Introduction to Data Science - K12 Group Assignment

Final Analysis: GDP Growth (Target 1) and Youth NEET (Target 2)

K12 Group

2025-12-11

Contents

1 Introduction

This analysis examines two key Sustainable Development Goals (SDGs):

SDG Target 8.1: Sustained GDP growth SDG Target 8.6: Reducing youth not in employment, education, or training (NEET) We analyze global data to understand economic growth patterns and youth employment across different continents and development levels.

2 1) Setup and Data Loading

2.1 a) Load Required Packages

```
library(tidyverse)
library(ggplot2)
library(ggthemes)
library(ggrepel)
library(ggpubr)
library(knitr)

knitr::opts_chunk$set(
  fig.width = 6,
  fig.height = 4,
  out.width = "90%"
)
```

2.2 b) Load and clean core datasets

```
### Set working directory to folder containing the CSV files
setwd('C:/Users/slee7/OneDrive - Imperial College London/Introduction to Data Science/IDS
↪  Midterm/data sets')

### Read and store 3 CSV files and 1 custom CSV file, separating data by comma delimiter
# - csv1: continent classification
# - csv2: GDP per capita (PPP 2017 USD $)
# - csv3: youth NEET share (% of youth 15-24)
# - csv4: LDC classification, ease of doing business index, HDI
csv1 <- read.csv(file = "continents-according-to-our-world-in-data.csv",
                 sep = ",")

csv2 <- read.csv(file = "gdp-per-capita-worldbank.csv",
                 sep = ",")


csv3 <- read.csv(file = "youth-not-in-education-employment-training.csv",
                 sep = ",")

csv4 <- read.csv(file = "Group K12 Custom Data.csv",
                 sep = ",")

### Remove Year Data and Entity Data - we drop Year and Entity because we only need each
↪  country's continent
csv1 <- csv1 %>% mutate(Year = NULL,
                        Entity = NULL)

### Primary key is the unique country code, given by column 'Code'
### Left join attaches continent information to GDP and NEET data using the 3-letter
↪  country code as the key
gdp <- csv2 %>% left_join(csv1, join_by(Code))
youth <- csv3 %>% left_join(csv1, join_by(Code))

### Rename long column name into shorter, clearer labels:
# - GDP Per Capita
```

```r
# - Youth NEET Share
gdp <- gdp %>% rename("GDP Per Capita" =
↪  "GDP.per.capita..PPP..constant.2017.international...")
youth <- youth %>% rename("Youth NEET Share" =
↪  "Share.of.youth.not.in.education..employment.or.training..total....of.youth.population."
↪  )
```

3 2) Data wrangling for GDP per capita and GDP growth rates

3.1 [Section 5 Data Wrangling] Entities classified into Continents and graphed by average GDP growth rate

```r
### For each country (Code), compute the year-on-year GDP per capita growth rate based on
↪  GDP per capita
# - This approximates the SDG Target 8.1 (sustained growth)
gdp <- gdp %>%
  group_by(Code) %>%
  mutate(`GDP Per Capita Growth Rate` = ((`GDP Per Capita` - lag(`GDP Per Capita`)) /
  ↪  lag(`GDP Per Capita`)) * 100) # calculates growth rate

### For each country (Code), compute the mean GDP growth rate across all years into a new
↪  data frame
### Then join continent info using left join so we can summarise and visualise by
↪  continent
growth_rate_country <- gdp %>%
  group_by(Code) %>%
  summarize("Mean GDP Per Capita Growth Rate" = mean(`GDP Per Capita Growth Rate`, na.rm
  ↪  = TRUE)) %>%
  left_join(csv1) %>%
  na.omit() # omitting any N/A values
```

3.2 [Section 6 Data Wrangling] Europe categorically classified by GDP and graphed by GDP growth rate

```r
### Filter European countries and restrict to 1990-2020, this period aligns with the
↪  availability of higher-quality data and the SDG focus
europe_data <- gdp %>%
  filter(Continent == "Europe",
         Year >= 1990,
         Year <= 2020)

### For each European country, compute the average GDP per capita (1990-2020)
# - This is used as a proxy for development level.
europe_avg_gdp <- europe_data %>%
  group_by(Entity, Code) %>%
  summarise(`Avg GDP Per Capita` = mean(`GDP Per Capita`, na.rm = TRUE),
            .groups = "drop")

### Use the median of average GDP per capita to split countries into:
# - "Upper Half GDP" (above or equal to median)
# - "Lower Half GDP" (below median)
```

```r
median_gdp <- median(europe_avg_gdp$`Avg GDP Per Capita`, na.rm = TRUE)

europe_avg_gdp <- europe_avg_gdp %>%
  mutate(`GDP Classification` = if_else(`Avg GDP Per Capita` >= median_gdp,
                                        "Upper Half GDP",
                                        "Lower Half GDP"))

### Attach development level back to the full panel (country-year) dataset
europe_data <- europe_data %>%
  left_join(europe_avg_gdp %>%
              select(Code, `GDP Classification`),
            by = "Code") %>%
  filter(!is.na(`GDP Classification`))

### For each continent, compute the mean GDP growth rate for each year into a new data
↪  frame
growth_rate_per_year <- gdp %>%
  group_by(Continent, Year) %>%
  summarize("Mean GDP Per Capita Growth Rate" = mean(`GDP Per Capita Growth Rate`, na.rm
  ↪  = TRUE), .groups = "drop") %>%
  na.omit() # omitting any N/A values
```

4 3) Data wrangling for share of youth NEET

4.1 [Section 8 Data Wrangling]

```r
# - Average NEET by continent and year (unweighted across countries)
# - We filter to Year <= 2020 to match the SDG 8.6 target horizon
youth_sum <- youth %>%
  group_by(Continent, Year) %>%
  filter(!is.na(Continent), Year <= 2020) %>%
  summarise(
    mean_neet = mean(`Youth NEET Share`, na.rm = TRUE),
    .groups = "drop"
  )
```

4.2 [Section 9 Data Wrangling] Bottom 5 and top 5 European countries by recent GDP and graphed by youth NEET share

```r
# - We filter to Year <= 2020 to match the SDG 8.6 target horizon
youth <- youth %>%
  group_by(Continent, Year) %>%
  filter(!is.na(Continent), Year <= 2020)

### Exclude San Marino due to lack of GDP data in 2021
### Exclude Ukraine due to lack of Youth NEET data
bot5top5_recent <- gdp %>%
  filter(Continent == "Europe") %>%
  filter(Code != "UKR") %>%
  filter(Year == 2021) %>%
  arrange(`GDP Per Capita`)
```

```r
### Filter top 5 and bottom 5 European countries by GDP Per Capita
bot5top5 <- bind_rows(head(bot5top5_recent, 5) %>%
                        mutate(Group = "Bottom 5"),
                      tail(bot5top5_recent, 5) %>%
                        mutate(Group = "Top 5"))

### Left join to include Youth NEET Share
bot5top5 <- youth %>%
  left_join(bot5top5 %>%
              mutate(Entity = NULL,
                     Year = NULL,
                     `GDP Per Capita` = NULL,
                     Continent= NULL,
                     `GDP Per Capita Growth Rate` = NULL), join_by(Code)) %>%
  na.omit() # omitting any N/A values
```

5 4) Data wrangling for custom K12 CSV

5.1 [Section 10 Data Wrangling]

```r
### Filter only LDC countries
### Delete year column that only has 2015
ldc <- csv4 %>%
  mutate(Year = NULL) %>%
  filter(LDC.Classification == TRUE) %>%
  left_join(gdp, join_by(Code))

### For each year, calculate the mean GDP growth rate for all LDC countries
# - This is used as a proxy for development level.
ldc_avg_growth <- ldc %>%
  group_by(Year, Continent.x) %>%
  summarise("Avg GDP Per Capita" = mean(`GDP Per Capita Growth Rate`, na.rm = TRUE),
            .groups = "drop") %>%
  na.omit()
```

5.2 [Section 11 Data Wrangling]

```r
### Left join adds ease of business index information necessary for graph
growth_rate_country_eodb <- growth_rate_country %>%
  left_join(csv4, join_by(Code))
```

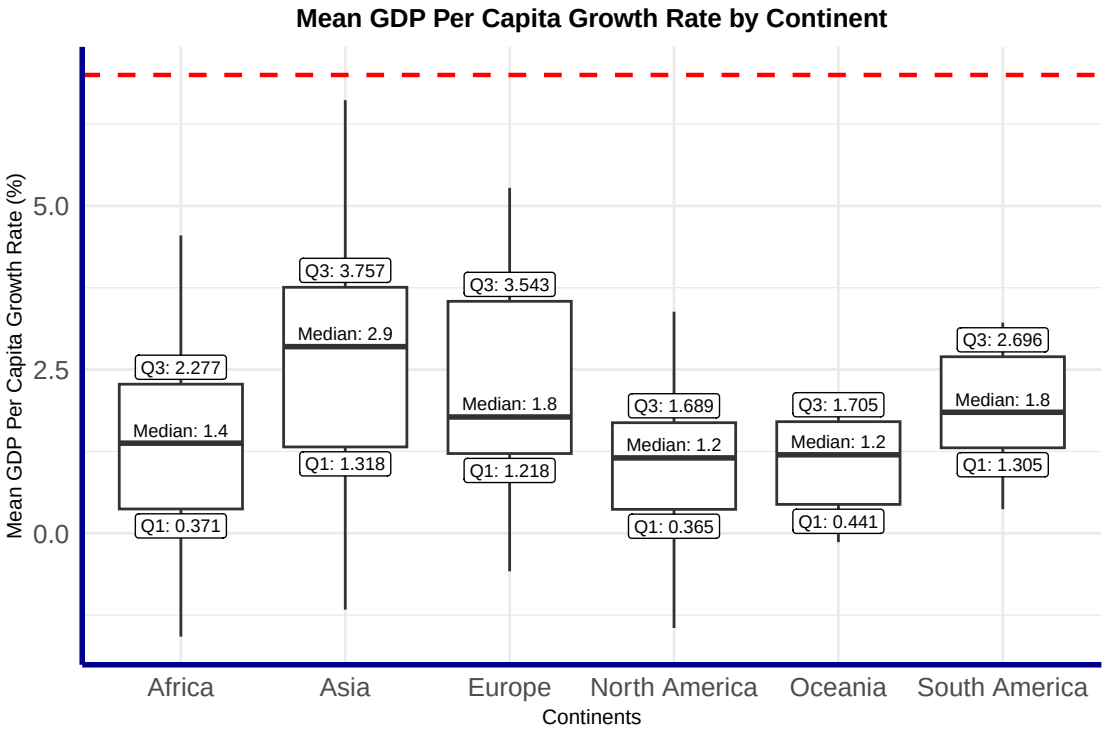
5.3 [Section 12 Data Wrangling]

```r
### Get most recent youth NEET share for each country
youth_recent <- youth %>%
  group_by(Code) %>%
  arrange(Year) %>%
  summarise(`Most Recent Youth NEET Share` = last(`Youth NEET Share`))

### Left join adds HDI index information necessary for graph
youth_recent <- youth_recent %>%
```

```
  left_join(csv4, join_by(Code))
```

6 5) Graph: Average GDP per capita across individual continents

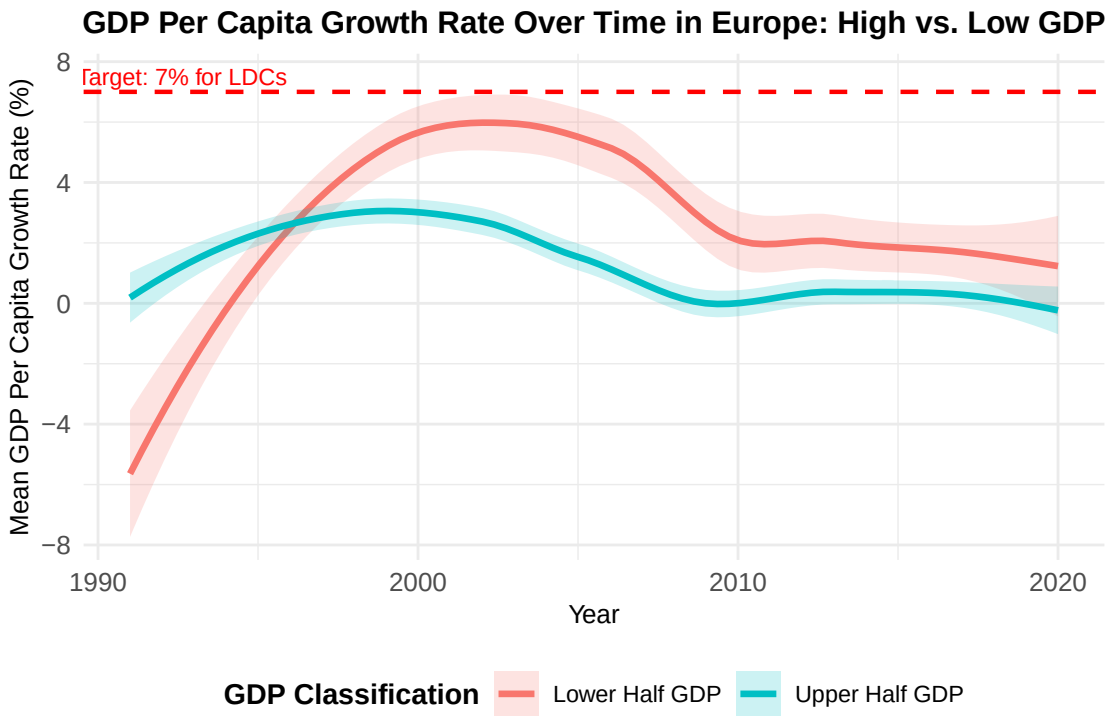
```
# - Remove outliers from visual (still used in calculating median and quartiles)
# - Stat summary adds labels of median, Q1, and Q3
# - Theme elements changes aesthetics
growth_rate_country %>%
  ggplot(aes(
    x = Continent,
    y = `Mean GDP Per Capita Growth Rate`)) +
  geom_boxplot(outliers = FALSE) +
  geom_hline(yintercept = 7,
             linetype = "dashed",
             color = "red",
             size = 0.8) +
  labs(
    x = "Continents",
    y = "Mean GDP Per Capita Growth Rate (%)",
    title = "Mean GDP Per Capita Growth Rate by Continent"
  ) +
  stat_summary(fun.data = function(x) data.frame(y=median(x),
  ↪  label=paste("Median:",round(median(x),1))),
               geom = "text", vjust = -0.5, size = 2.5) +
  stat_summary(fun.data = ~ data.frame(quarts = quantile(.x, probs = .25)),
               aes(y = stage(`Mean GDP Per Capita Growth Rate`, after_stat = quarts),
                   label = paste("Q1:",round(after_stat(quarts), digits = 3))),
               geom = "label", vjust = 1.2, size = 2.5) +
  stat_summary(fun.data = ~ data.frame(quarts = quantile(.x, probs = .75)),
               aes(y = stage(`Mean GDP Per Capita Growth Rate`, after_stat = quarts),
                   label = paste("Q3:",round(after_stat(quarts), digits = 3))),
               geom = "label", vjust = -0.2, size = 2.5) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 10, face = "bold"),
    axis.text = element_text(size = 10),
    axis.title = element_text(size = 8),
    axis.line = element_line(colour = "darkblue", size = 1),
    legend.title = element_text(face = "bold")
  )
```



7 6) Graph: European countries GDP growth breakdown

7.1 a) Graph: Europe GDP growth over time by average GDP

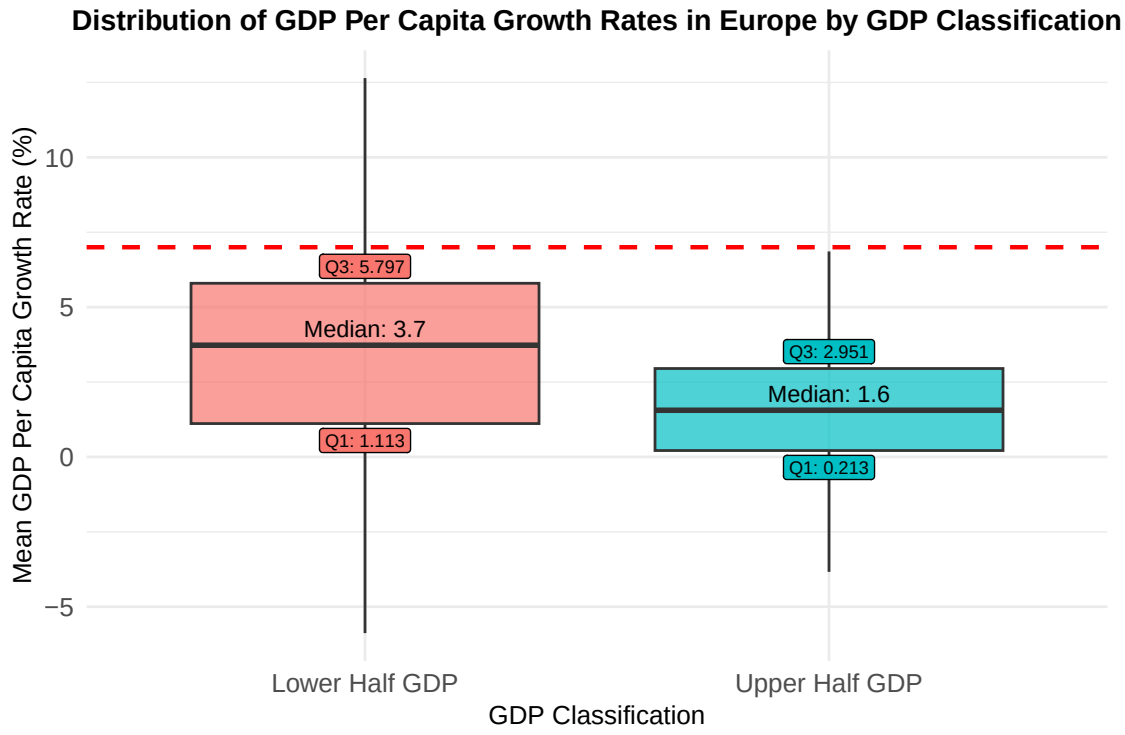
```
# - Smoothed graph paths for high vs low GDP European countries.
# - The dashed red line at 7% marks the UN target for LDCs in SDG 8.1.
europe_data %>%
  ggplot(aes(x = Year, y = `GDP Per Capita Growth Rate`, color = `GDP Classification`)) +
  geom_smooth(method = "loess", aes(fill = `GDP Classification`), alpha = 0.2, size =
  ↪  1.2) +
  geom_hline(yintercept = 7, linetype = "dashed", color = "red", size = 0.8) +
  annotate("text", x = min(europe_data$Year, na.rm = TRUE) + 2, y = 7.5, label = "UN
  ↪  Target: 7% for LDCs", color = "red", size = 3) +
  labs(
    title = "GDP Per Capita Growth Rate Over Time in Europe: High vs. Low GDP",
    x = "Year",
    y = "Mean GDP Per Capita Growth Rate (%)",
    color = "GDP Classification",
    fill  = "GDP Classification"
  ) +
  theme_minimal() +
  theme(
    plot.title    = element_text(hjust = 0.5, size = 12, face = "bold"),
    axis.text     = element_text(size = 10),
    axis.title    = element_text(size = 10),
    legend.title = element_text(face = "bold"),
    legend.position = "bottom"
  )
```



7.2 b) Graph: Distribution of GDP growth in Europe by development level

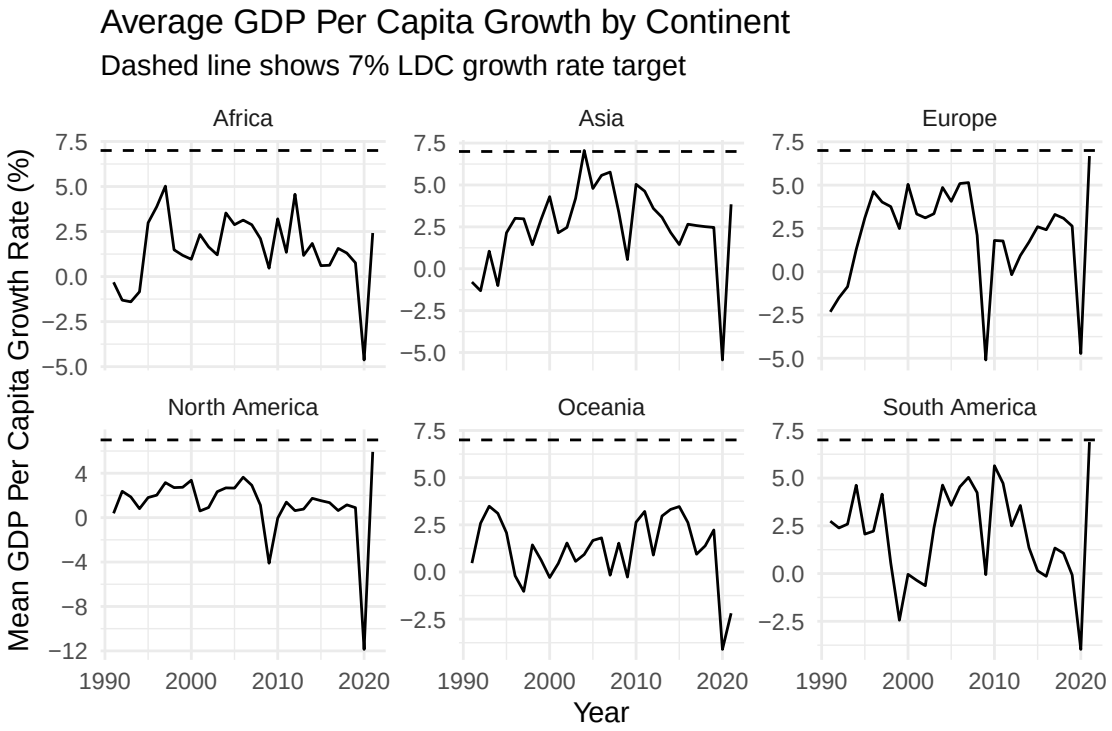
```
# - Boxplot visually removes extreme outliers to focus on the bulk of the distribution.
# - Compares typical growth volatility for more vs less developed countries.

europe_data %>%
  ggplot(aes(x = `GDP Classification`,
             y = `GDP Per Capita Growth Rate`,
             fill = `GDP Classification`)) +
  geom_boxplot(outliers = FALSE, alpha = 0.7) +
  geom_hline(yintercept = 7,
             linetype = "dashed",
             color = "red",
             size = 0.8) +
  labs(
    title = "Distribution of GDP Per Capita Growth Rates in Europe by GDP
    ↪  Classification",
    x = "GDP Classification",
    y = "Mean GDP Per Capita Growth Rate (%)",
    fill = "GDP Classification"
  ) +
  stat_summary(fun.data = function(x) data.frame(y=median(x),
  ↪  label=paste("Median:",round(median(x),1))),
               geom = "text", vjust = -0.5, size = 3) +
  stat_summary(fun.data = ~ data.frame(quarts = quantile(.x, probs = .25)),
               aes(y = stage(`GDP Per Capita Growth Rate`, after_stat = quarts),
                   label = paste("Q1:",round(after_stat(quarts), digits = 3))),
               geom = "label", vjust = 1.2, size = 2.5) +
  stat_summary(fun.data = ~ data.frame(quarts = quantile(.x, probs = .75)),
```

```
              aes(y = stage(`GDP Per Capita Growth Rate`, after_stat = quarts),
                  label = paste("Q3:",round(after_stat(quarts), digits = 3))),
              geom = "label", vjust = -0.2, size = 2.5) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 11, face = "bold"),
    axis.text  = element_text(size = 10),
    axis.title = element_text(size = 10),
    legend.position = "none"
  )
```



8 7) Graph: Average GDP per capita growth by continent (1990-)

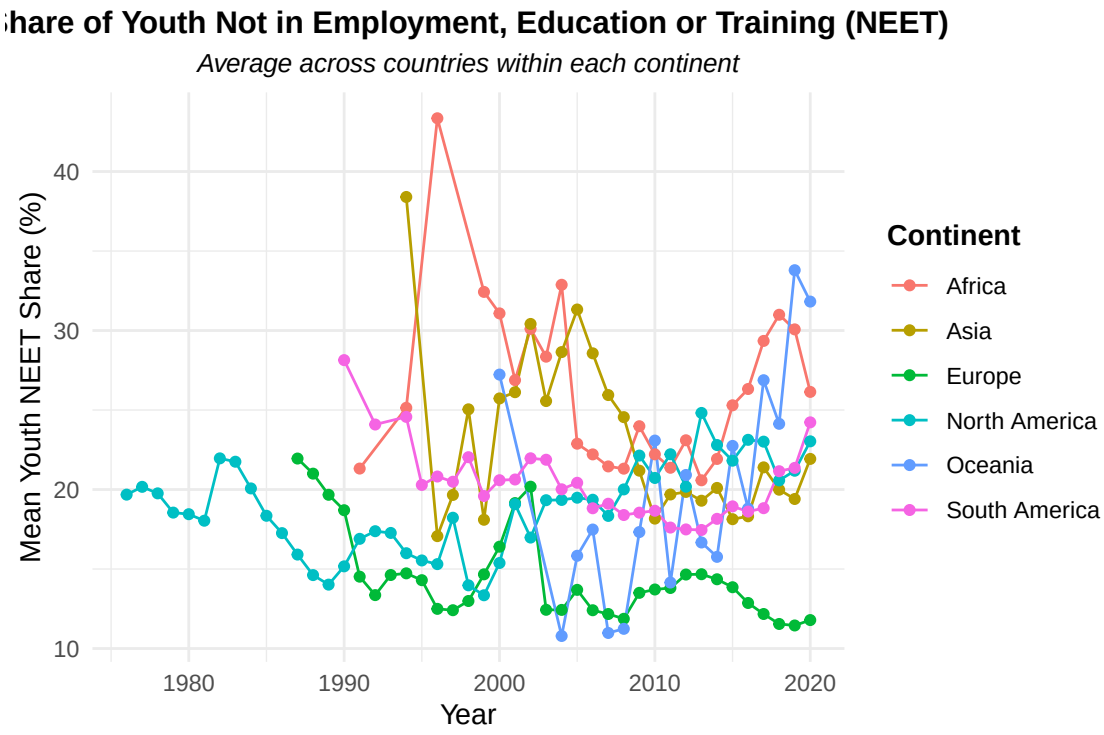
```
### Graphs mean GDP growth rate of each continent by year
# - Geom_hline creates dashed line, showing 7% LDC growth rate target
# - Facet_wrap groups the six separate continent graphs into one graph
ggplot(growth_rate_per_year,
       aes(x = Year, y = `Mean GDP Per Capita Growth Rate`)) +
  geom_hline(yintercept = 7, linetype = "dashed") +
  geom_line() +
  facet_wrap(~ Continent, scales = "free_y") +
  labs(
    title = "Average GDP Per Capita Growth by Continent",
    subtitle = "Dashed line shows 7% LDC growth rate target",
    x = "Year",
    y = "Mean GDP Per Capita Growth Rate (%)"
  ) +
  theme_minimal()
```



9 8) Youth NEET across all continents

9.1 8a) Graph: share of youth NEET across all the continents over time

```
### Line plot: evolution of average NEET share across continents over time
# - Line plot chosen to clearly see trends over years

ggplot(youth_sum,
       aes(x = Year, y = mean_neet, colour = Continent)) +
  geom_line() +
  geom_point() +
  labs(
    title = "Share of Youth Not in Employment, Education or Training (NEET)",
    subtitle = "Average across countries within each continent",
    x = "Year",
    y = "Mean Youth NEET Share (%)",
    colour = "Continent"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 12, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5, size = 10, face = "italic"),
    legend.title = element_text(face = "bold")
  )
```



9.2 8b) Table summary of absolute and percentage change of youth NEET share

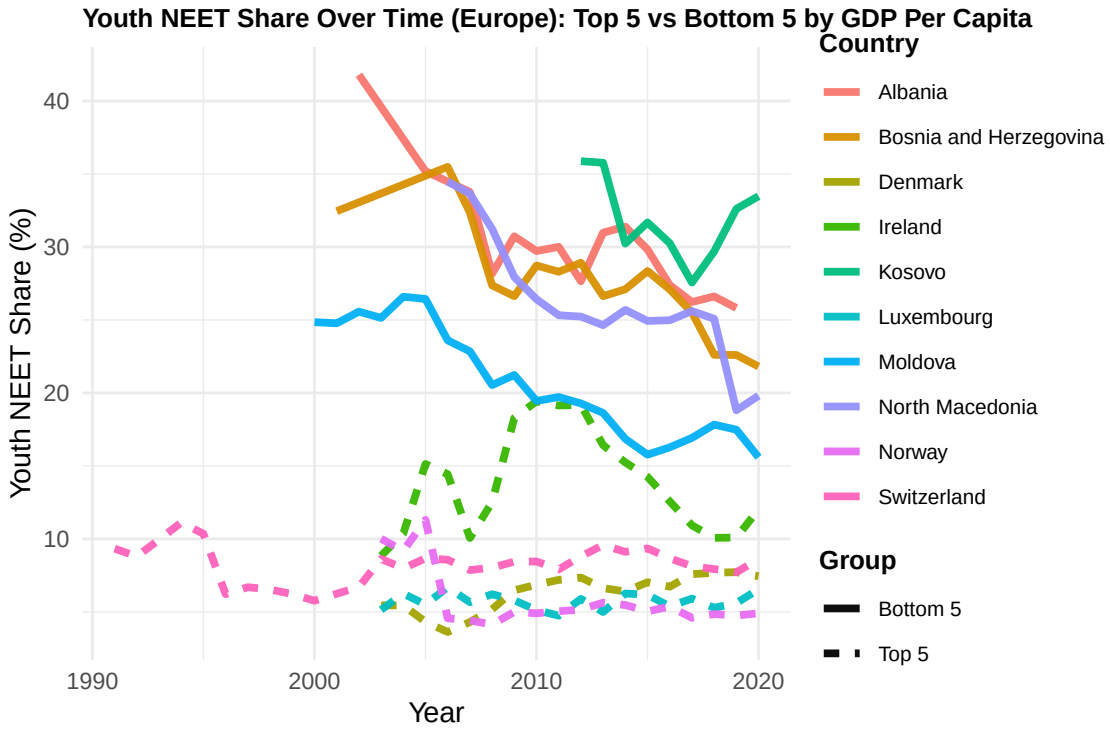
```
### Summarise change up to 2020 by continent
### For each continent, we:
# - Find the earliest year with data
# - Find 2020 (or the latest year <= 2020 if 2020 is missing)
# - Compute absolute and percentage change

earliest_neet <- youth_sum %>%
  group_by(Continent) %>%
  slice_min(Year, n = 1, with_ties = FALSE) %>%
  rename(
    earliest_year = Year,
    neet_earliest = mean_neet
  ) %>%
  ungroup()

latest_neet <- youth_sum %>%
  group_by(Continent) %>%
  slice_max(Year, n = 1, with_ties = FALSE) %>%
  rename(
    latest_year = Year,
    neet_latest = mean_neet
  ) %>%
  ungroup()

neet_change_summary <- earliest_neet %>%
  inner_join(latest_neet, by = "Continent") %>%
  mutate(
```

```
    absolute_change = neet_latest - neet_earliest,
    percent_change = 100 * (neet_latest - neet_earliest) / neet_earliest
  ) %>%
  arrange(absolute_change) %>%
  select(Continent, absolute_change, percent_change, neet_earliest)
kable(neet_change_summary, format = "latex", booktabs = TRUE)
```

Continent	absolute_change	percent_change	neet_earliest
Asia	-16.473043	-42.89855	38.40
Europe	-10.158649	-46.28086	21.95
South America	-3.915000	-13.91258	28.14
North America	3.351818	17.03160	19.68
Oceania	4.592000	16.86375	27.23
Africa	4.820000	22.60788	21.32

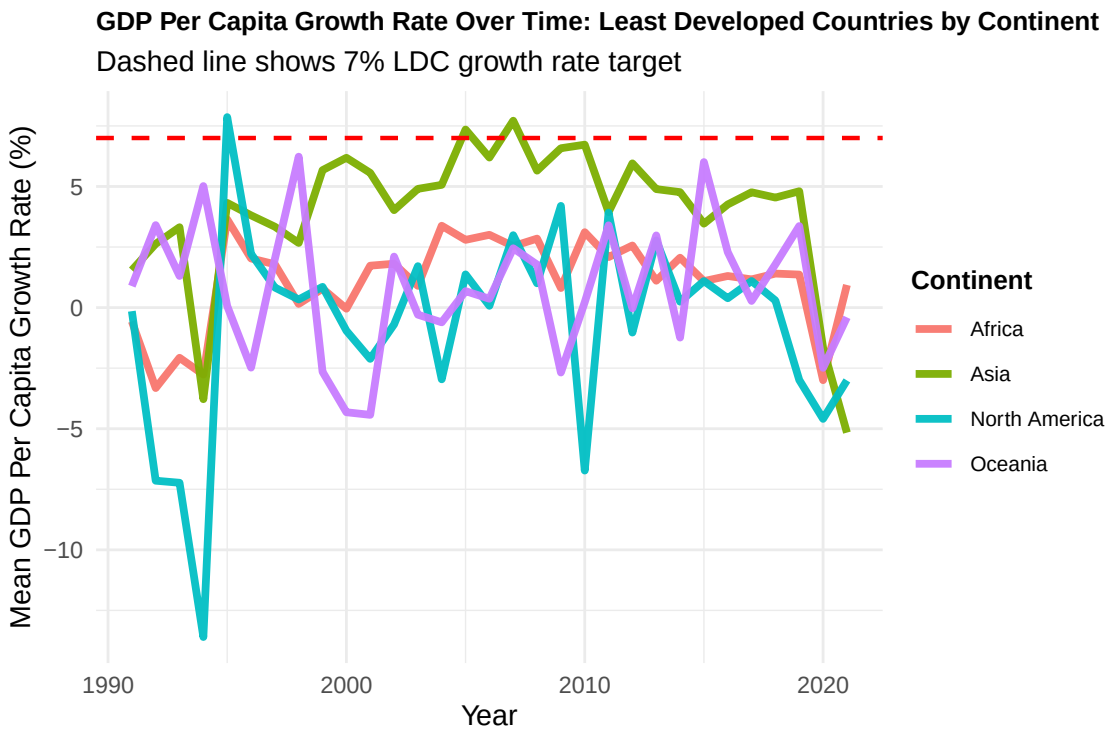
10 9) NEET in Europe: Top 5 vs Bottom 5 by GDP per capita

```
### Create line plot of NEET
### Remove outliers from visual (still used in calculating median and quartiles)
### Stat summary adds labels of median, Q1, and Q3
### Theme elements changes aesthetics
bot5top5 %>% ggplot(aes(x = Year, y = `Youth NEET Share`, color = Entity, linetype =
↪  Group, group = Entity)) +
  geom_line(size = 1.4, alpha = 0.95) +
  labs(
    title = "Youth NEET Share Over Time (Europe): Top 5 vs Bottom 5 by GDP Per Capita",
    x = "Year",
    y = "Youth NEET Share (%)",
    color = "Country"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0, size = 10, face = "bold"),
    legend.position = "right",
    legend.margin = margin(0,0,0,0),
    legend.title = element_text(size = 10, face = "bold"),
    legend.text  = element_text(size = 8)
  )
```



11 10) Least Developed Countries Target Comparison

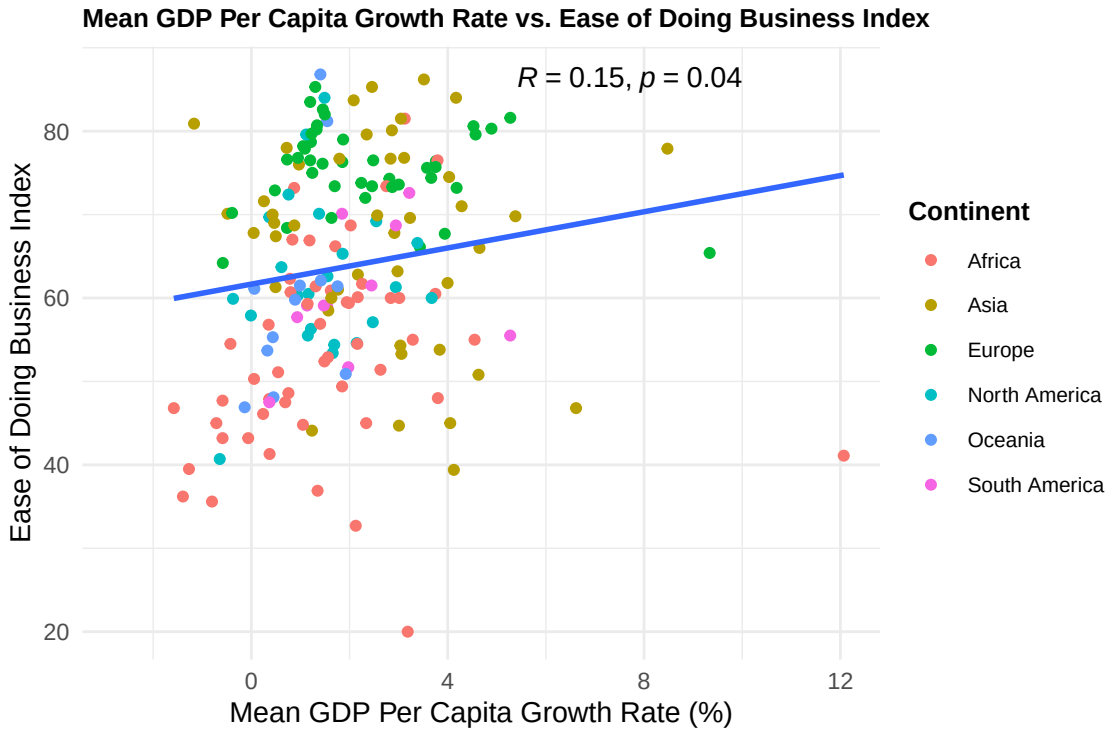
```
### Create line plot of GDP growth rate
# - Geom_hline creates dashed line, showing 7% LDC growth rate target
# - Color and legend shows continent

ldc_avg_growth %>% ggplot(aes(x = Year, y = `Avg GDP Per Capita`, color = Continent.x,
↪  group = Continent.x)) +
  geom_line(size = 1.4, alpha = 0.95) +
  geom_hline(yintercept = 7,
             linetype = "dashed",
             color = "red",
             size = 0.8) +
  labs(
    title = "GDP Per Capita Growth Rate Over Time: Least Developed Countries by
    ↪  Continent",
    subtitle = "Dashed line shows 7% LDC growth rate target",
    x = "Year",
    y = "Mean GDP Per Capita Growth Rate (%)",
    color = "Continent"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0, size = 10, face = "bold"),
    legend.position = "right",
    legend.margin = margin(0,0,0,0),
    legend.title = element_text(size = 10, face = "bold"),
    legend.text  = element_text(size = 8)
  )
```



12 11) Ease of Doing Business and GDP Per Capita Growth Rate Graph

```
### Scatterplot of GDP Per Capita Growth Rate and Ease of Doing Business
# - geom_smooth adds linear regression line of best fit
# - stat_cor adds p-value and Pearson correlation coefficient
growth_rate_country_eodb %>% ggplot(aes(x = `Mean GDP Per Capita Growth Rate`, y =
↪  Ease.of.Doing.Business, color = Continent.x, group = Continent.x)) +
  geom_point() +
  geom_smooth(aes(x = `Mean GDP Per Capita Growth Rate`, y = Ease.of.Doing.Business),
  ↪  method = "lm", se = FALSE, inherit.aes = FALSE) +
  labs(
    title = "Mean GDP Per Capita Growth Rate vs. Ease of Doing Business Index",
    x = "Mean GDP Per Capita Growth Rate (%)",
    y = "Ease of Doing Business Index",
    color = "Continent"
  ) +
  stat_cor(
    aes(x = `Mean GDP Per Capita Growth Rate`, y = Ease.of.Doing.Business), method =
    ↪  "pearson", inherit.aes = FALSE,
    label.x.npc = 0.55,
    label.y.npc = 1.0
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0, size = 10, face = "bold"),
    legend.position = "right",
```

```
    legend.margin = margin(0,0,0,0),
    legend.title = element_text(size = 10, face = "bold"),
    legend.text  = element_text(size = 8)
  )
```



13 12) HDI and Youth NEET Share Graph

```
### Scatterplot of Youth NEET and HDI
# - geom_smooth adds linear regression line of best fit
# - stat_cor adds p-value and Pearson correlation coefficient
youth_recent %>% ggplot(aes(x = `Most Recent Youth NEET Share`, y = HDI, color =
↪  Continent, group = Continent)) +
  geom_point() +
  geom_smooth(aes(x = `Most Recent Youth NEET Share`, y = HDI), method = "lm", se =
  ↪  FALSE, inherit.aes = FALSE) +
  labs(
    title = "Youth NEET Share vs. Human Development Index (HDI)",
    x = "Most Recent Youth NEET Share (%)",
    y = "Human Development Index",
    color = "Continent"
  ) +
  stat_cor(
    aes(x = `Most Recent Youth NEET Share`, y = HDI), method = "pearson", inherit.aes =
    ↪  FALSE,
    label.x.npc = 0.35,
    label.y.npc = 1.0
  ) +
  theme_minimal() +
```

```
theme(
  plot.title = element_text(hjust = 0, size = 10, face = "bold"),
  legend.position = "right",
  legend.margin = margin(0,0,0,0),
  legend.title = element_text(size = 10, face = "bold"),
  legend.text  = element_text(size = 8)
)
```