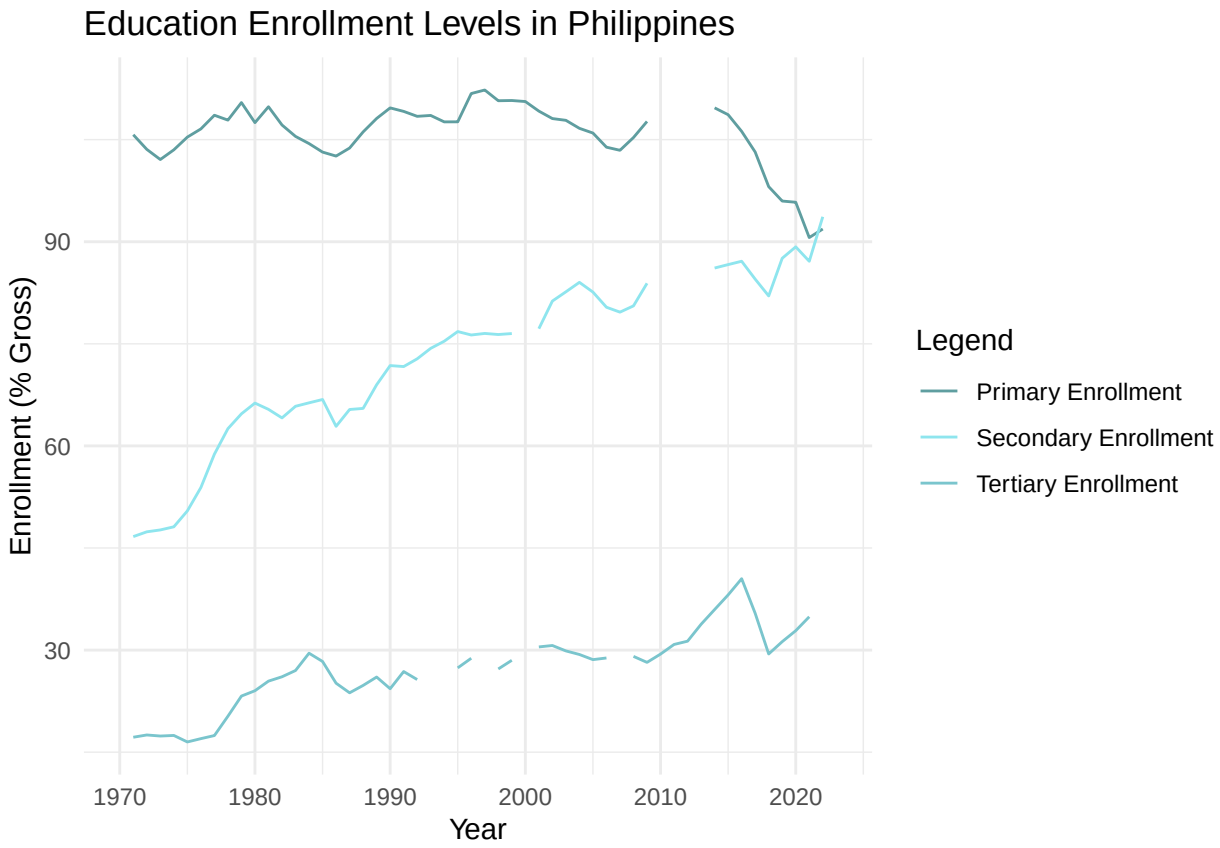

```
# Mortality Rates Time Series
ggplot(data %>% filter(wb_countryname == country & !is.na(qog_who_mrt)),
       aes(x = year, y = qog_who_mrt)) +
  geom_line(color = "cadetblue2") +
  labs(title = paste("Mortality Rates in", country),
       y = "Adult Mortality Rate, Total (per 1000 population)",
       x = "Year") +
  theme_minimal()
```



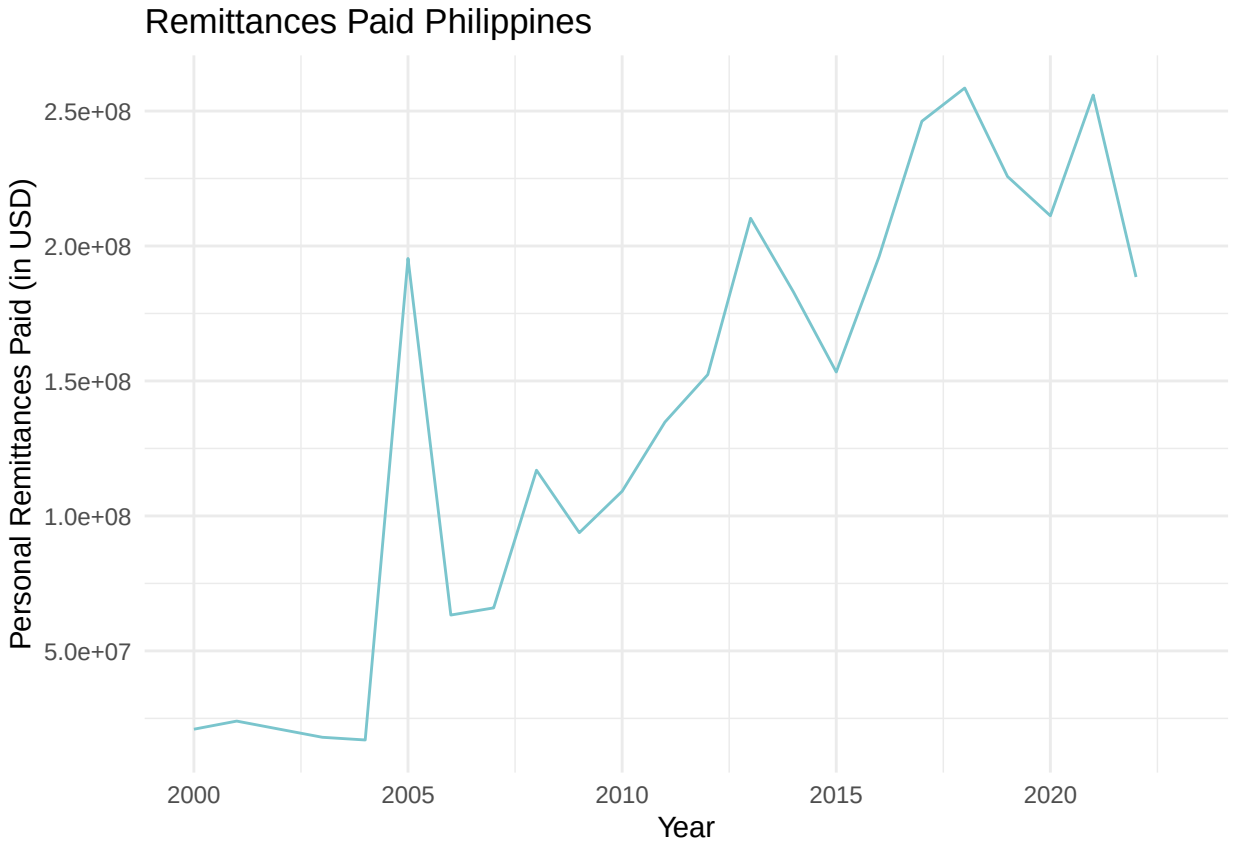

```
# Life Expectancy Time Series
ggplot(data %>% filter(wb_countryname == country & !is.na(wdi_sp_dyn_le00_in)),
       aes(x = year, y = wdi_sp_dyn_le00_in)) +
  geom_line(color = "cadetblue3") +
  labs(title = paste("Life Expectancy in", country),
       y = "Life Expectancy, Total (Years)",
       x = "Year") +
  theme_minimal()
```




```r
# Plot Education Levels (Primary, Secondary, Tertiary)
ggplot(data %>% filter(wb_countryname == country & year >= 1970 & year <= 2023),
       aes(x = year)) +
  geom_line(aes(y = wdi_se_prm_enrr, color = "Primary Enrollment")) +
  geom_line(aes(y = wdi_se_sec_enrr, color = "Secondary Enrollment")) +
  geom_line(aes(y = wdi_se_ter_enrr, color = "Tertiary Enrollment")) +
  scale_color_manual(values = c("Primary Enrollment" = "cadetblue",
                                "Secondary Enrollment" = "cadetblue2",
                                "Tertiary Enrollment" = "cadetblue3")) +
  labs(title = paste("Education Enrollment Levels in", country),
       y = "Enrollment (% Gross)",
       x = "Year",
       color = "Legend") +
  theme_minimal()
```



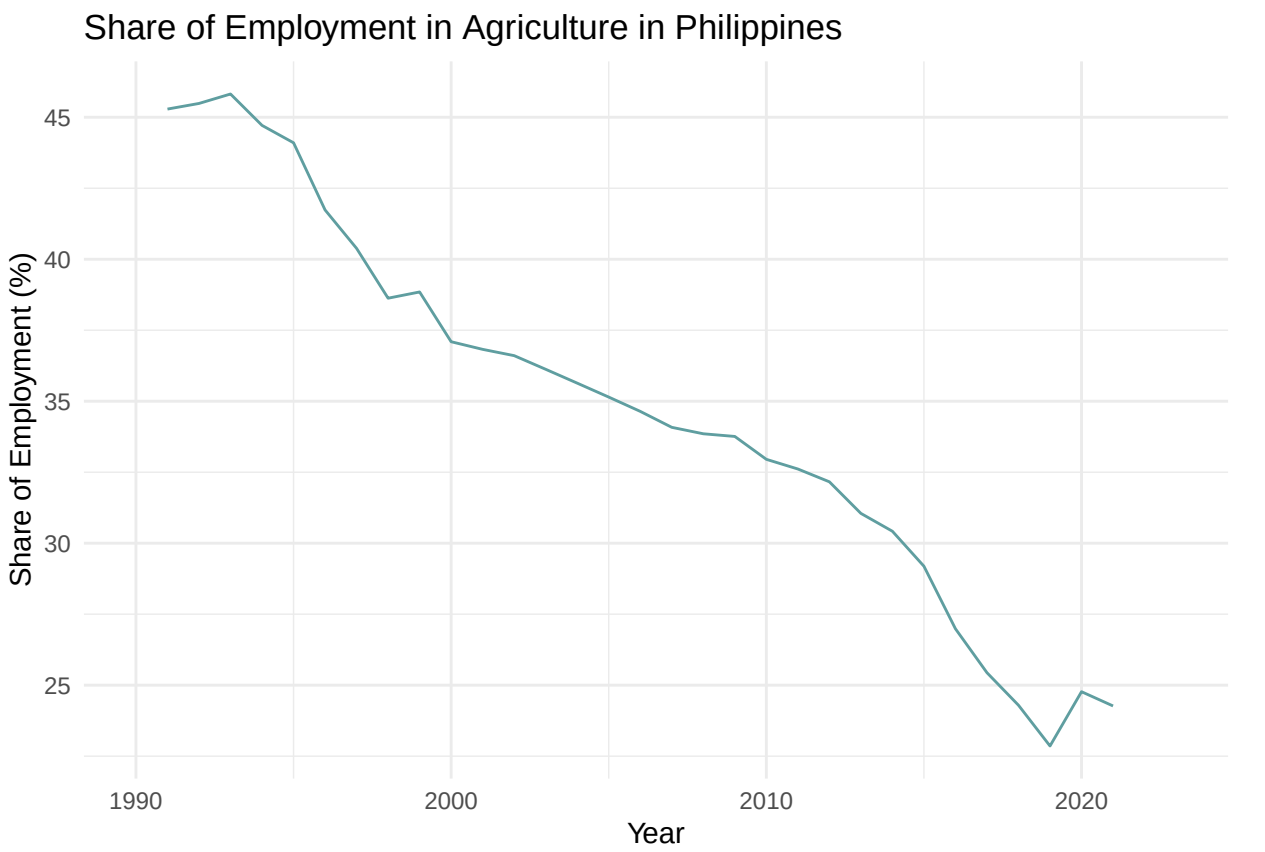

```r
# Additional Plot: Remittances
ggplot(data %>% filter(wb_countryname == country & year >= 2000 & year <= 2023),
       aes(x = year)) +
  geom_line(aes(y = wdi_bm_trf_pwkr_cd_dt), color = "cadetblue3") +
  labs(title = paste("Remittances Paid", country),
       y = "Personal Remittances Paid (in USD)",
       x = "Year") +
  theme_minimal()
```



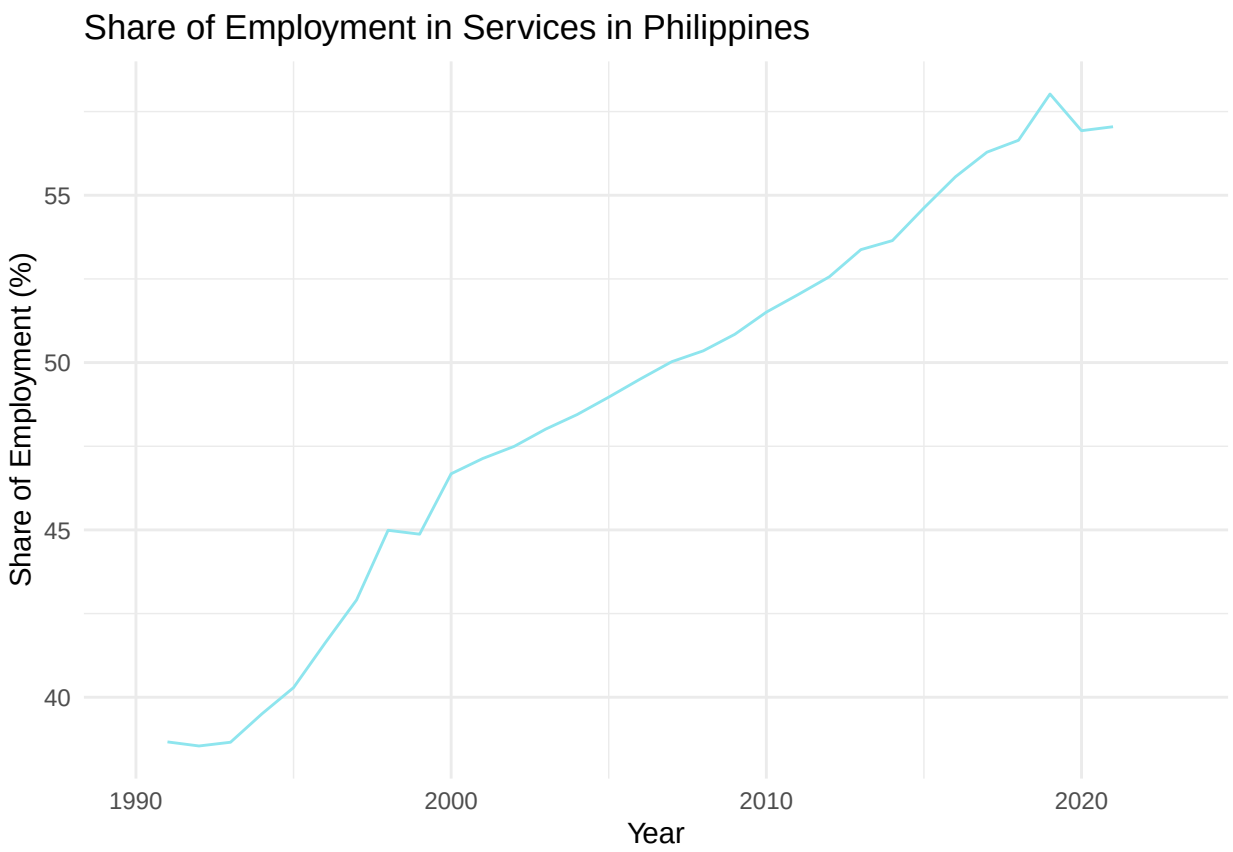
Here I write the interpretation of my results (maximum 200 words).

Question 3: History of Structural Transformation (20 points)

```
# Employment Share in Agriculture
ggplot(data %>% filter(wb_countryname == country & year >= 1990 & year <= 2023),
       aes(x = year, y = wdi_sl_agr_empl_zs)) +
  geom_line(color = "cadetblue") +
  labs(title = paste("Share of Employment in Agriculture in", country),
       y = "Share of Employment (%)",
       x = "Year") +
  theme_minimal()
```




```
# Plot Employment Share in Services
ggplot(data %>% filter(wb_countryname == country & year >= 1990 & year <= 2023),
       aes(x = year, y = wdi_sl_srv_empl_zs)) +
  geom_line(color = "cadetblue2") +
  labs(title = paste("Share of Employment in Services in", country),
       y = "Share of Employment (%)",
       x = "Year") +
  theme_minimal()
```

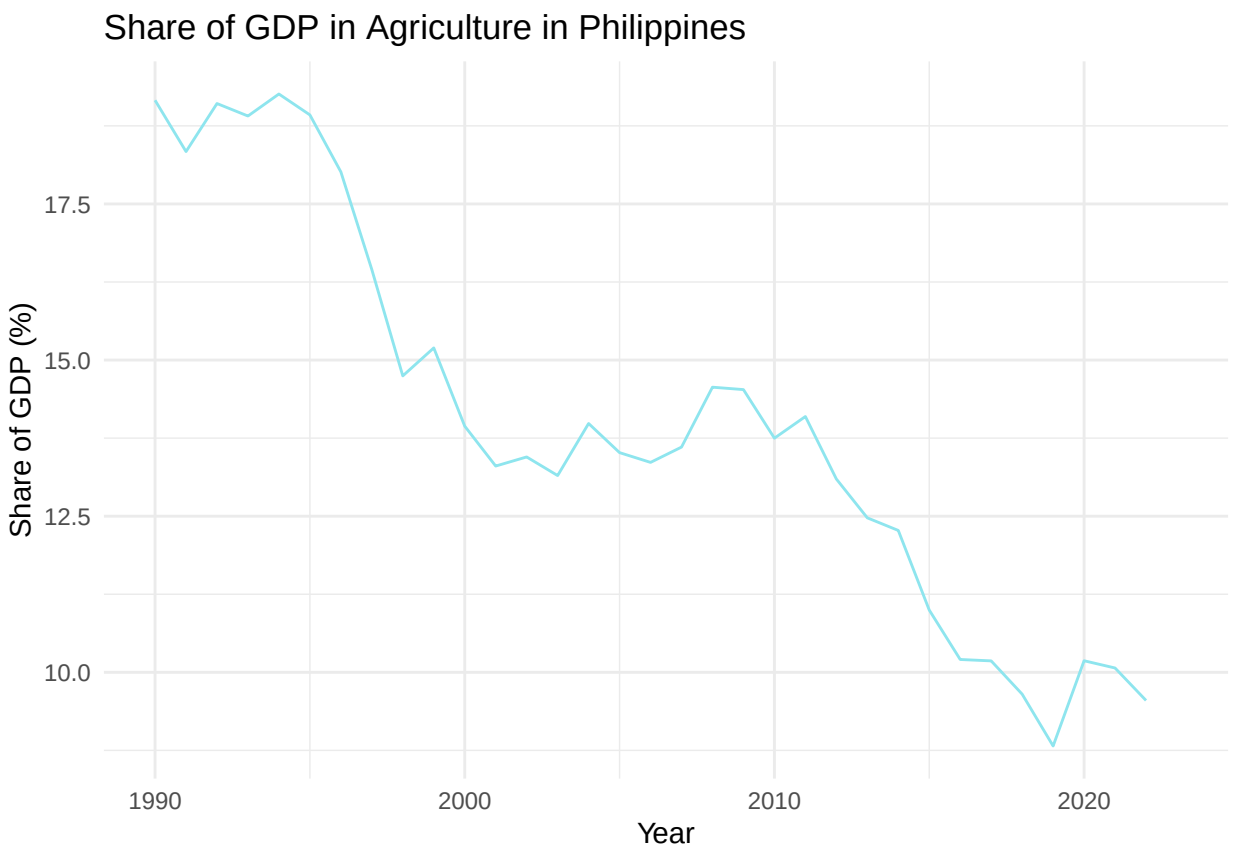



```
# Plot Employment Share in Manufacturing
ggplot(data %>% filter(wb_countryname == country & year >= 1990 & year <= 2023),
       aes(x = year, y = wdi_sl_ind_empl_zs)) +
  geom_line(color = "cadetblue3") +
  labs(title = paste("Employment in Manufacturing in", country),
       y = "Share of Employment (%)",
       x = "Year") +
  theme_minimal()
```


```r
# Convert total GDP from billions to US$
data <- data %>%
  mutate(total_gdp_usd = weo_ngdpd * 1e9)
```

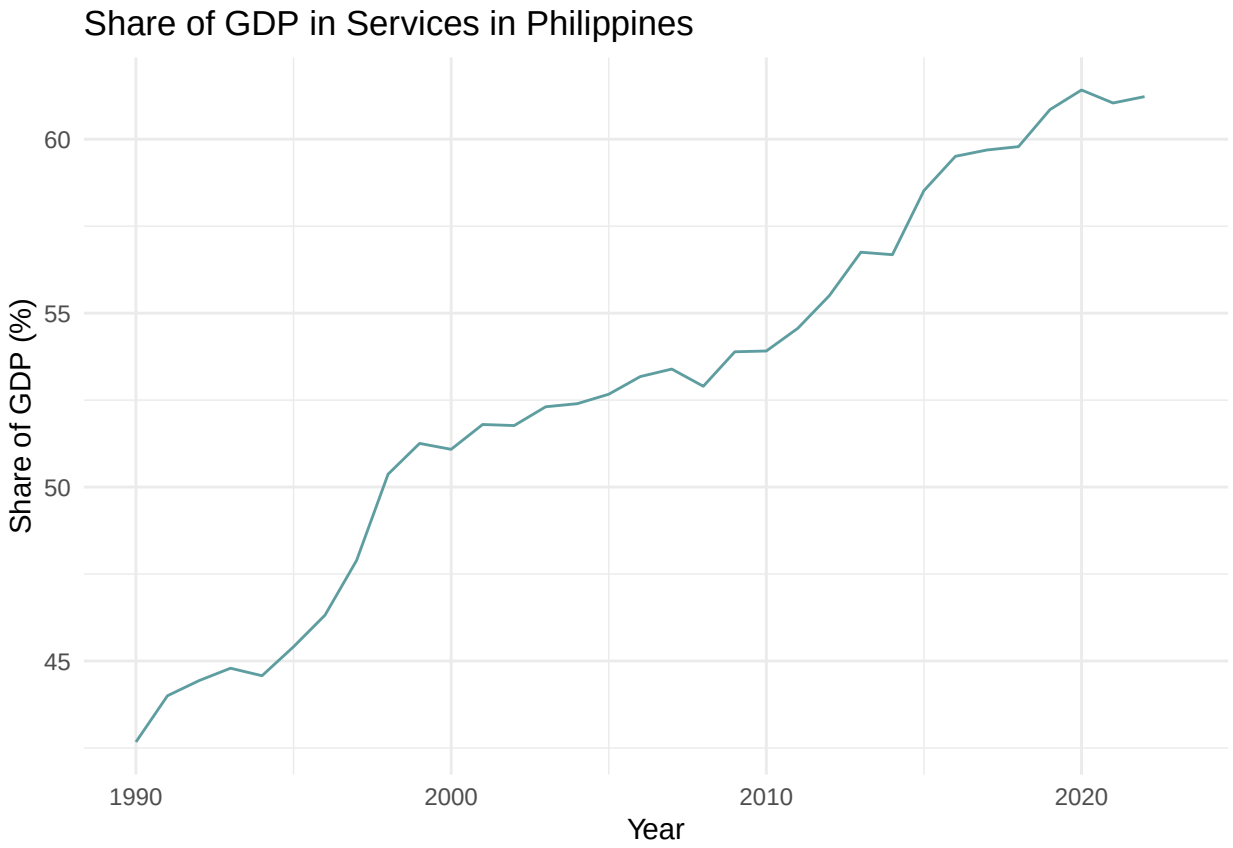
```r
# Calculate Shares of GDP
data <- data %>%
  mutate(
    gdp_share_agriculture = (wdi_nv_agr_totl_cd / total_gdp_usd) * 100,
    gdp_share_industry = (wdi_nv_ind_totl_cd / total_gdp_usd) * 100,
    gdp_share_services = (wdi_nv_srv_totl_cd / total_gdp_usd) * 100
  )
```

```r
# Plot Share of GDP in Agriculture
ggplot(data %>% filter(wb_countryname == country & year >= 1990 & year <= 2023),
       aes(x = year, y = gdp_share_agriculture)) +
  geom_line(color = "cadetblue2") +
  labs(title = paste("Share of GDP in Agriculture in", country),
       y = "Share of GDP (%)",
       x = "Year") +
  theme_minimal()
```




```
# Plot Share of GDP in Industry
ggplot(data %>% filter(wb_countryname == country & year >= 1990 & year <= 2023),
       aes(x = year, y = gdp_share_industry)) +
  geom_line(color = "cadetblue3") +
  labs(title = paste("Share of GDP in Industry in", country),
       y = "Share of GDP (%)",
       x = "Year") +
  theme_minimal()
```

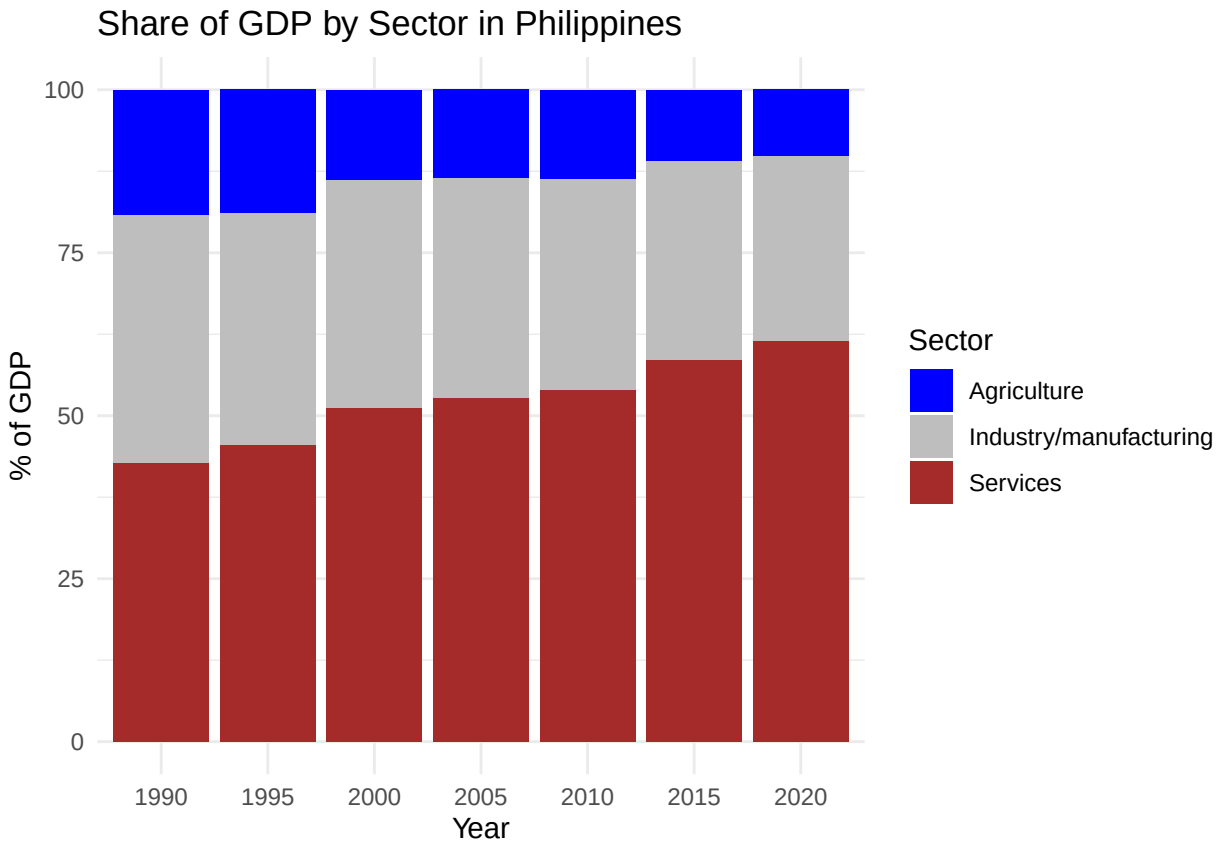

```
# Plot Share of GDP in Services
ggplot(data %>% filter(wb_countryname == country & year >= 1990 & year <= 2023),
       aes(x = year, y = gdp_share_services)) +
  geom_line(color = "cadetblue") +
  labs(title = paste("Share of GDP in Services in", country),
       y = "Share of GDP (%)",
       x = "Year") +
  theme_minimal()
```




```r
# Reshape GDP shares data of agriculture, industry, and services sector

data_long <- data %>%
  filter(wb_countryname == country & year >= 1990 & year <= 2023) %>%
  select(year, gdp_share_agriculture, gdp_share_industry, gdp_share_services) %>%
  pivot_longer(cols = starts_with("gdp_share"),
               names_to = "Sector",
               values_to = "Share") %>%
  mutate(Sector = recode(Sector,
                         gdp_share_agriculture = "Agriculture",
                         gdp_share_industry = "Industry/manufacturing",
                         gdp_share_services = "Services"))
```

```r
# Plot Share of GDP by sector in stacked bar
data_long_filtered <- data_long %>%
  filter(year %% 5 == 0)  # Display data every five years

ggplot(data_long_filtered, aes(x = factor(year), y = Share, fill = Sector)) +
  geom_bar(stat = "identity", position = "stack") +
  scale_fill_manual(values = c("Agriculture" = "blue",
                               "Industry/manufacturing" = "gray",
                               "Services" = "brown")) +
  labs(
    title = paste("Share of GDP by Sector in", country),
    x = "Year",
    y = "% of GDP",
    fill = "Sector"
```

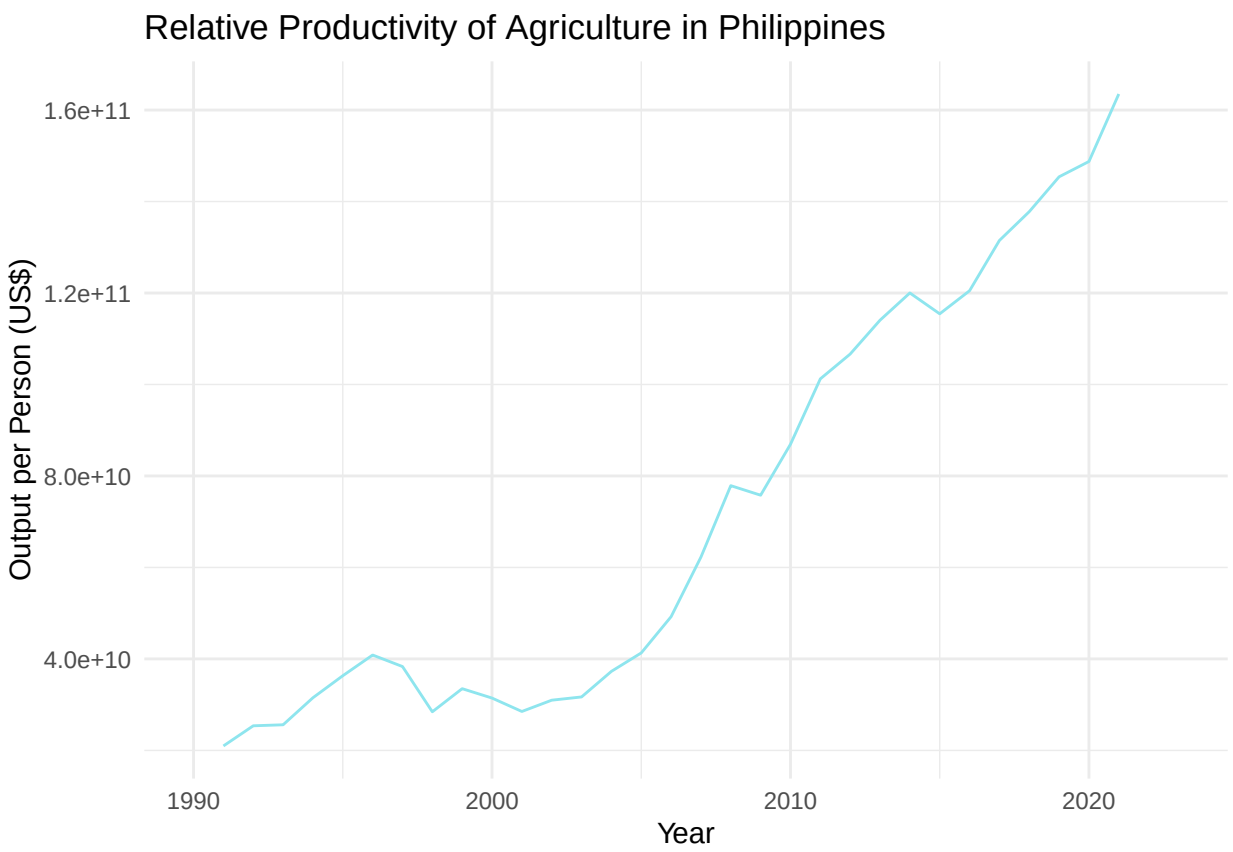
```
  ) +
  theme_minimal()
```



Plot Relative Productivity

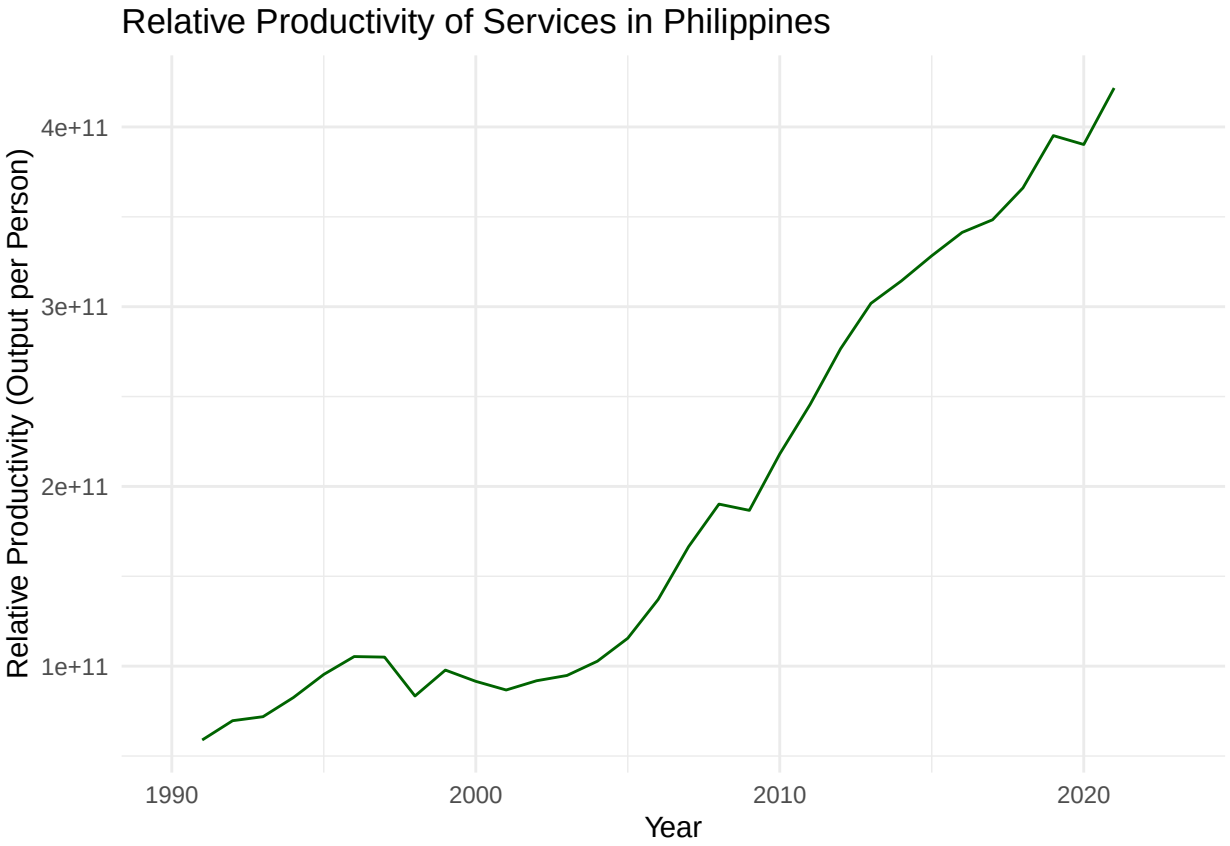
```
# Calculate relative productivity of agriculture
data <- data %>%
  mutate(
    rel_prod_agriculture = wdi_nv_agr_totl_cd / (wdi_sl_agr_empl_zs / 100)
  )
```

```
# Plot relative productivity in agriculture

ggplot(data %>% filter(wb_countryname == country & year >= 1990 & year <= 2023),
       aes(x = year, y = rel_prod_agriculture)) +
  geom_line(color = "cadetblue2") +
  labs(title = paste("Relative Productivity of Agriculture in", country),
       y = "Output per Person (US$)",
       x = "Year") +
  theme_minimal()
```




```
# Calculate Relative Productivity for Manufacturing (Industry)
data <- data %>%
  mutate(
    manf_rp = wdi_nv_ind_manf_cd / (wdi_sl_ind_empl_zs / 100)
  )

# Plot Relative Productivity for Manufacturing
ggplot(data %>% filter(wb_countryname == country & year >= 2000 & year <= 2023),
       aes(x = year, y = manf_rp)) +
  geom_line(color = "cadetblue") +
  labs(
    title = paste("Relative Productivity of Manufacturing in", country),
    y = "Relative Productivity (Output per Person)",
    x = "Year"
  ) +
  theme_minimal()
```


```
# Calculate Relative Productivity for Services
data <- data %>%
  mutate(
    manf_rp = wdi_nv_srv_totl_cd / (wdi_sl_srv_empl_zs / 100)
  )

# Plot Relative Productivity for Services
ggplot(data %>% filter(wb_countryname == country & year >= 1990 & year <= 2023),
       aes(x = year, y = manf_rp)) +
  geom_line(color = "darkgreen") +
  labs(
    title = paste("Relative Productivity of Services in", country),
    y = "Relative Productivity (Output per Person)",
    x = "Year"
  ) +
  theme_minimal()
```



Here I write the interpretation of my results (maximum 200 words).

Question 4: Structural Breaks (10 points)

Filter Data for Chosen Country

```
# Filter data for the chosen country
data_filtered <- data %>%
  filter(wb_countryname == country) %>%
  filter(!is.na(wdi_ny_gdp_pcap_kd_zg)) %>%
  rename(gdp_pc_g = wdi_ny_gdp_pcap_kd_zg)
```

Model: All Years

```
model1 <- lm(gdp_pc_g ~ year, data = data_filtered)
summary(model1)
```

```
##
## Call:
## lm(formula = gdp_pc_g ~ year, data = data_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5189  -0.6438   0.7174   1.8595   4.5351
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -39.11446   47.08122  -0.831    0.409
## year          0.02062    0.02364   0.872    0.387
##
## Residual standard error: 3.331 on 60 degrees of freedom
## Multiple R-squared:  0.01252,    Adjusted R-squared:  -0.003935
## F-statistic: 0.7609 on 1 and 60 DF,  p-value: 0.3865
```

Identify Structural Breaks

```
break_test1 <- breakpoints(gdp_pc_g ~ year, data = data_filtered)
summary(break_test1)
```

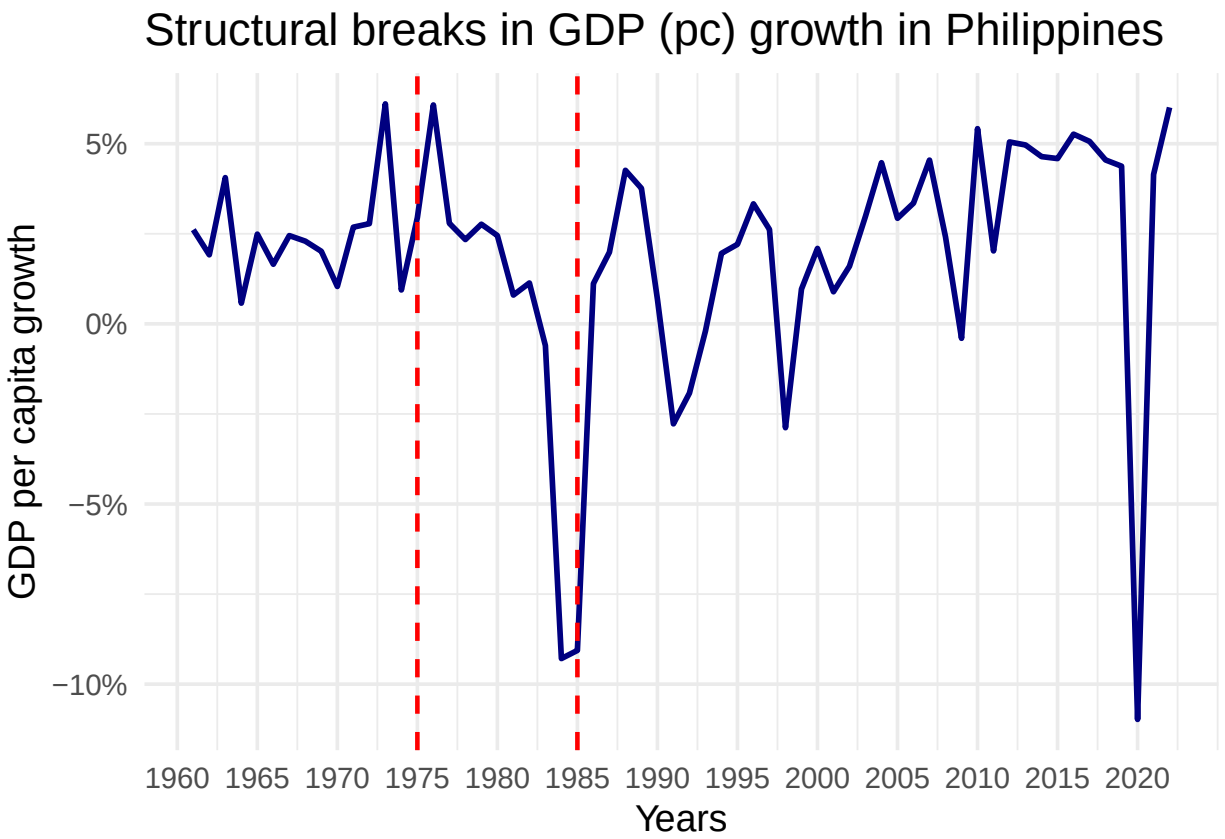
```
##
##   Optimal (m+1)-segment partition:
##
## Call:
## breakpoints.formula(formula = gdp_pc_g ~ year, data = data_filtered)
##
## Breakpoints at observation number:
##
## m = 1      22
## m = 2   15 25
## m = 3   15 25       51
## m = 4   15 25    42 51
## m = 5   15 25 34 43 52
##
## Corresponding to breakdates:
##
## m = 1                     0.354838709677419
## m = 2   0.241935483870968 0.403225806451613
## m = 3   0.241935483870968 0.403225806451613
## m = 4   0.241935483870968 0.403225806451613                    0.67741935483871
## m = 5   0.241935483870968 0.403225806451613 0.548387096774194 0.693548387096774
##
## m = 1
## m = 2
## m = 3   0.82258064516129
## m = 4   0.82258064516129
## m = 5   0.838709677419355
##
## Fit:
##
## m   0     1     2     3     4     5
## RSS 665.8 551.5 425.4 397.7 384.3 382.8
## BIC 335.5 336.2 332.5 340.7 351.0 363.1
```

```
# Extract break years
breaks <- break_test1$breakpoints
break_years <- data_filtered$year[breaks]
break_years
```

```
## [1] 1975 1985
```

Structural Breaks Plot

```
plot_breaks <- ggplot(data_filtered, aes(x = year, y = gdp_pc_g)) +
  geom_line(color = "navy", linewidth = 1) +
  geom_vline(xintercept = break_years, linetype = "dashed", color = "red",
             linewidth = 0.8) +
  scale_y_continuous(labels = scales::percent_format(scale = 1),
                     breaks = scales::pretty_breaks(n = 6)) +
  scale_x_continuous(breaks = seq(1960, 2021, by = 5)) +
  labs(y = "GDP per capita growth", x = "Years",
       title = paste0("Structural breaks in GDP (pc) growth in ", country)) +
  theme_minimal(base_size = 14)
plot_breaks
```

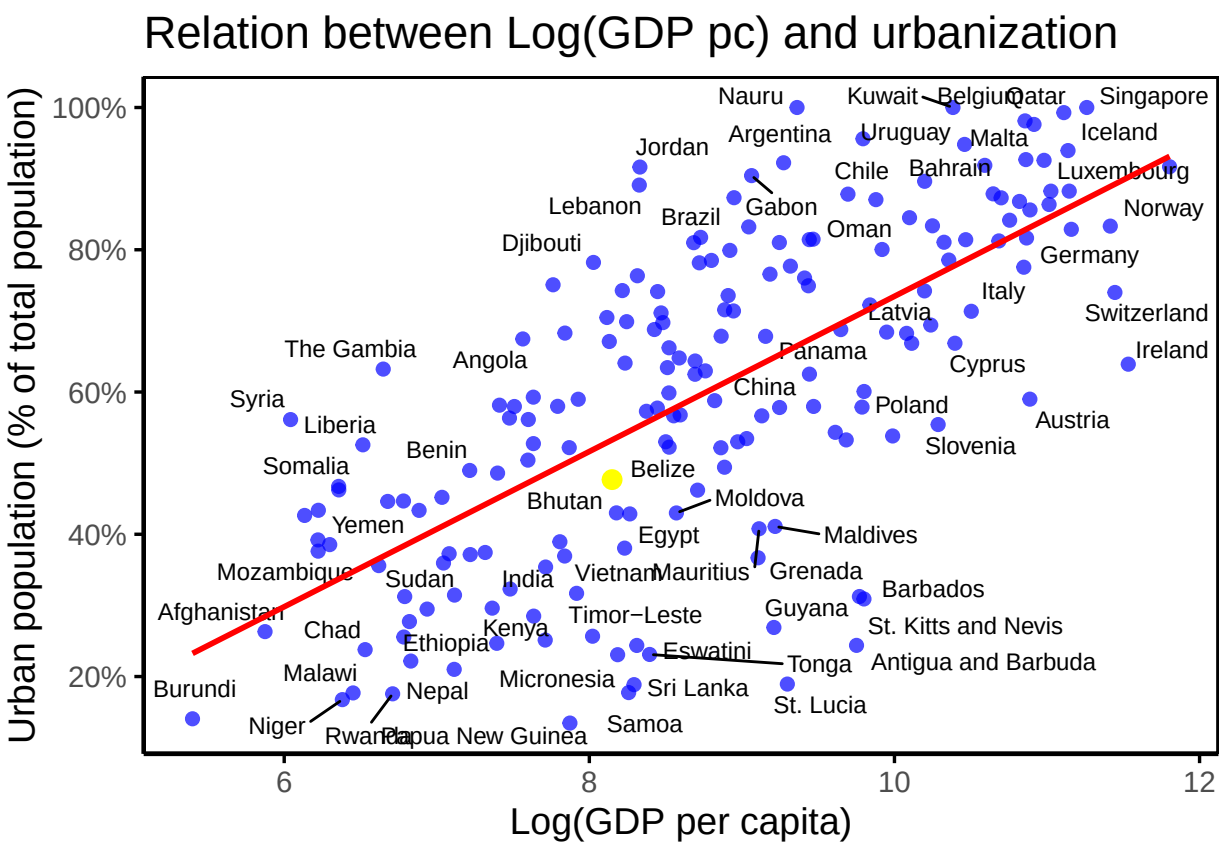


Here I write the interpretation of my results (maximum 200 words).

Question 5: Structural transformation comparison with other countries (10 points)

```
# Add log of GDP per Capita
data$log_gdp_pc <- log(data$wdi_ny_gdp_pcap_cd)

# Add dummy column for the chosen country
data <- data %>%
  mutate(dummy_chosen_country = ifelse(wb_countryname == country, 1, 0))
```

```r
# I am using 1960 as it is the first year for Philippines
data_filtered_1960 <- data %>%
  filter(year == 1960) %>%
  filter(!is.na(qog_wdi_popurb) & !is.na(wdi_ny_gdp_pcap_cd))

data_filtered_2021 <- data %>%
  filter(year == 2021) %>%
  filter(!is.na(qog_wdi_popurb) & !is.na(wdi_ny_gdp_pcap_cd))

#Plot Log(GDP) and Urban Population
ggplot(data_filtered_2021, aes(x = log_gdp_pc, y = qog_wdi_popurb)) +
  geom_point(color = "blue", size = 2, alpha = 0.7) +  # Scatter plot points
  geom_point(
    data = data_filtered_2021[data_filtered_2021$dummy_chosen_country == 1, ],
    aes(x = log_gdp_pc, y = qog_wdi_popurb),
    color = "yellow",
    size = 3
  ) +
  geom_text_repel(aes(label = wb_countryname), size = 3) +  # Repelled country labels
  geom_smooth(method = "lm", color = "red", se = FALSE, linetype = "solid") +
  scale_y_continuous(labels = scales::percent_format(scale = 1),
                     breaks = scales::pretty_breaks(n = 6)) +
  labs(
    title = "Relation between Log(GDP pc) and urbanization",
    y = "Urban population (% of total population)",
    x = "Log(GDP per capita)"
  ) +
  theme_minimal(base_size = 14) +  # Minimal theme for a clean look
  theme(
    panel.border = element_rect(color = "black", fill = NA, linewidth = 0.8),
    axis.line = element_line(color = "black", linewidth = 0.8),
    axis.ticks = element_line(color = "black", linewidth = 0.6),
    panel.grid.major = element_blank(),  # Remove major grid lines
    panel.grid.minor = element_blank()   # Remove minor grid lines
  )
```

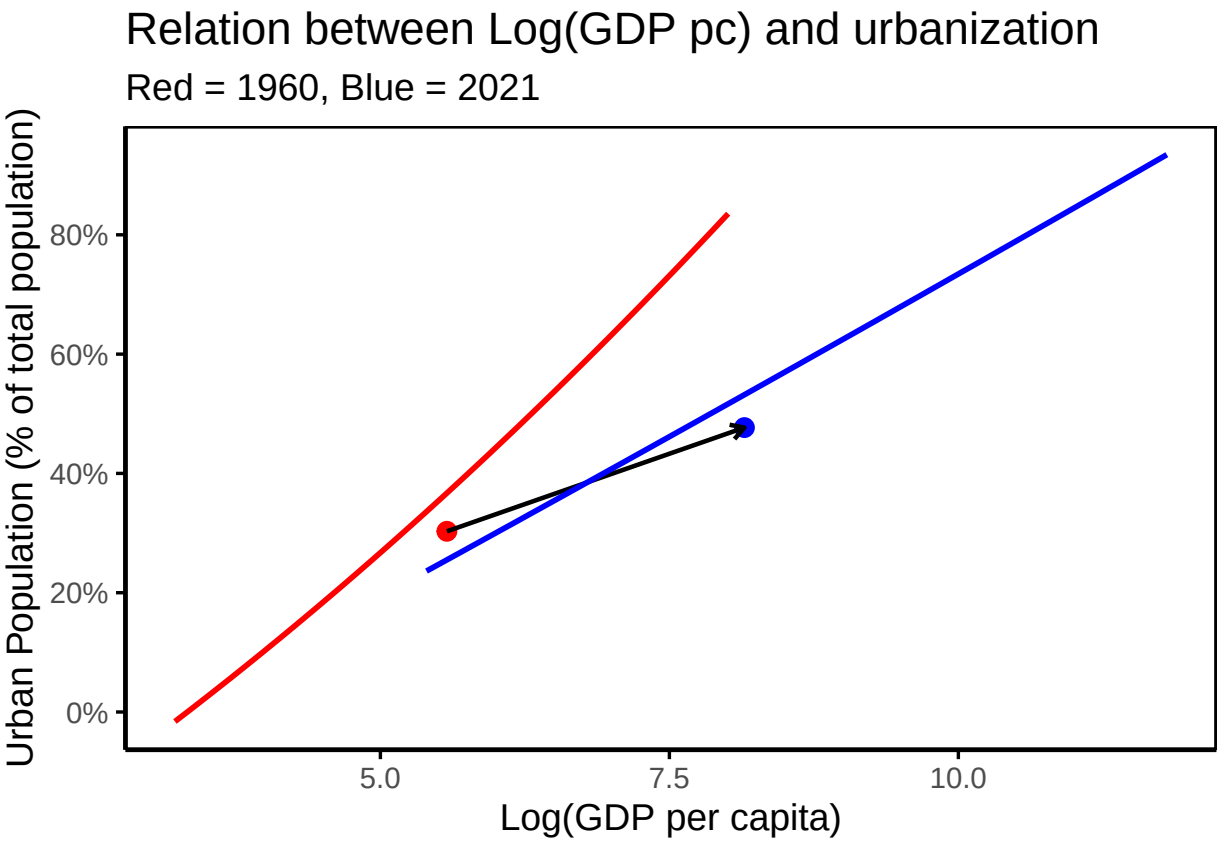


Here I write the interpretation of my results (maximum 200 words).

Question 6: Comparative Trends in Structural Transformation (10 points)

```
ggplot() +
  # Highlight chosen country in 1960 (red point)
  geom_point(
    data = data_filtered_1960[data_filtered_1960$dummy_chosen_country == 1, ],
    aes(x = log_gdp_pc, y = qog_wdi_popurb),
    color = "red",
    size = 3
  ) +
  # Highlight chosen country in 2021 (blue point)
  geom_point(
    data = data_filtered_2021[data_filtered_2021$dummy_chosen_country == 1, ],
    aes(x = log_gdp_pc, y = qog_wdi_popurb),
    color = "blue",
    size = 3
  ) +
  # Add arrow connecting 1960 and 2021 points
  geom_segment(
    data = merge(
      data_filtered_1960[data_filtered_1960$dummy_chosen_country == 1, ],
      data_filtered_2021[data_filtered_2021$dummy_chosen_country == 1, ],
      by = "dummy_chosen_country"
```

```
  ),
  aes(
    x = log_gdp_pc.x, y = qog_wdi_popurb.x,
    xend = log_gdp_pc.y, yend = qog_wdi_popurb.y
  ),
  arrow = arrow(length = unit(0.2, "cm")),
  color = "black",
  linewidth = 0.8
) +
# Polynomial regression for 1988 (red curve)
geom_smooth(
  data = data_filtered_1960,
  aes(x = log_gdp_pc, y = qog_wdi_popurb),
  method = "lm",
  formula = y ~ poly(x, 2),
  color = "red",
  se = FALSE,
  linetype = "solid"
) +
# Polynomial regression for 2021 (blue curve)
geom_smooth(
  data = data_filtered_2021,
  aes(x = log_gdp_pc, y = qog_wdi_popurb),
  method = "lm",
  formula = y ~ poly(x, 2),
  color = "blue",
  se = FALSE,
  linetype = "solid"
) +
scale_y_continuous(labels = scales::percent_format(scale = 1),
                   breaks = scales::pretty_breaks(n = 6)) +
# Axis labels
labs(
  title = "Relation between Log(GDP pc) and urbanization",
  subtitle = "Red = 1960, Blue = 2021",
  y = "Urban Population (% of total population)",
  x = "Log(GDP per capita)"
) +
# Minimal theme
theme_minimal(base_size = 14) +
theme(
  panel.border = element_rect(color = "black", fill = NA, linewidth = 0.8),
  axis.line = element_line(color = "black", linewidth = 0.8),
  axis.ticks = element_line(color = "black", linewidth = 0.6),
  panel.grid.major = element_blank(),
  panel.grid.minor = element_blank()
)
```



Here I write the interpretation of my results (maximum 200 words).

Question 7: Conclusion (30 points)

Here I write the conclusion (maximum 300-500 words).
