C10 IDS MIDTERM

Elisa, Fiona, Mae, Meshack, Rinta

2025-12-05

Hello. This RMarkdown document will compile all codes used into one singular document, reproducible from scratch.

Cleaning & Compiling Data - Mae

Before any analysis was possible, the data needed to be organised, cleaned, merged and compiled. Here is the code that made this possible.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.6
## v forcats   1.0.1     v stringr   1.5.2
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
# CLEANING THE YOUTH DATA SET
youth_raw <- read.csv("youth-not-in-education-employment-training.csv")

youth <- youth_raw %>%
  clean_names() #convert column names to snake case


youth_clean <- youth %>% #renaming the columns
```

```
  rename(
    country = entity,
    youth_NIEET = share_of_youth_not_in_education_employment_or_training_total_of_youth_population
  )


# str(youth_clean) #check structure of current data set


youth_clean <- youth_clean %>% #removing rows that are missing or values which are impossible
  filter(!is.na(youth_NIEET),
         youth_NIEET >= 0, youth_NIEET <= 100)


youth_clean %>% #check for duplicates - we want only one observation per country per year
  count(country, year) %>%
  filter(n > 1)
```

```
## [1] country year    n
## <0 rows> (or 0-length row.names)
```

```
# CLEANING THE GDPC DATA SET

gdp_raw <- read_csv("gdp-per-capita-worldbank.csv", show_col_types = FALSE)

gdp <- gdp_raw %>% #convert column names to snake case
  clean_names()

gdp_clean <- gdp %>% #renaming the columns
  rename(
    country = entity,
    gdp_pc_2017 = gdp_per_capita_ppp_constant_2017_international
  )

#str(gdp_clean) #check structure of data set

gdp_clean <- gdp_clean %>%  #remove impossible or missing values
  filter(!is.na(gdp_pc_2017),
         gdp_pc_2017 > 0)

gdp_clean %>% #check for duplicates
  count(country, year) %>%
  filter(n > 1)
```

```
## # A tibble: 0 x 3
## # i 3 variables: country <chr>, year <dbl>, n <int>
```

```
# CLEANING THE CONTINENTS DATA SET

continents_raw <- read.csv("continents-according-to-our-world-in-data.csv")

continents_clean <- continents_raw %>% #convert column names to snake case
```

```
  clean_names()

continents_clean <- continents_clean %>% #renaming columns
  rename(
    country   = entity,
  )


continents_clean <- continents_clean %>% #removing the year because year is 2015 for all
  select(country, code, continent)


continents_clean %>% #checking for duplicates
  count(country) %>%
    filter(n > 1)
```

```
## [1] country n
## <0 rows> (or 0-length row.names)
```

```
#JOINING ALL 3 DATA SETS

full_data <- gdp_clean %>%
  left_join(continents_clean, by = c("country", "code")) %>%
  left_join(youth_clean, by = c("country", "year", "code"))


# CREATING A NEW FULL DATA SET WITH NO NA's

full_data_no_nas <- full_data %>%
  filter(!is.na(youth_NIEET))



write_csv(full_data_no_nas, "full_data_no_nas.csv")
write_csv(gdp_clean, "gdp_clean.csv")
write_csv(youth_clean, "youth_clean.csv")
write_csv(continents_clean, "continents_clean.csv")
```

QUESTION ONE - Rinta & Meshack

Average GDP Growth Rate for all continents:

First, we must prepare all the data files which we will be using.

```
library(tidyverse)
library(ggplot2)
library(dplyr)
library(tidyr)

full_data <- read.csv("full_data.csv")

continent_growth <- full_data %>%
```

```
  group_by(continent, year) %>%
  summarise(
    cont_avg_gdp_growth = mean(gdp_growth, na.rm = TRUE)
  )
```

```
## `summarise()` has grouped output by 'continent'. You can override using the
## `.groups` argument.
```

```
country_growth_africa <- full_data %>%
  filter(continent == "Africa")

continent_growth_africa <- continent_growth %>%
  filter(continent == "Africa")


country_growth_asia <- full_data %>%
  filter(continent == "Asia")

continent_growth_asia <- continent_growth %>%
  filter(continent == "Asia")


country_growth_Europe <- full_data %>%
  filter(continent == "Europe")

continent_growth_Europe <- continent_growth %>%
  filter(continent == "Europe")


country_growth_North_America <- full_data %>%
  filter(continent == "North America")

continent_growth_North_America <- continent_growth %>%
  filter(continent == "North America")


country_growth_Oceania <- full_data %>%
  filter(continent == "Oceania")

continent_growth_Oceania <- continent_growth %>%
  filter(continent == "Oceania")


country_growth_South_America <- full_data %>%
  filter(continent == "South America")

continent_growth_South_America <- continent_growth %>%
  filter(continent == "South America")
```

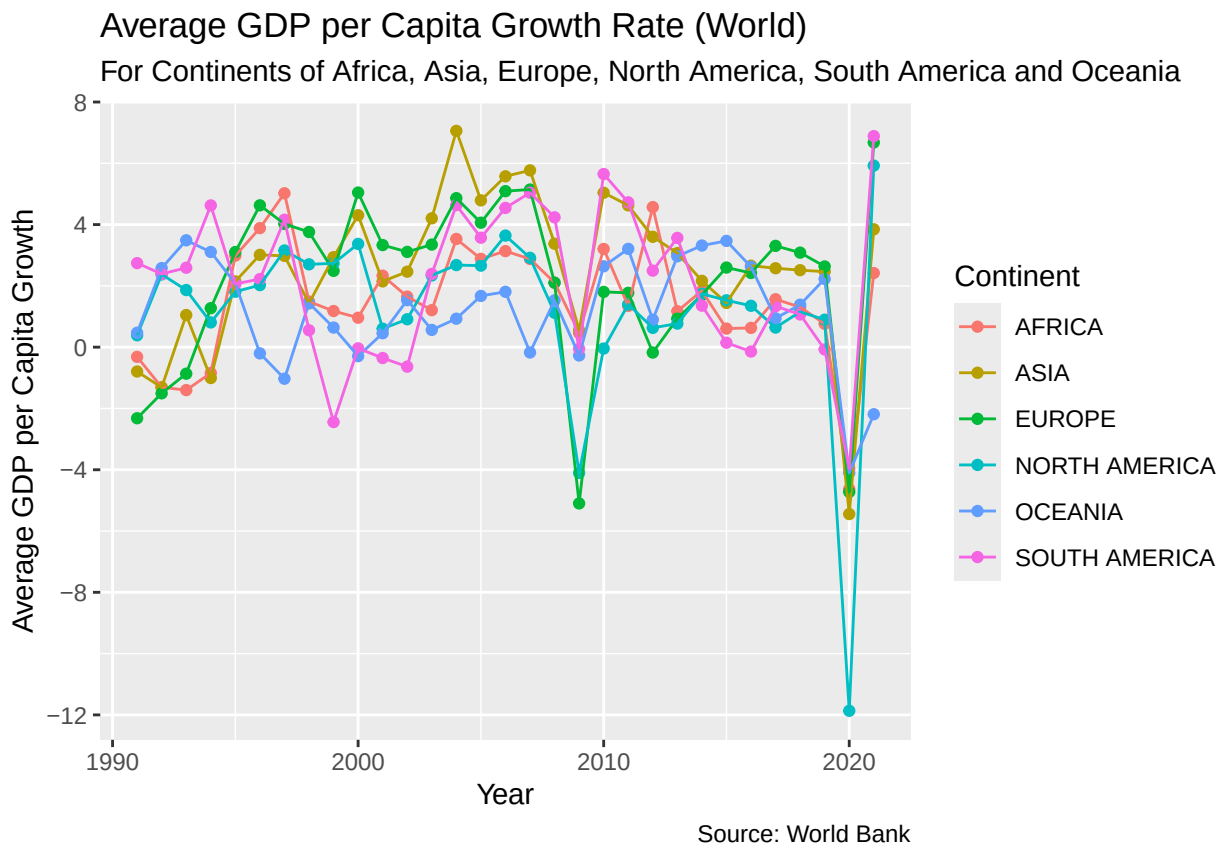
```r
# Create output directory
output_dir <- "continent_data"
dir.create(output_dir, showWarnings = FALSE)

# Save ONLY the 12 CSV files
write_csv(continent_growth, file.path(output_dir, "continent_growth.csv"))

write_csv(country_growth_africa, file.path(output_dir, "country_growth_africa.csv"))
write_csv(continent_growth_africa, file.path(output_dir, "continent_growth_africa.csv"))

write_csv(country_growth_asia, file.path(output_dir, "country_growth_asia.csv"))
write_csv(continent_growth_asia, file.path(output_dir, "continent_growth_asia.csv"))

write_csv(country_growth_Europe, file.path(output_dir, "country_growth_europe.csv"))
write_csv(continent_growth_Europe, file.path(output_dir, "continent_growth_europe.csv"))

write_csv(country_growth_North_America, file.path(output_dir, "country_growth_north_america.csv"))
write_csv(continent_growth_North_America, file.path(output_dir, "continent_growth_north_america.csv"))

write_csv(country_growth_Oceania, file.path(output_dir, "country_growth_oceania.csv"))
write_csv(continent_growth_Oceania, file.path(output_dir, "continent_growth_oceania.csv"))

write_csv(country_growth_South_America, file.path(output_dir, "country_growth_south_america.csv"))
write_csv(continent_growth_South_America, file.path(output_dir, "continent_growth_south_america.csv"))
```

Now, we can move into plotting these graphs,

```r
library(tidyverse)
library(ggplot2)
library(dplyr)
library(tidyr)
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.5.2
```

```r
p <- ggplot(data = na.omit(continent_growth), aes(year, cont_avg_gdp_growth, col = continent)) +
  geom_point() +
  geom_line() + labs(title = "Average GDP per Capita Growth Rate (World)", subtitle = "For Continents o
  scale_color_discrete(labels = toupper, name = "Continent")

p
```



We are now going to compare each continent, including and not including its respective "Least Developed" countries. We took this list of countries from the UN Trade and Development database (UNCTAD), which is linked in the references page.

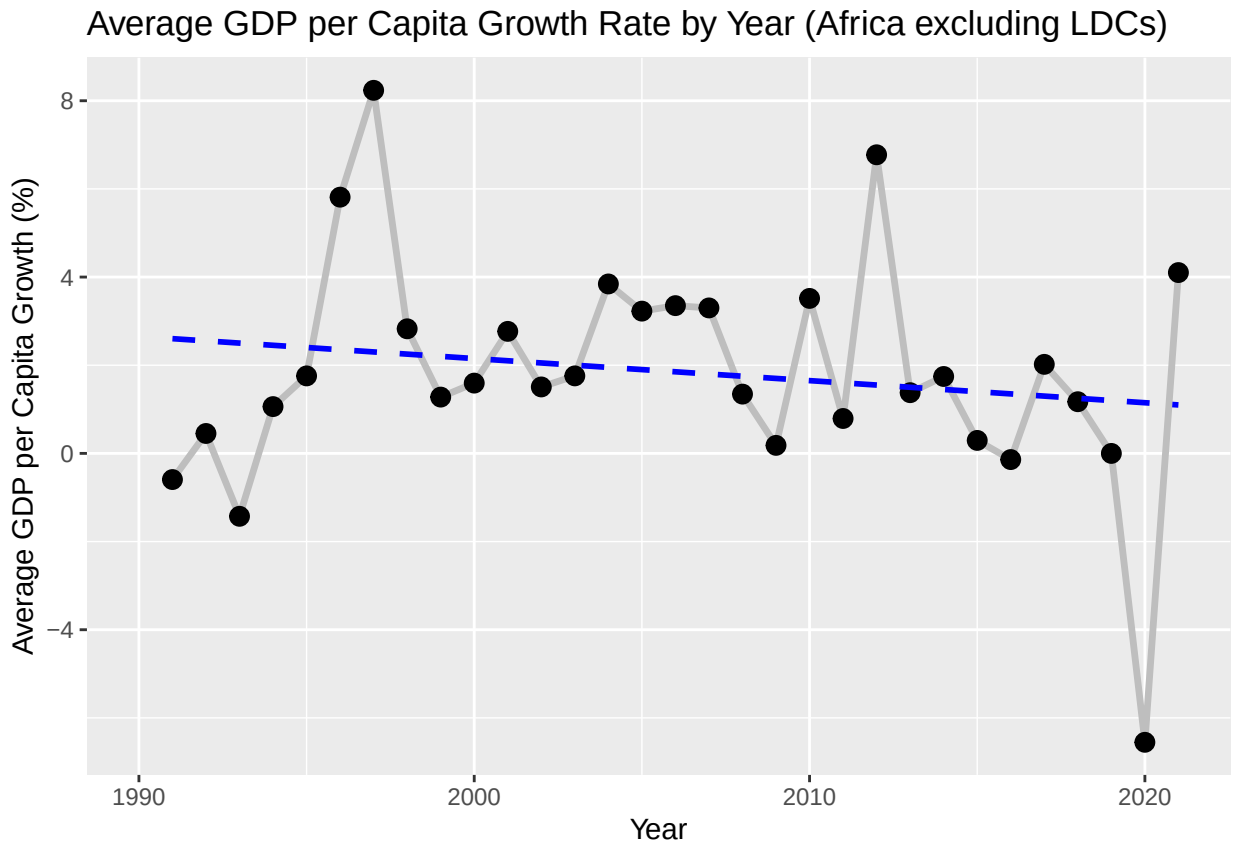
```
# WE WILL FIRST LABEL THE COUNTRIES LDC, AND THEN USE THAT IN OUR ANALYSIS
ldc_country_names <-
  c(
    "Afghanistan", "Angola", "Bangladesh", "Benin", "Burkina Faso", "Burundi", "Cambodia", "Central Afri
    "Djibouti", "Ethiopia", "Gambia", "Guinea", "Guinea-Bissau", "Haiti",
    "Kiribati", "Laos", "Lesotho", "Liberia", "Madagascar",
    "Malawi", "Mali", "Mauritania", "Mozambique", "Myanmar", "Nepal", "Niger",
    "Rwanda", "Senegal", "Sierra Leone", "Solomon Islands", "Somalia",
    "Sudan", "East Timor", "Togo", "Tuvalu", "Uganda", "Tanzania",
    "Zambia")

country_codes <- c("AFG","AGO","BGD","BEN","BFA","BDI","KHM","CAF","TCD","COM", "COD",
                   "DJI","ETH","GMB","GIN","GNB","HTI","KIR", "LAO","LSO","LBR","MDG",
                   "MWI","MLI","MRT","MOZ","MMR","NPL","NER","RWA","SEN","SLE",
                   "SLB","SOM","SDN", "TLS", "TGO","TUV","UGA", "TZA","ZMB")


ldc_countries <- data.frame(country = ldc_country_names)

ldc_countries <- ldc_countries %>%
  mutate(
    code = country_codes,
```

```
    Status = "Least Developed Country"
  )

ldc_full_data <- full_data %>% semi_join(ldc_countries, join_by(code))
```

Now with the countries assigned and labelled, we can graph out the plots so we can use *exploratory data analysis* on our continents, and compare them fairly.

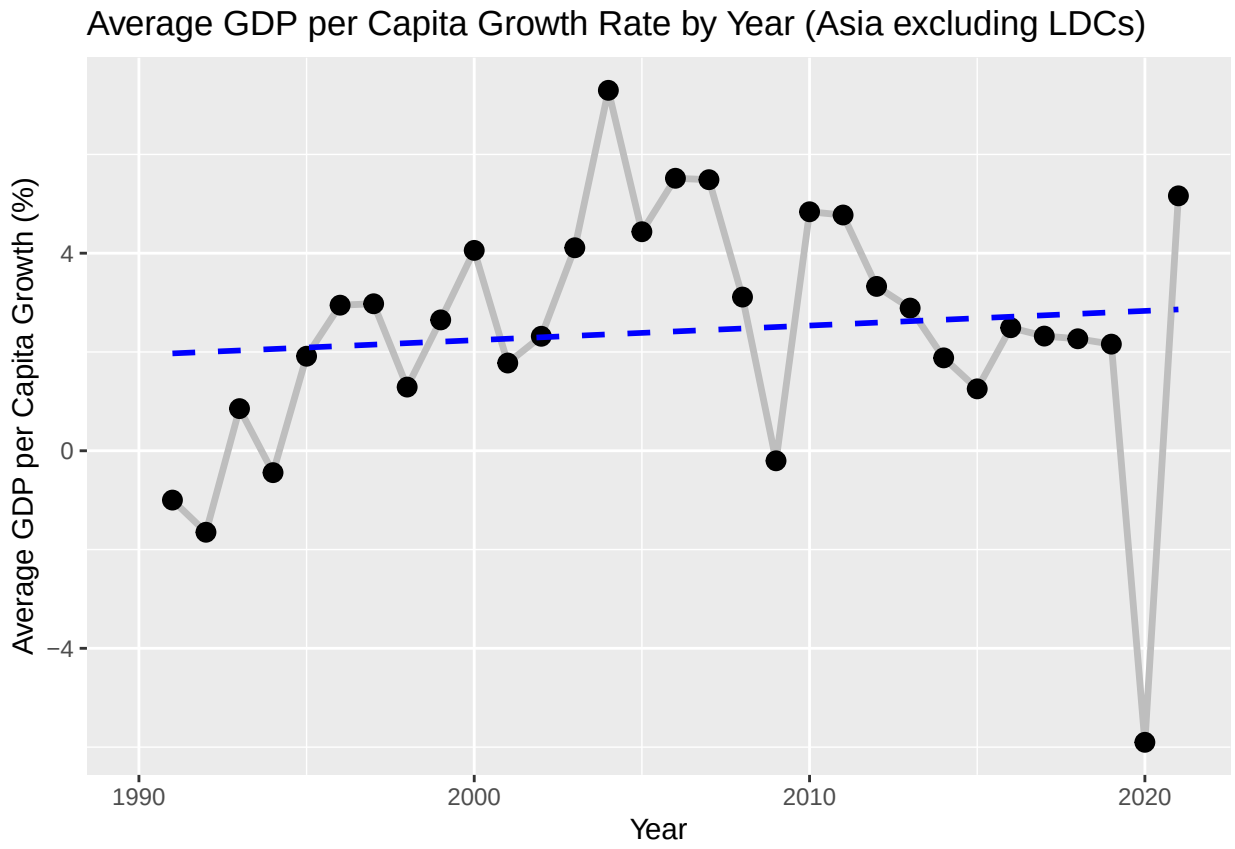
```
#In this R document, Meshack will ascertain growth rates for continents excluding LDCs.
library(dplyr)
library(tidyverse)
library(ggplot2)

#CONTINENTAL AVERAGE GROWTH RATE EXCLUDING LDCs

noldc_full_data <- full_data %>% anti_join(ldc_countries, join_by(code))

#AFRICA

africa_noLDC <- anti_join(country_growth_africa, ldc_countries, by = "country")

africa_avg_growth_by_year <- africa_noLDC %>%
  group_by(year) %>%
  summarise(Average_Growth = mean(gdp_growth, na.rm = TRUE))

africa_plot_noLDC <- ggplot(africa_avg_growth_by_year, aes(x = year, y = Average_Growth)) +
  geom_line(colour = "grey",size = 1.2) +
  geom_point(colour = "black", size = 3) +
  labs(title = "Average GDP per Capita Growth Rate by Year (Africa excluding LDCs)",
       x = "Year",
       y = "Average GDP per Capita Growth (%)") + geom_smooth(method = "lm", se = FALSE, color = "blue"
africa_plot_noLDC
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



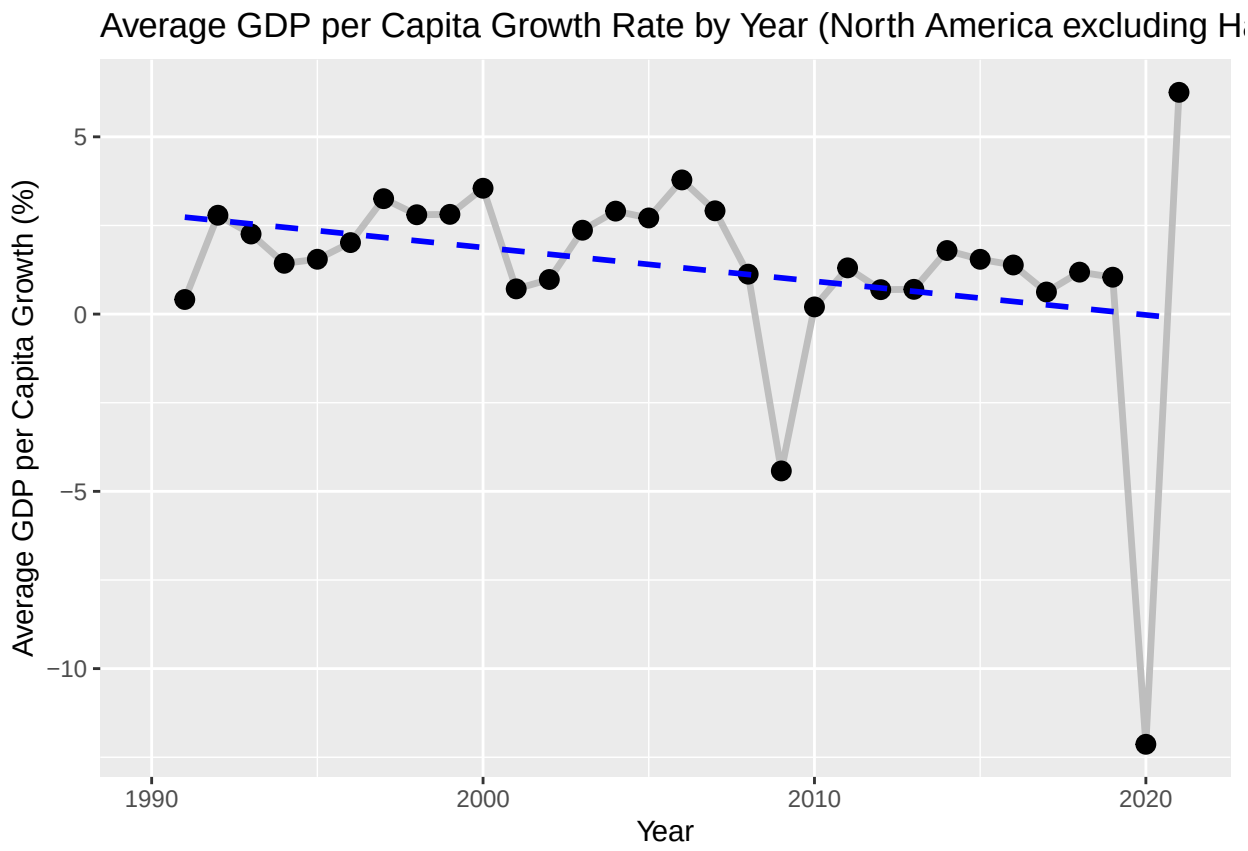

```
#ASIA

asia_noLDC <- anti_join(country_growth_asia, ldc_countries, by = "country")

asia_avg_growth_by_year <- asia_noLDC %>%
  group_by(year) %>%
  summarise(Average_Growth = mean(gdp_growth, na.rm = TRUE))

asia_plot_noLDC <- ggplot(asia_avg_growth_by_year, aes(x = year, y = Average_Growth)) +
  geom_line(colour = "grey",size = 1.2) +
  geom_point(colour = "black", size = 3) +
  labs(title = "Average GDP per Capita Growth Rate by Year (Asia excluding LDCs)",
       x = "Year",
       y = "Average GDP per Capita Growth (%)") + geom_smooth(method = "lm", se = FALSE, color = "blue"
asia_plot_noLDC
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



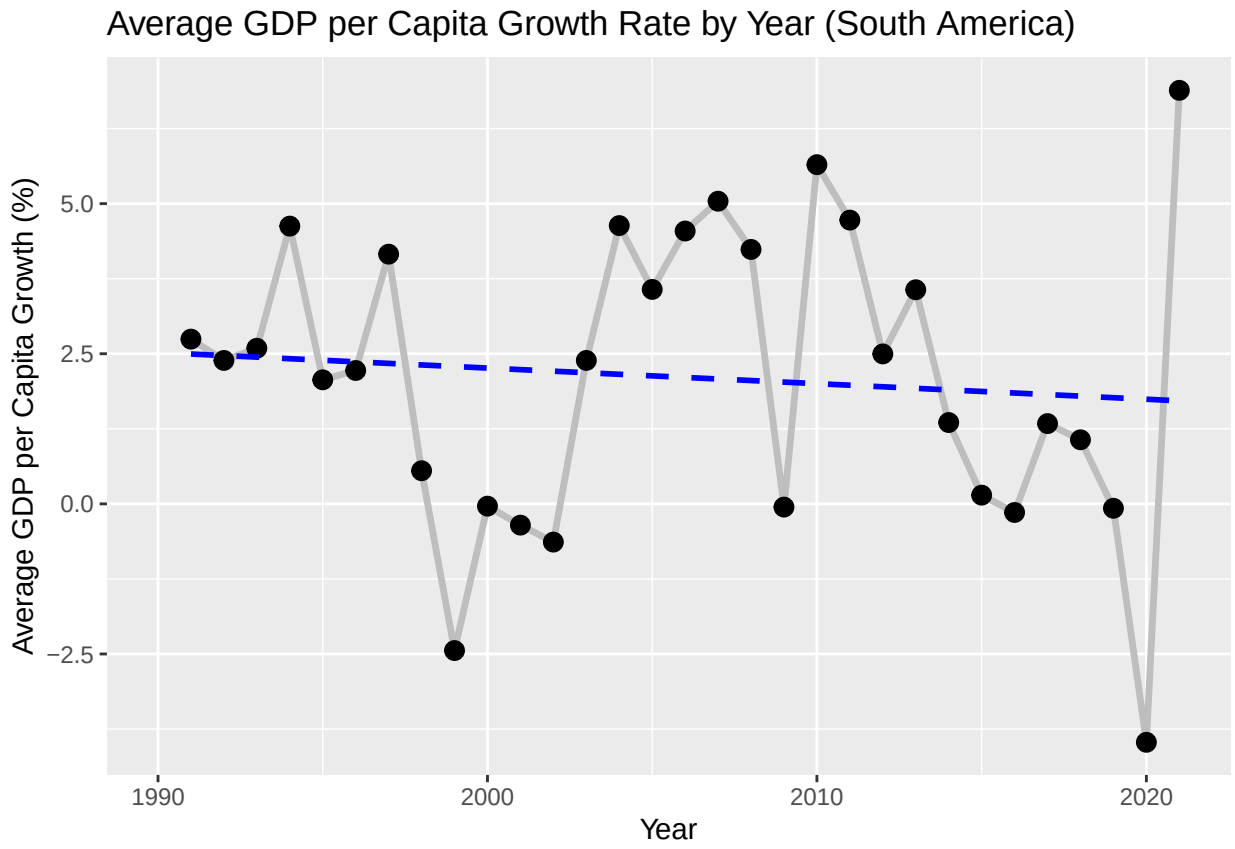

```
#FOR NORTH AMERICA, ITS ONLY HAITI NOT MENTIONED

North_America_noLDC <- anti_join(country_growth_North_America, ldc_countries, by = "country")

North_America_avg_growth_by_year <- North_America_noLDC %>%
  group_by(year) %>%
  summarise(Average_Growth = mean(gdp_growth, na.rm = TRUE))

North_America_noLDC <- ggplot(North_America_avg_growth_by_year, aes(x = year, y = Average_Growth)) +
  geom_line(colour = "grey",size = 1.2) +
  geom_point(colour = "black", size = 3) +
  labs(title = "Average GDP per Capita Growth Rate by Year (North America excluding Haiti)",
       x = "Year",
       y = "Average GDP per Capita Growth (%)") + geom_smooth(method = "lm", se = FALSE, color = "blue"
North_America_noLDC
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



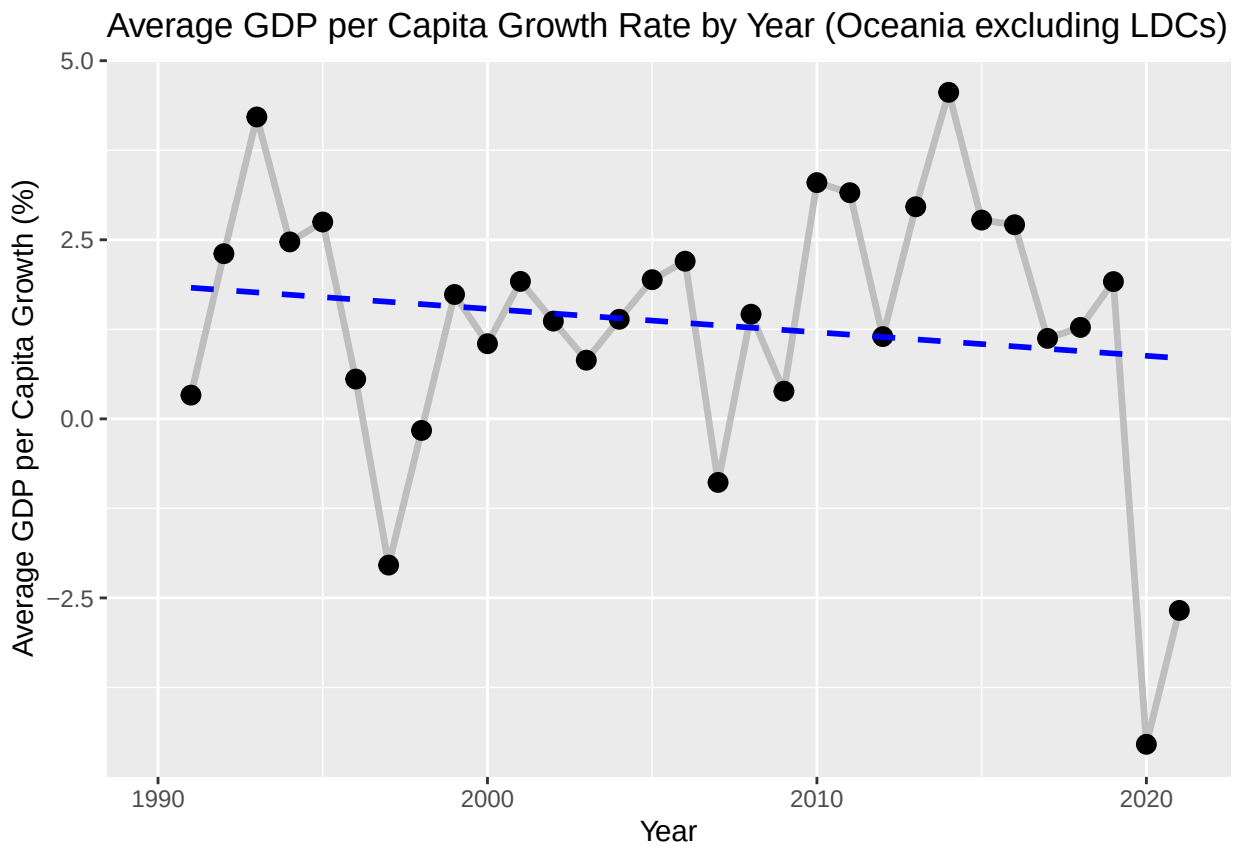

```
#South America

South_America_noLDC <- anti_join(country_growth_South_America, ldc_countries, by = "country")

South_America_avg_growth_by_year <- South_America_noLDC %>%
  group_by(year) %>%
  summarise(Average_Growth = mean(gdp_growth, na.rm = TRUE))

South_America_noLDC <- ggplot(South_America_avg_growth_by_year, aes(x = year, y = Average_Growth)) +
  geom_line(colour = "grey",size = 1.2) +
  geom_point(colour = "black", size = 3) +
  labs(title = "Average GDP per Capita Growth Rate by Year (South America)",
       x = "Year",
       y = "Average GDP per Capita Growth (%)") + geom_smooth(method = "lm", se = FALSE, color = "blue"
South_America_noLDC
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



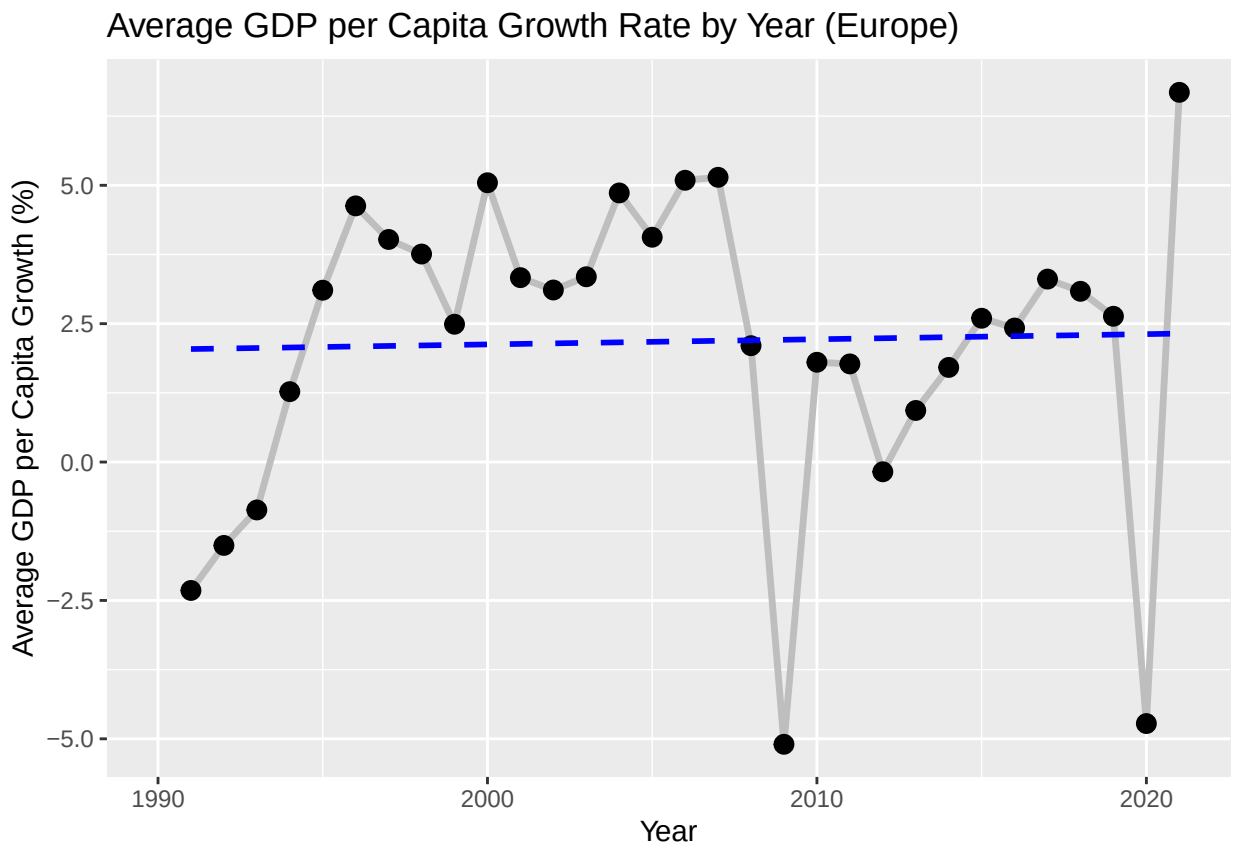

```
#OCEANIA

oceania_noLDC <- anti_join(country_growth_Oceania, ldc_countries, by = "country")

oceania_avg_growth_by_year <- oceania_noLDC %>%
  group_by(year) %>%
  summarise(Average_Growth = mean(gdp_growth, na.rm = TRUE))

oceania_plot_noLDC <- ggplot(oceania_avg_growth_by_year, aes(x = year, y = Average_Growth)) +
  geom_line(colour = "grey",size = 1.2) +
  geom_point(colour = "black", size = 3) +
  labs(title = "Average GDP per Capita Growth Rate by Year (Oceania excluding LDCs)",
       x = "Year",
       y = "Average GDP per Capita Growth (%)") + geom_smooth(method = "lm", se = FALSE, color = "blue"
oceania_plot_noLDC
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# EUROPE

europe_noLDC <- anti_join(country_growth_Europe, ldc_countries, by = "country")

europe_avg_growth_by_year <- europe_noLDC %>%
  group_by(year) %>%
  summarise(Average_Growth = mean(gdp_growth, na.rm = TRUE))

europe_plot_noLDC <- ggplot(europe_avg_growth_by_year, aes(x = year, y = Average_Growth)) +
  geom_line(colour = "grey",size = 1.2) +
  geom_point(colour = "black", size = 3) +
  labs(title = "Average GDP per Capita Growth Rate by Year (Europe)",
       x = "Year",
       y = "Average GDP per Capita Growth (%)") + geom_smooth(method = "lm", se = FALSE, color = "blue"
europe_plot_noLDC
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# avg noLDC

continent_noldc_growth <- noldc_full_data %>%
  group_by(continent, year) %>%
  summarise(
    cont_avg_gdp_growth = mean(gdp_growth, na.rm = TRUE)
  )
```

```
## `summarise()` has grouped output by 'continent'. You can override using the
## `.groups` argument.
```

```
# Taking the liberty to analyse Pearson regression statistics
print(cor(africa_avg_growth_by_year$year, africa_avg_growth_by_year$Average_Growth, use = "complete.obs
```

```
## [1] -0.1739771
```

```
print(cor(asia_avg_growth_by_year$year, asia_avg_growth_by_year$Average_Growth, use = "complete.obs", m
```

```
## [1] 0.1065993
```

```
print(cor(North_America_avg_growth_by_year$year, North_America_avg_growth_by_year$Average_Growth, use =
```

```
## [1] -0.2863833
```

```r
print(cor(South_America_avg_growth_by_year$year, South_America_avg_growth_by_year$Average_Growth, use =
```

```
## [1] -0.09580125
```

```r
print(cor(oceania_avg_growth_by_year$year, oceania_avg_growth_by_year$Average_Growth, use = "complete.ob
```

```
## [1] -0.1550122
```

```r
print(cor(europe_avg_growth_by_year$year, europe_avg_growth_by_year$Average_Growth, use = "complete.obs
```

```
## [1] 0.0307284
```

We then compare these same plots to those continents including LDCs, and we want to analyse specifically those LDC nations. But first, a bit of data preparation.

```r
library(ggpubr)



continent_ldc_growth <- ldc_full_data %>%
  group_by(continent, year) %>%
  summarise(
    cont_avg_gdp_growth = mean(gdp_growth, na.rm = TRUE)
  )
```

```
## `summarise()` has grouped output by 'continent'. You can override using the
## `.groups` argument.
```

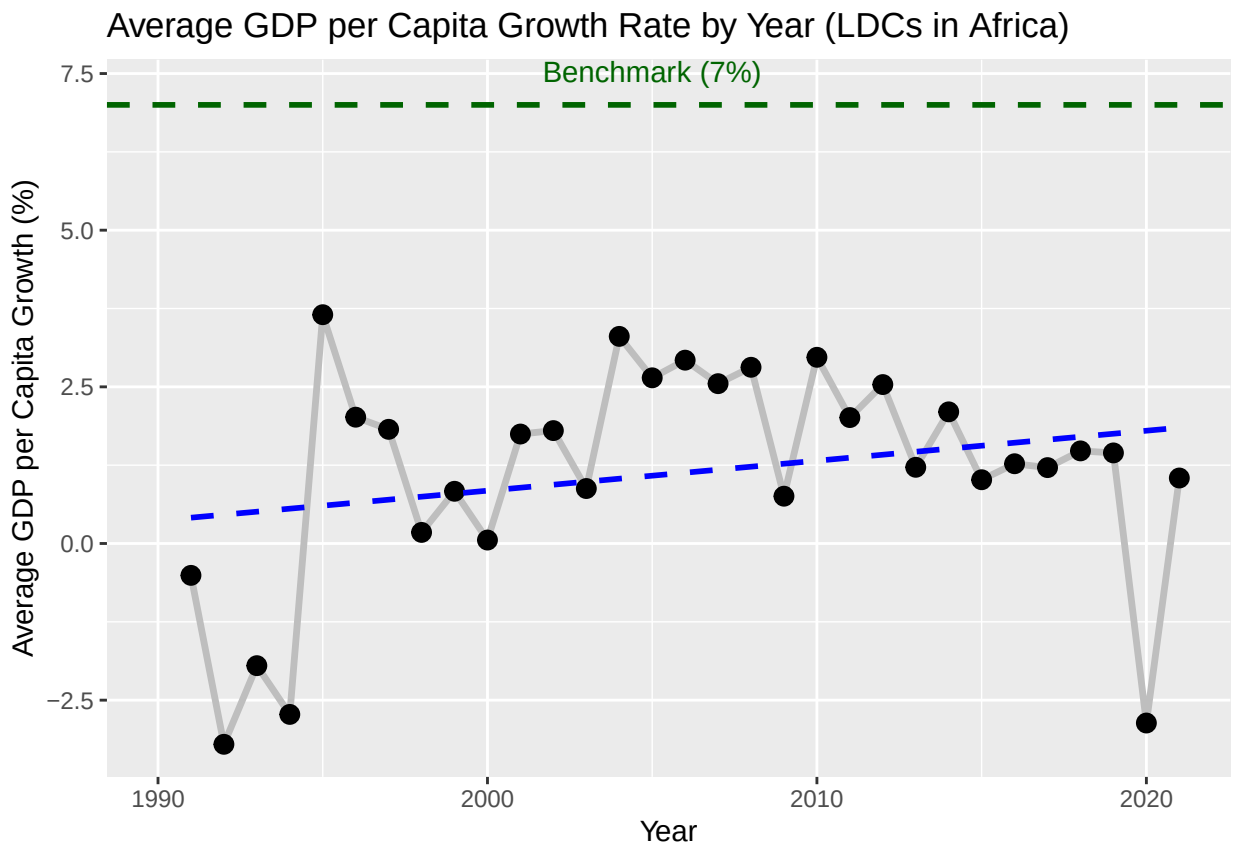
```r
ldc_growth_africa <- country_growth_africa %>% semi_join(ldc_countries, join_by(code))

continent_ldc_growth_africa <- continent_ldc_growth %>%
  filter(continent == "Africa")



ldc_growth_asia <- country_growth_asia %>% semi_join(ldc_countries, join_by(code))

continent_ldc_growth_asia <- continent_ldc_growth %>%
  filter(continent == "Asia")



### no LDC in europe



ldc_growth_North_America <- country_growth_North_America %>% semi_join(ldc_countries, join_by(code))

continent_ldc_growth_North_America <- continent_ldc_growth %>%
  filter(continent == "North America")
```

```r
ldc_growth_Oceania <- country_growth_Oceania %>% semi_join(ldc_countries, join_by(code))

continent_ldc_growth_Oceania <- continent_ldc_growth %>%
  filter(continent == "Oceania")
```

Now to create the plots,

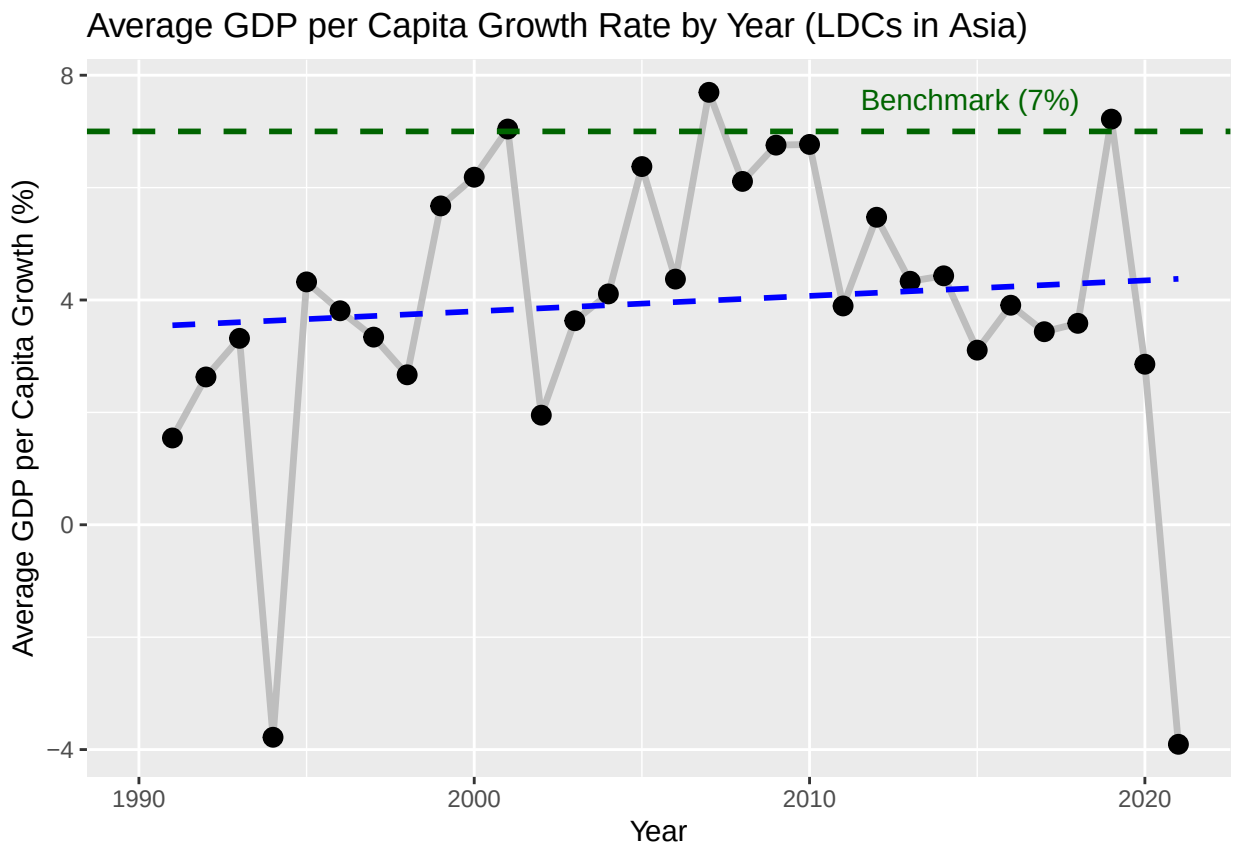
```r
#In this R Document, we will ascertain GDP per capita growth in Continents/countries considered LDC (Le
library(dplyr)
library(tidyverse)
library(ggplot2)


#AFRICA (LDCs only)


ldc_africa_df <- read.csv("continent_ldc_growth_africa.csv")

ldc_africa_avg_growth_by_year <- ldc_africa_df %>%
  group_by(year) %>%
  summarise(Average_Growth = mean(cont_avg_gdp_growth, na.rm = TRUE))

ldc_africa_plot <- ggplot(ldc_africa_avg_growth_by_year, aes(x = year, y = Average_Growth)) +
  geom_line(colour = "grey",size = 1.2) +
  geom_point(colour = "black", size = 3) +
  labs(title = "Average GDP per Capita Growth Rate by Year (LDCs in Africa)",
       x = "Year",
       y = "Average GDP per Capita Growth (%)") + geom_smooth(method = "lm", se = FALSE, color = "blue"
  geom_hline(yintercept = 7, linetype = "dashed", color = "darkgreen", size = 1) +
  annotate("text", x = 2005, y = 7.2,
           label = "Benchmark (7%)", color = "darkgreen", vjust = -0.5)
ldc_africa_plot
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



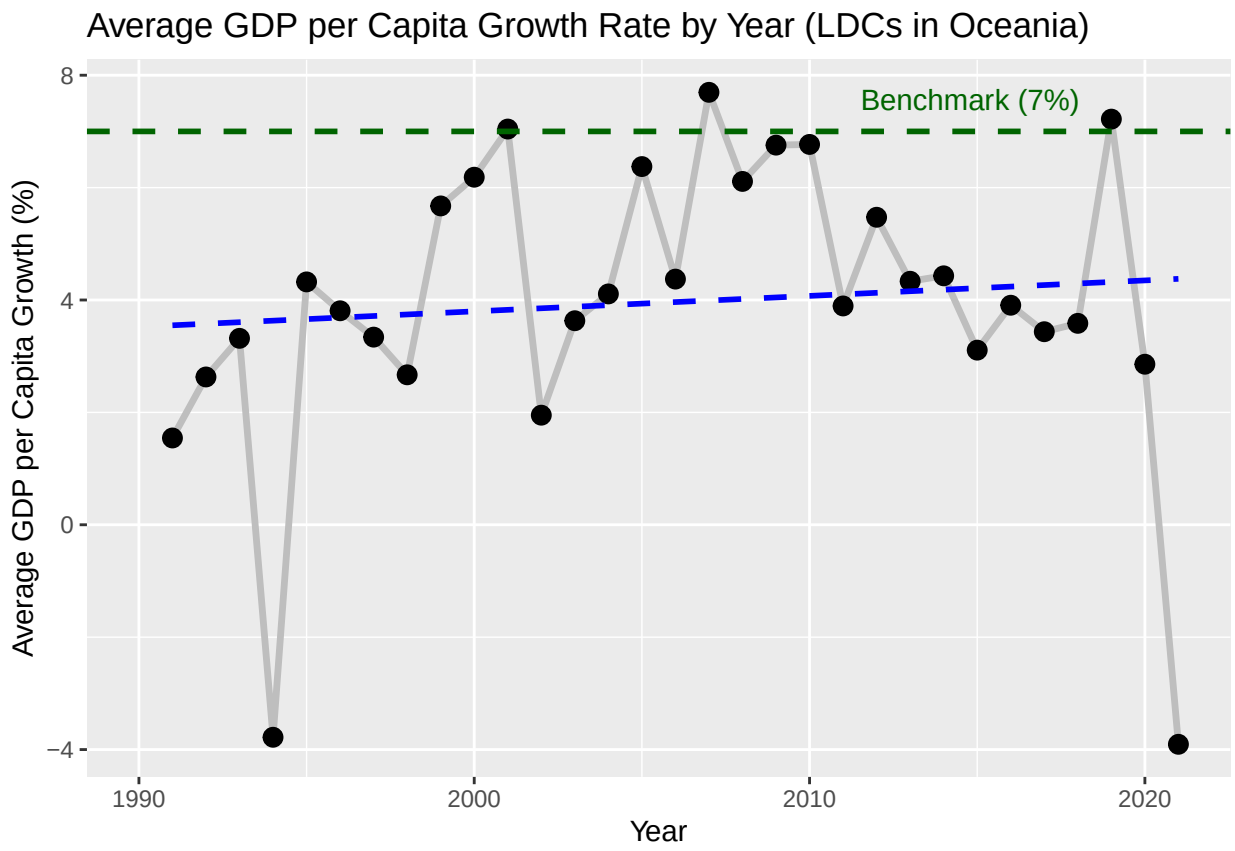

```
# ASIA (LDCs only)

ldc_asia_df <- read.csv("continent_ldc_growth_asia.csv")

ldc_asia_avg_growth_by_year <- ldc_asia_df %>%
  group_by(year) %>%
  summarise(Average_Growth = mean(cont_avg_gdp_growth, na.rm = TRUE))

ldc_asia_plot <- ggplot(ldc_asia_avg_growth_by_year, aes(x = year, y = Average_Growth)) +
  geom_line(colour = "grey",size = 1.2) +
  geom_point(colour = "black", size = 3) +
  labs(title = "Average GDP per Capita Growth Rate by Year (LDCs in Asia)",
       x = "Year",
       y = "Average GDP per Capita Growth (%)") + geom_smooth(method = "lm", se = FALSE, color = "blue"
  geom_hline(yintercept = 7, linetype = "dashed", color = "darkgreen", size = 1) +
  annotate("text", x = 2005, y = 7.2,
           label = "Benchmark (7%)", color = "darkgreen", vjust = -0.5, hjust = -1)
ldc_asia_plot
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



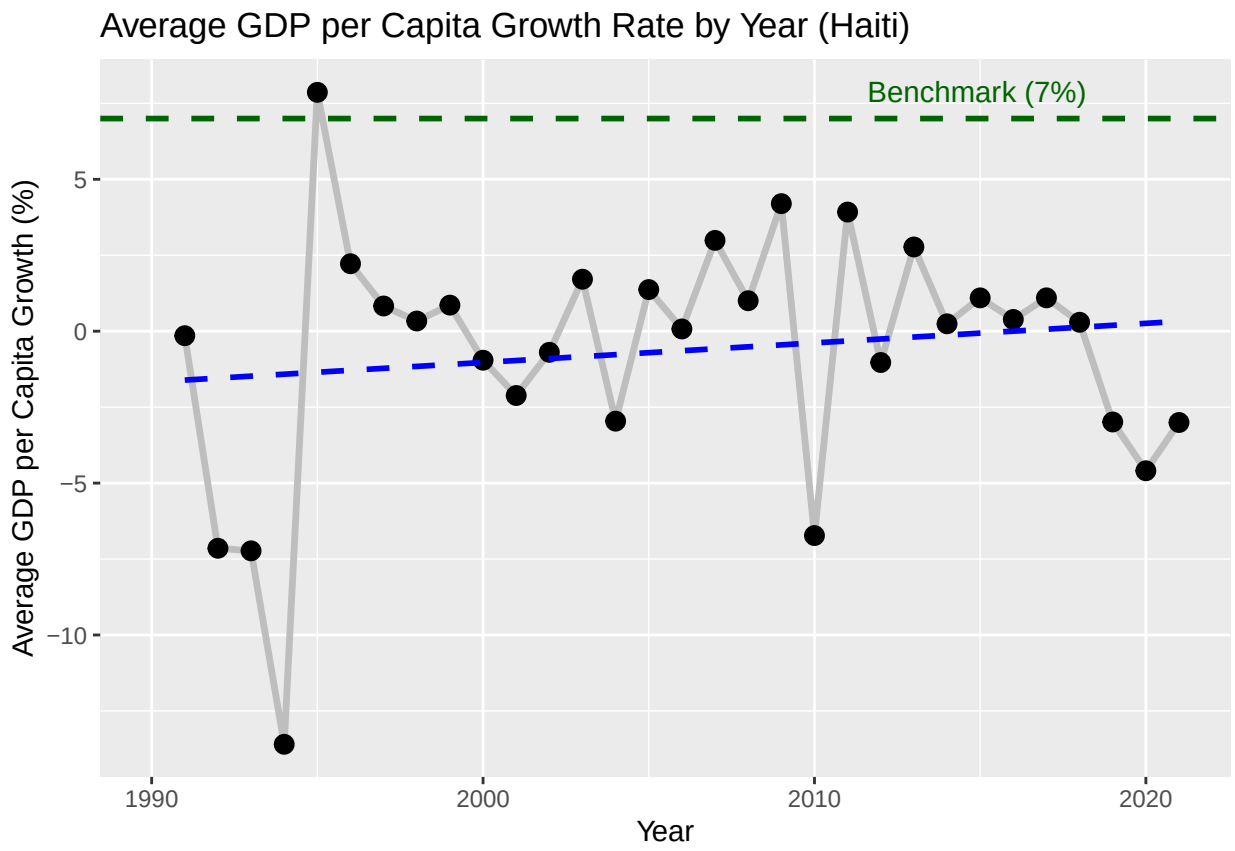

```
#OCEANIA (LDCs only)

ldc_oceania_df <- read.csv("continent_ldc_growth_asia.csv")

ldc_oceania_avg_growth_by_year <- ldc_oceania_df %>%
  group_by(year) %>%
  summarise(Average_Growth = mean(cont_avg_gdp_growth, na.rm = TRUE))

ldc_oceania_plot <- ggplot(ldc_oceania_avg_growth_by_year, aes(x = year, y = Average_Growth)) +
  geom_line(colour = "grey",size = 1.2) +
  geom_point(colour = "black", size = 3) +
  labs(title = "Average GDP per Capita Growth Rate by Year (LDCs in Oceania)",
       x = "Year",
       y = "Average GDP per Capita Growth (%)") + geom_smooth(method = "lm", se = FALSE, color = "blue"
  geom_hline(yintercept = 7, linetype = "dashed", color = "darkgreen", size = 1) +
  annotate("text", x = 2005, y = 7.2,
           label = "Benchmark (7%)", color = "darkgreen", vjust = -0.5, hjust = -1)
ldc_oceania_plot
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
#Haiti (Only North American LDC)

haiti_df <- read.csv("continent_ldc_growth_North.csv")

haiti_avg_growth_by_year <- haiti_df %>%
  group_by(year) %>%
  summarise(Average_Growth = mean(cont_avg_gdp_growth, na.rm = TRUE))

haiti_plot <- ggplot(haiti_avg_growth_by_year, aes(x = year, y = Average_Growth)) +
  geom_line(colour = "grey",size = 1.2) +
  geom_point(colour = "black", size = 3) +
  labs(title = "Average GDP per Capita Growth Rate by Year (Haiti)",
       x = "Year",
       y = "Average GDP per Capita Growth (%)") + geom_smooth(method = "lm", se = FALSE, color = "blue"
  geom_hline(yintercept = 7, linetype = "dashed", color = "darkgreen", size = 1) +
  annotate("text", x = 2005, y = 7.2,
           label = "Benchmark (7%)", color = "darkgreen", vjust = -0.5, hjust = -1)
haiti_plot
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Pearson Regression Analysis

print(cor(ldc_africa_avg_growth_by_year$year, ldc_africa_avg_growth_by_year$Average_Growth, use = "compl

## [1] 0.2443803

print(cor(ldc_asia_avg_growth_by_year$year, ldc_asia_avg_growth_by_year$Average_Growth, use = "complete

## [1] 0.0947886

print(cor(haiti_avg_growth_by_year$year, haiti_avg_growth_by_year$Average_Growth, use = "complete.obs",

## [1] 0.1430906

print(cor(ldc_oceania_avg_growth_by_year$year, ldc_oceania_avg_growth_by_year$Average_Growth, use = "co

## [1] 0.0947886
```

Now with the plots created and (Pearson) correlation coefficients calculated, we can dive deeper and look for more interesting statistics which could be useful for our analysis.

In order to see trends in volatility, we have the code for plotting 5 year rolling standard deviations to compare standard deviations in our continents.

First, we will calculate the data for the standard deviations.

```r
library(tidyverse)
library(ggplot2)
library(dplyr)
library(tidyr)
library(ggpubr)


cal_5yr_sd <- function(avg_growth) {
  result <- rep(NA, length(avg_growth))

  for (i in 1:length(avg_growth)) {
    if(i >= 5){
      span <- avg_growth[(i-4):i]
      if(!any(is.na(span))){
        result[i] <- sd(span)
      }
    }
  }

  return(result)
}

sd_5yr_all <- continent_growth %>%
  group_by(continent) %>%
  mutate(sd = cal_5yr_sd(cont_avg_gdp_growth))
sd_5yr_all
```

```
## # A tibble: 224 x 4
## # Groups:   continent [7]
##    continent  year cont_avg_gdp_growth    sd
##    <chr>     <int>               <dbl> <dbl>
##  1 Africa     1990             NaN     NA
##  2 Africa     1991              -0.316 NA
##  3 Africa     1992              -1.31  NA
##  4 Africa     1993              -1.40  NA
##  5 Africa     1994              -0.843 NA
##  6 Africa     1995               2.99   1.82
##  7 Africa     1996               3.89   2.56
##  8 Africa     1997               5.02   2.88
##  9 Africa     1998               1.49   2.28
## 10 Africa     1999               1.18   1.62
## # i 214 more rows
```

```r
sd_5yr_ldc <- continent_ldc_growth %>%
  group_by(continent) %>%
  mutate(sd = cal_5yr_sd(cont_avg_gdp_growth))
sd_5yr_ldc
```

```
## # A tibble: 128 x 4
## # Groups:   continent [4]
##    continent  year cont_avg_gdp_growth    sd
##    <chr>     <int>               <dbl> <dbl>
```

```
##  1 Africa     1990             NaN     NA
##  2 Africa     1991              -0.508 NA
##  3 Africa     1992              -3.20  NA
##  4 Africa     1993              -1.95  NA
##  5 Africa     1994              -2.73  NA
##  6 Africa     1995               3.65   2.77
##  7 Africa     1996               2.02   3.08
##  8 Africa     1997               1.82   2.76
##  9 Africa     1998               0.177  2.41
## 10 Africa     1999               0.833  1.32
## # i 118 more rows
```

```
sd_5yr_noldc <- continent_noldc_growth %>%
  group_by(continent) %>%
  mutate(sd = cal_5yr_sd(cont_avg_gdp_growth))
sd_5yr_noldc
```

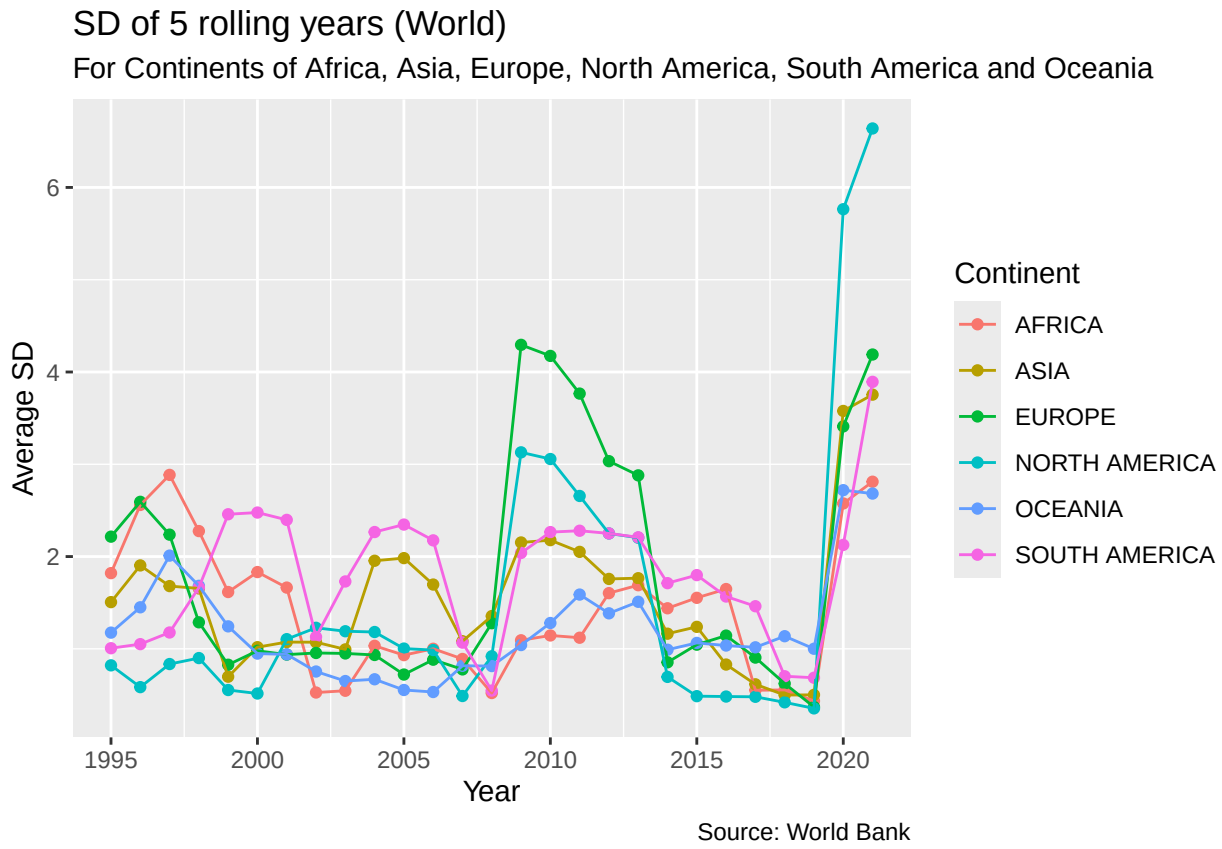
```
## # A tibble: 224 x 4
## # Groups:   continent [7]
##    continent  year cont_avg_gdp_growth    sd
##    <chr>     <int>               <dbl> <dbl>
##  1 Africa     1990            NaN      NA
##  2 Africa     1991             -0.0572 NA
##  3 Africa     1992              1.25   NA
##  4 Africa     1993             -0.662  NA
##  5 Africa     1994              1.70   NA
##  6 Africa     1995              2.09    1.18
##  7 Africa     1996              6.41    2.60
##  8 Africa     1997              9.34    4.02
##  9 Africa     1998              3.25    3.25
## 10 Africa     1999              1.65    3.26
## # i 214 more rows
```

```
# Create output directory
output_dir <- "sd_data"
dir.create(output_dir, showWarnings = FALSE)

# Save ONLY the 12 CSV files
write_csv(sd_5yr_all, file.path(output_dir, "sd_5yr_all.csv"))

write_csv(sd_5yr_ldc, file.path(output_dir, "sd_5yr_ldc.csv"))

write_csv(sd_5yr_noldc, file.path(output_dir, "sd_5yr_noldc.csv"))

# changed the name of files
```

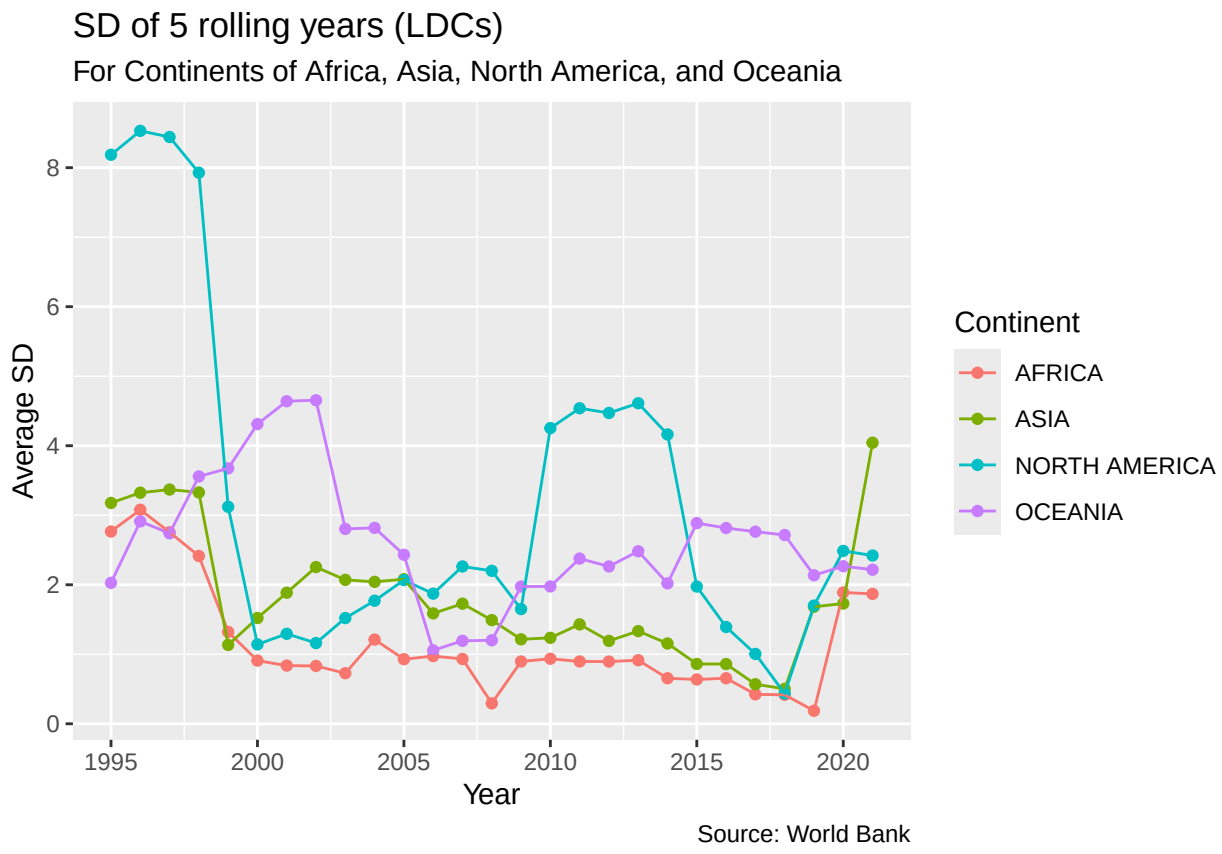
Now we have the data calculated, we can plot them with the following code:

```
library(tidyverse)
library(ggplot2)
library(dplyr)
library(tidyr)
library(ggpubr)
```

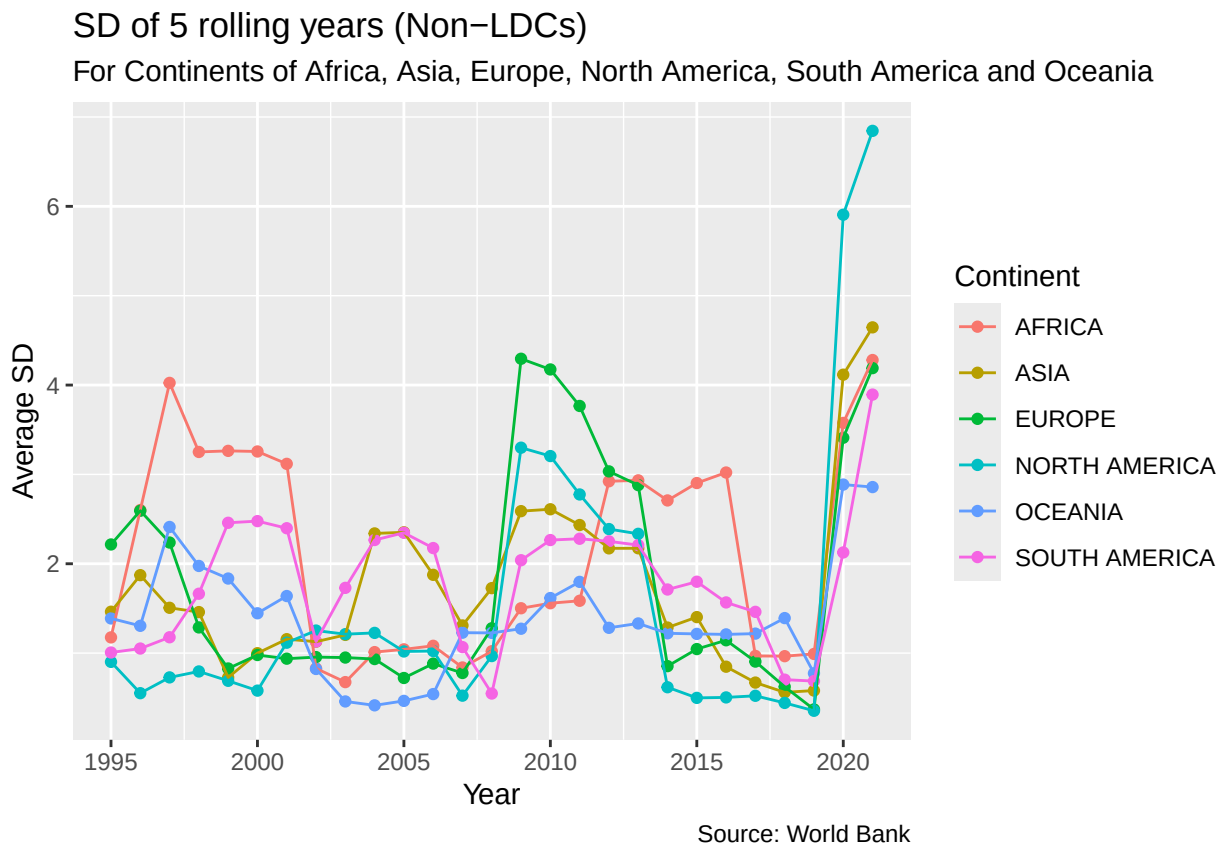
```
p_sd_5yr_all <- ggplot(data = na.omit(sd_5yr_all), aes(year, sd, col = continent)) +
  geom_point() +
  geom_line() +
  labs(title = "SD of 5 rolling years (World)", subtitle = "For Continents of Africa, Asia, Europe, Nor
  scale_color_discrete(labels = toupper, name = "Continent")
p_sd_5yr_all
```




```
p_sd_5yr_LDCs <- ggplot(data = na.omit(sd_5yr_ldc), aes(year, sd, col = continent)) +
  geom_point() +
  geom_line() +
  labs(title = "SD of 5 rolling years (LDCs)", subtitle = "For Continents of Africa, Asia, North America
  scale_color_discrete(labels = toupper, name = "Continent")
p_sd_5yr_LDCs
```




```
p_sd_5yr_nonLDCs <- ggplot(data = na.omit(sd_5yr_noldc), aes(year, sd, col = continent)) +
  geom_point() +
  geom_line() +
  labs(title = "SD of 5 rolling years (Non-LDCs)", subtitle = "For Continents of Africa, Asia, Europe, N
  scale_color_discrete(labels = toupper, name = "Continent")
p_sd_5yr_nonLDCs
```



It is also worth exploring this statistical data in different formats. Let us consider experimenting with the data using box plots.

```
library(tidyverse)
library(ggplot2)
library(dplyr)
library(tidyr)
library(ggpubr)


p_box_all_pre <- ggplot(data = filter(continent_growth, year <= 2014), aes(continent, cont_avg_gdp_grow
  geom_boxplot() +
  labs(title = "Box Plot for Continents Including All Countries", subtitle = "Pre-2015", x = "", y = "A

p_box_all_post <- ggplot(data = filter(continent_growth, year >= 2015), aes(continent, cont_avg_gdp_gro
  geom_boxplot() +
  labs(subtitle = "Post-2015", x = "", y = "", caption = "Source: World Bank")

p_box_all <- ggarrange(p_box_all_pre, p_box_all_post, ncol=1, nrow = 2)



p_box_ldc_pre <- ggplot(data = filter(continent_ldc_growth, year <= 2014), aes(continent, cont_avg_gdp_
  geom_boxplot() +
  labs(title = "Box Plot for Continents Including LDCs", subtitle = "Pre-2015", x = "", y = "Average GD
```

```
p_box_ldc_post <- ggplot(data = filter(continent_ldc_growth, year >= 2015), aes(continent, cont_avg_gdp
  geom_boxplot() +
  labs(subtitle = "Post-2015", x = "", y = "", caption = "Source: World Bank")

p_box_ldc <- ggarrange(p_box_ldc_pre, p_box_ldc_post, ncol=1, nrow = 2)



p_box_noldc_pre <- ggplot(data = filter(continent_noldc_growth, year <= 2014), aes(continent, cont_avg_
  geom_boxplot() +
  labs(title = "Box Plot for Continents Including Non-LDCs", subtitle = "Pre-2015", x = "", y = "Averag

p_box_noldc_post <- ggplot(data = filter(continent_noldc_growth, year >= 2015), aes(continent, cont_avg
  geom_boxplot() +
  labs(subtitle = "Post-2015", x = "", y = "", caption = "Source: World Bank")

p_box_noldc <- ggarrange(p_box_noldc_pre, p_box_noldc_post, ncol=1, nrow = 2)
```

That is all the code for question one. Let us now consider question two.

QUESTION TWO - Elisa & Fiona

Here is the code used in creating the graphs and data used in Question Two, courtesy of Elisa and Fiona.

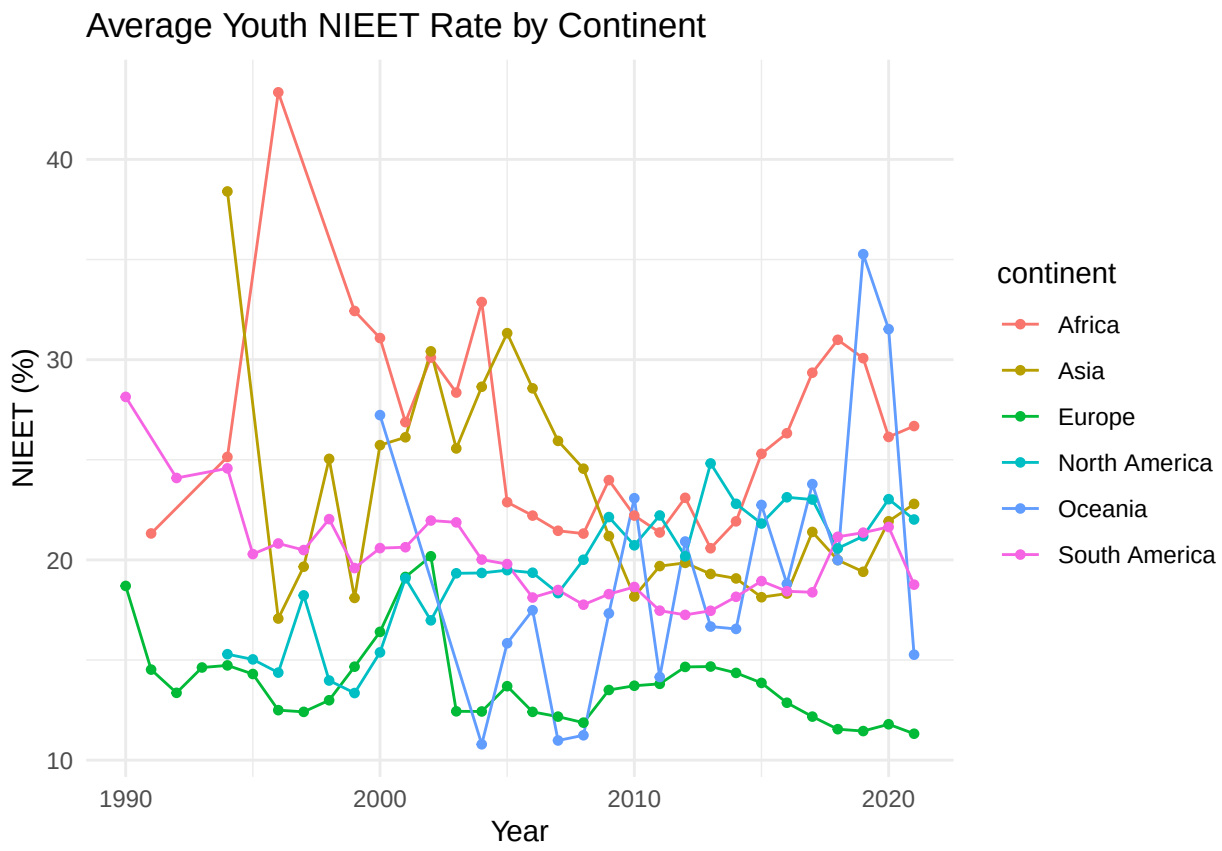
Question two is all about ascertaining continent's and country's targets in "substantially reducing the proportion of youth not in employment, education or training" (NEET).

First, the data preparation:

```
library(tidyverse)
library(ggplot2)

full <- read.csv("full_data.csv",stringsAsFactors = F)

nieet_by_continent <- full %>%
  filter(!is.na(youth_NIEET)) %>%
  group_by(continent, year) %>%
  summarise(nieet = mean(youth_NIEET, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'continent'. You can override using the
## `.groups` argument.
```

```
nieet_by_continent %>%
  ggplot(aes(x = year, y = nieet, colour = continent)) +
  geom_line()+
  geom_point(size = 1.2)+
  labs(
    title = "Average Youth NIEET Rate by Continent",
    x = "Year",
    y = "NIEET (%)"
  ) +
  theme_minimal()
```




```
data_2000 <- nieet_by_continent %>%
  filter(year == 2000) %>%
  select(continent, nieet) %>%
  rename(nieet_2000 = nieet)

data_2020 <- nieet_by_continent %>%
  filter(year == 2020) %>%
  select(continent, nieet) %>%
  rename(nieet_2020 = nieet)

nieet_change <- data_2000 %>%
  inner_join(data_2020, by = "continent") %>%
  mutate(change = nieet_2020 - nieet_2000)
```

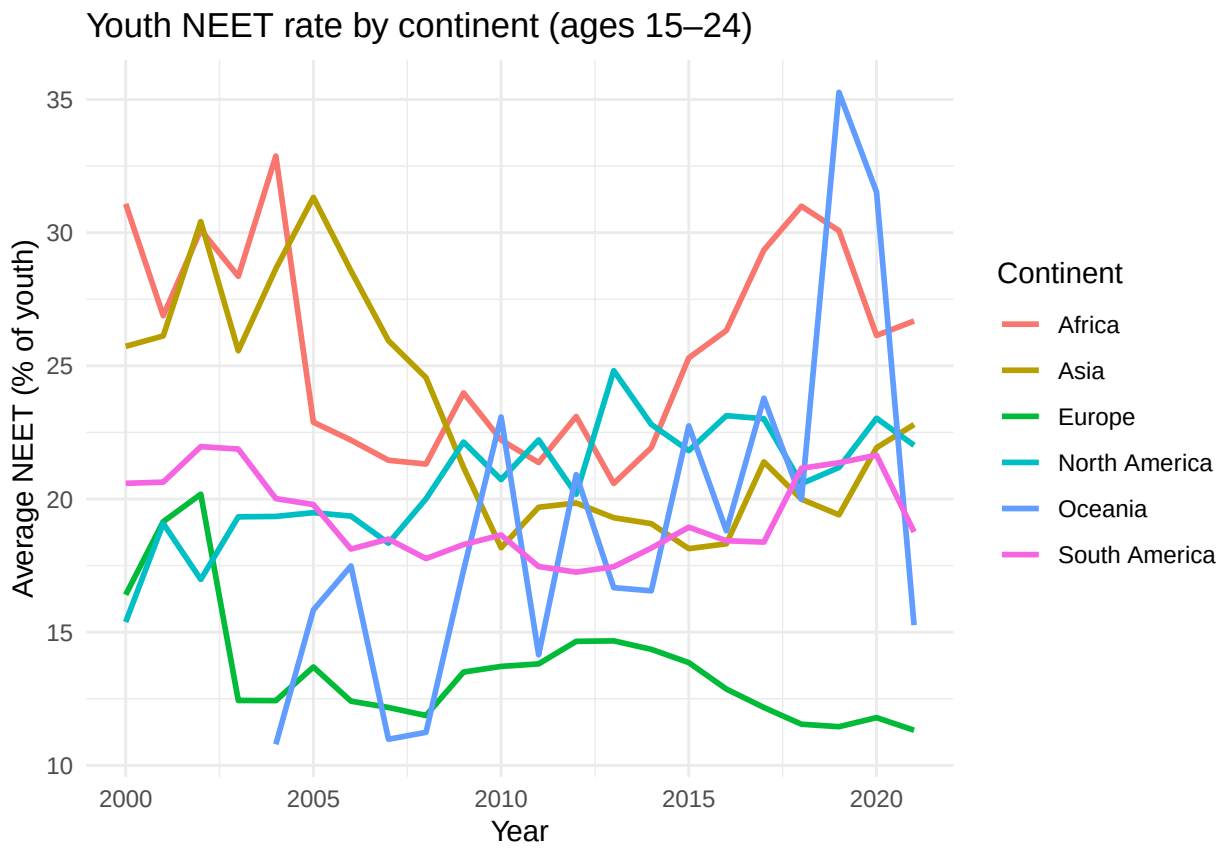
Now, the plots:

```
library(tidyverse)
full_data <- read_csv("full_data.csv")
```

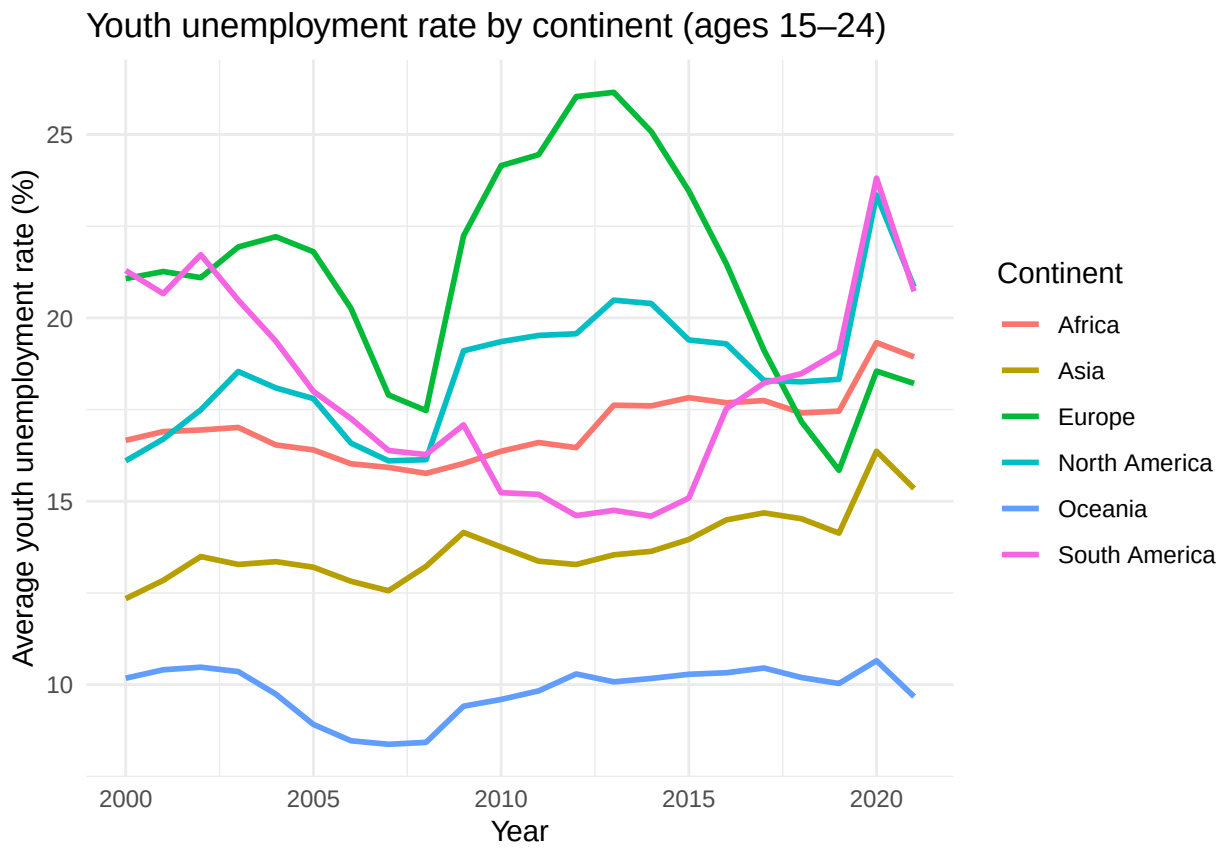
```
## Rows: 6346 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): country, code, continent
## dbl (5): year, gdp_pc_2017, youth_NIEET, gdp_growth, youth_unemployment_rate
##
```

```
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
full_data_q2 <- full_data %>%
  filter(year >= 2000, year <= 2023) %>%
  filter(!is.na(continent),
         continent != "Antarctica") %>%
  rename(
    neet_youth  = youth_NIEET,
    youth_unemp = youth_unemployment_rate
  )
continent_year <- full_data_q2 %>%
  group_by(continent, year) %>%
  summarise(
    neet_mean  = mean(neet_youth,  na.rm = TRUE),
    unemp_mean = mean(youth_unemp, na.rm = TRUE),
    n_countries = n(),
    .groups    = "drop"
  )
ggplot(continent_year,
       aes(x = year, y = neet_mean, colour = continent)) +
  geom_line(linewidth = 1) +
  labs(
    title  = "Youth NEET rate by continent (ages 15-24)",
    x      = "Year",
    y      = "Average NEET (% of youth)",
    colour = "Continent"
  ) +
  theme_minimal()
```




```r
ggplot(continent_year,
       aes(x = year, y = unemp_mean, colour = continent)) +
  geom_line(linewidth = 1) +
  labs(
    title  = "Youth unemployment rate by continent (ages 15-24)",
    x      = "Year",
    y      = "Average youth unemployment rate (%)",
    colour = "Continent"
  ) +
  theme_minimal()
```




```
neet_change_2000_2020 <- continent_year %>%
  filter(year %in% c(2000, 2020)) %>%
  select(continent, year, neet_mean) %>%
  pivot_wider(
    names_from  = year,
    values_from = neet_mean,
    names_prefix = "neet_"
  ) %>%
  mutate(
    abs_change = neet_2020 - neet_2000,
    rel_change = 100 * (neet_2020 - neet_2000) / neet_2000,
    substantial_reduction = case_when(
      rel_change <= -20 ~ "Yes (>= 20% drop)",
      TRUE              ~ "No"
    )
  )

neet_change_2000_2020
```

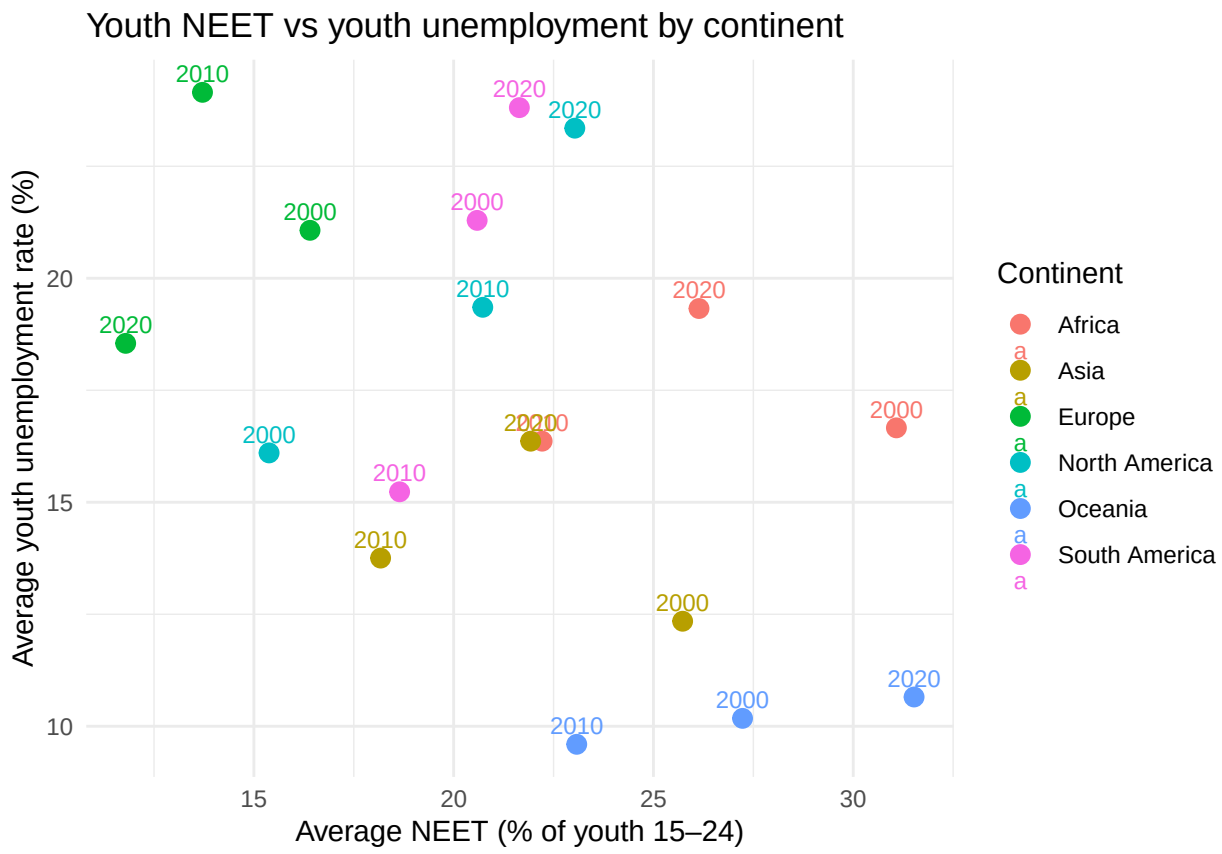
```
## # A tibble: 6 x 6
##   continent     neet_2000 neet_2020 abs_change rel_change substantial_reduction
##   <chr>             <dbl>     <dbl>      <dbl>      <dbl> <chr>
## 1 Africa             31.1      26.1      -4.94      -15.9  No
## 2 Asia               25.7      21.9      -3.80      -14.8  No
## 3 Europe             16.4      11.8      -4.61      -28.1  Yes (>= 20% drop)
```

```
## 4 North America      15.4      23.0       7.65      49.8  No
## 5 Oceania            27.2      31.5       4.29      15.8  No
## 6 South America      20.6      21.6       1.06       5.13 No
```

```
unemp_change_2000_2020 <- continent_year %>%
  filter(year %in% c(2000, 2020)) %>%
  select(continent, year, unemp_mean) %>%
  pivot_wider(
    names_from  = year,
    values_from = unemp_mean,
    names_prefix = "unemp_"
  ) %>%
  mutate(
    abs_change = unemp_2020 - unemp_2000,
    rel_change = 100 * (unemp_2020 - unemp_2000) / unemp_2000
  )

unemp_change_2000_2020
```

```
## # A tibble: 6 x 5
##   continent     unemp_2000 unemp_2020 abs_change rel_change
##   <chr>              <dbl>      <dbl>      <dbl>      <dbl>
## 1 Africa              16.7       19.3      2.66       16.0
## 2 Asia                12.3       16.4      4.02       32.6
## 3 Europe              21.1       18.5     -2.52      -12.0
## 4 North America       16.1       23.4      7.25       45.0
## 5 Oceania             10.2       10.7      0.477       4.69
## 6 South America       21.3       23.8      2.52       11.8
```

```
key_years <- c(2000, 2010, 2020)

continent_key_years <- continent_year %>%
  filter(year %in% key_years)

ggplot(continent_key_years,
       aes(x = neet_mean,
           y = unemp_mean,
           colour = continent)) +
  geom_point(size = 3) +
  geom_text(aes(label = year), vjust = -0.6, size = 3) +
  labs(
    title  = "Youth NEET vs youth unemployment by continent",
    x      = "Average NEET (% of youth 15-24)",
    y      = "Average youth unemployment rate (%)",
    colour = "Continent"
  ) +
  theme_minimal()
```



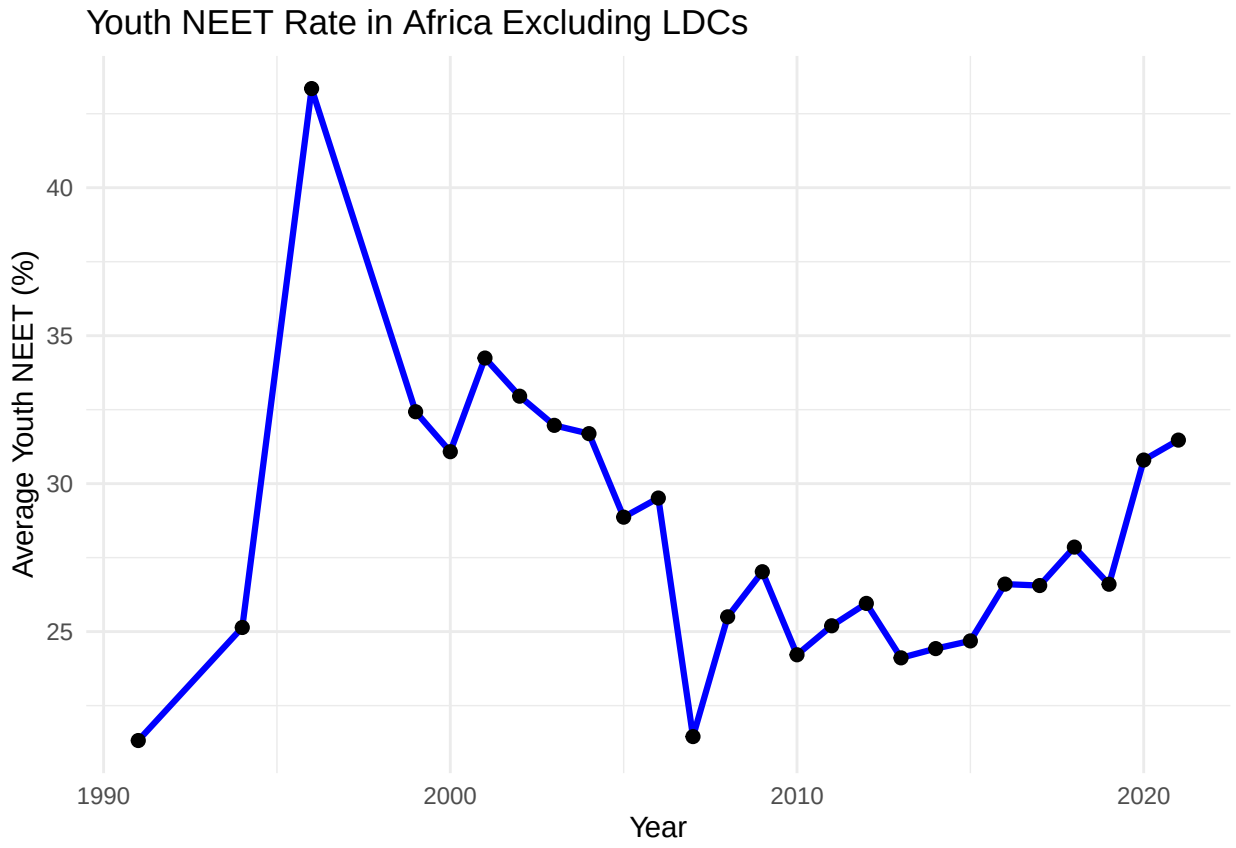
Here we also have code for NEET per continent, excluding LDCs.

```
library(tidyverse)
data <- read_csv("noldc_full_data.csv")
```

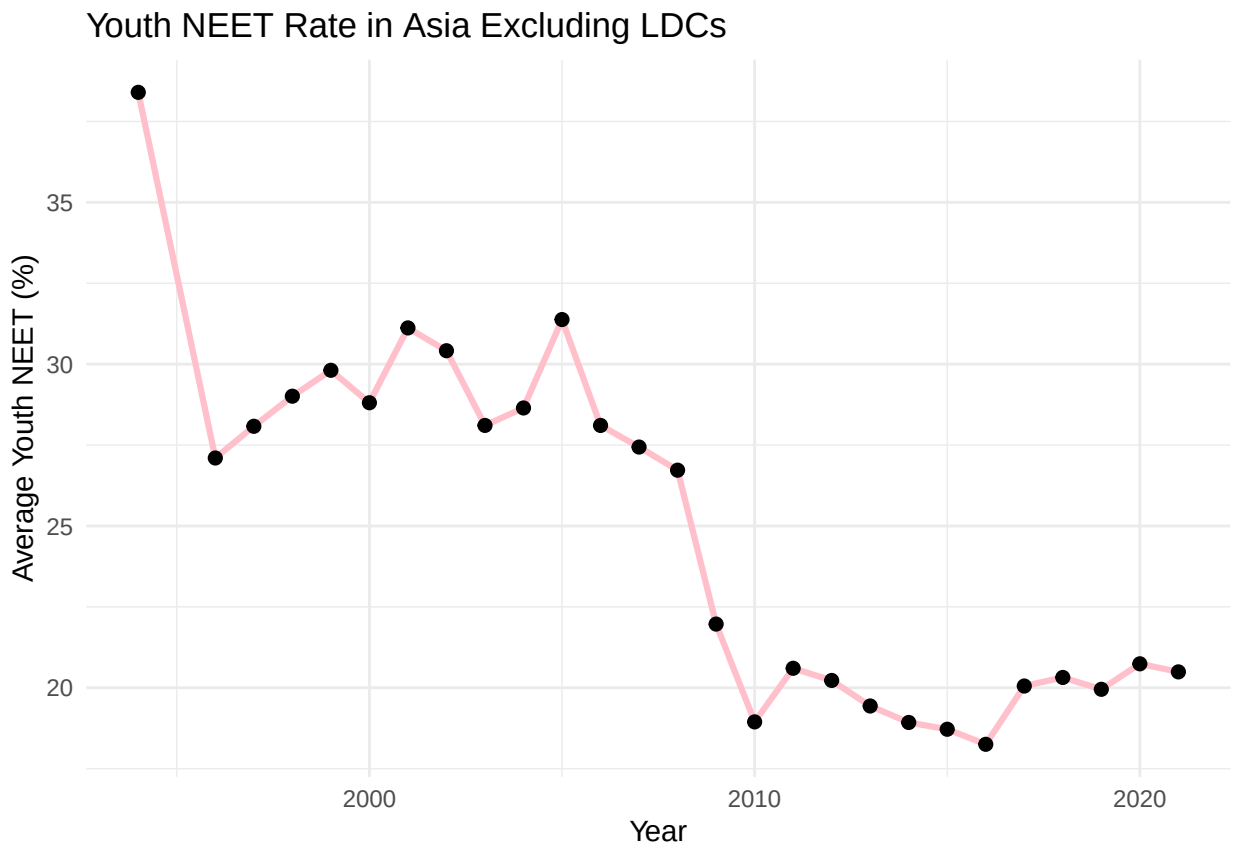
```
## Rows: 5115 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): country, code, continent
## dbl (5): year, gdp_pc_2017, youth_NIEET, gdp_growth, youth_unemployment_rate
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
plot_neet_by_continent <- function(df, continent_name, line_colour, plot_title) {
  df %>%
    filter(continent == continent_name,
           !is.na(youth_NIEET)) %>%
    group_by(year) %>%
    summarise(neet_mean = mean(youth_NIEET, na.rm = TRUE), .groups = "drop") %>%
    ggplot(aes(x = year, y = neet_mean)) +
    geom_line(color = line_colour, linewidth = 1.1) +
    geom_point(size = 2) +
    labs(
      title = plot_title,
```

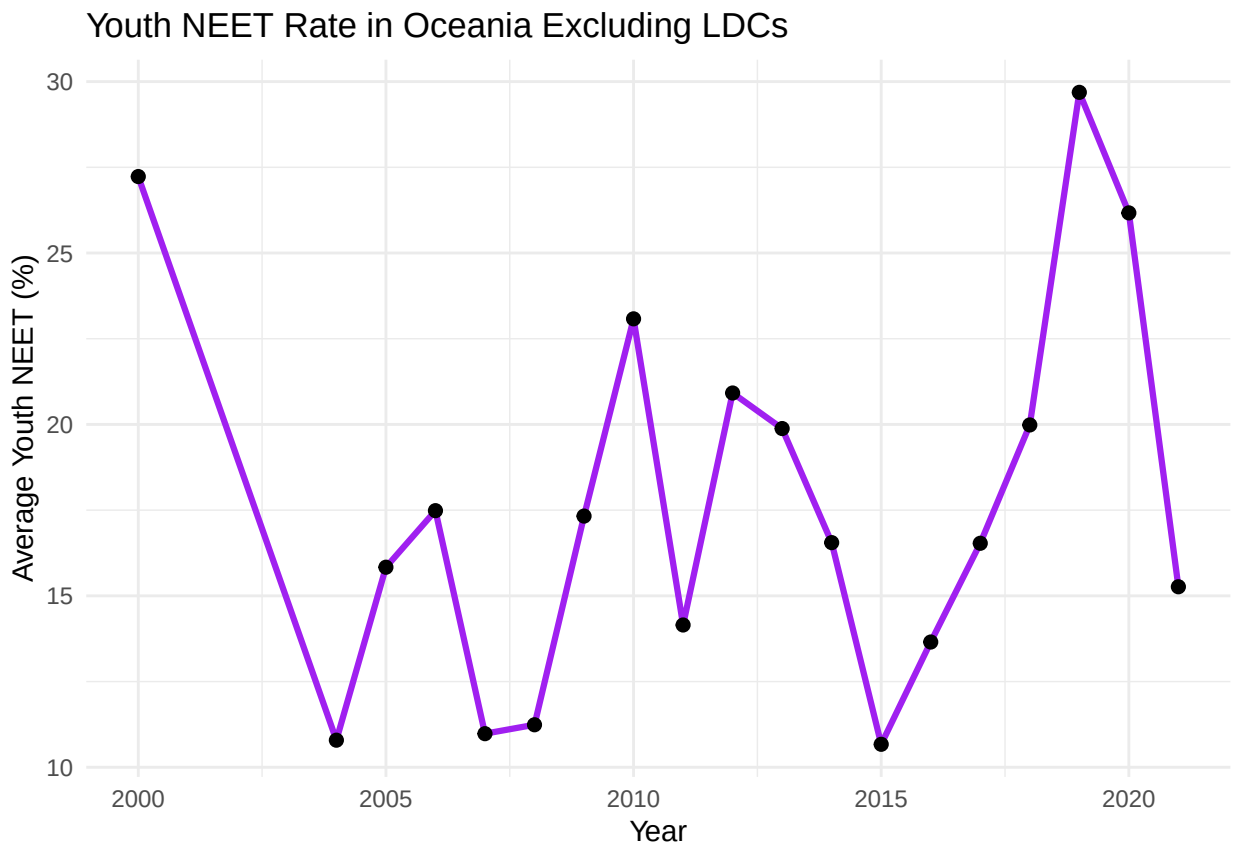
```
      x = "Year",
      y = "Average Youth NEET (%)"
    ) +
    theme_minimal()
}

plot_neet_by_continent(
  df = data,
  continent_name = "Africa",
  line_colour = "blue",
  plot_title = "Youth NEET Rate in Africa Excluding LDCs"
)
```



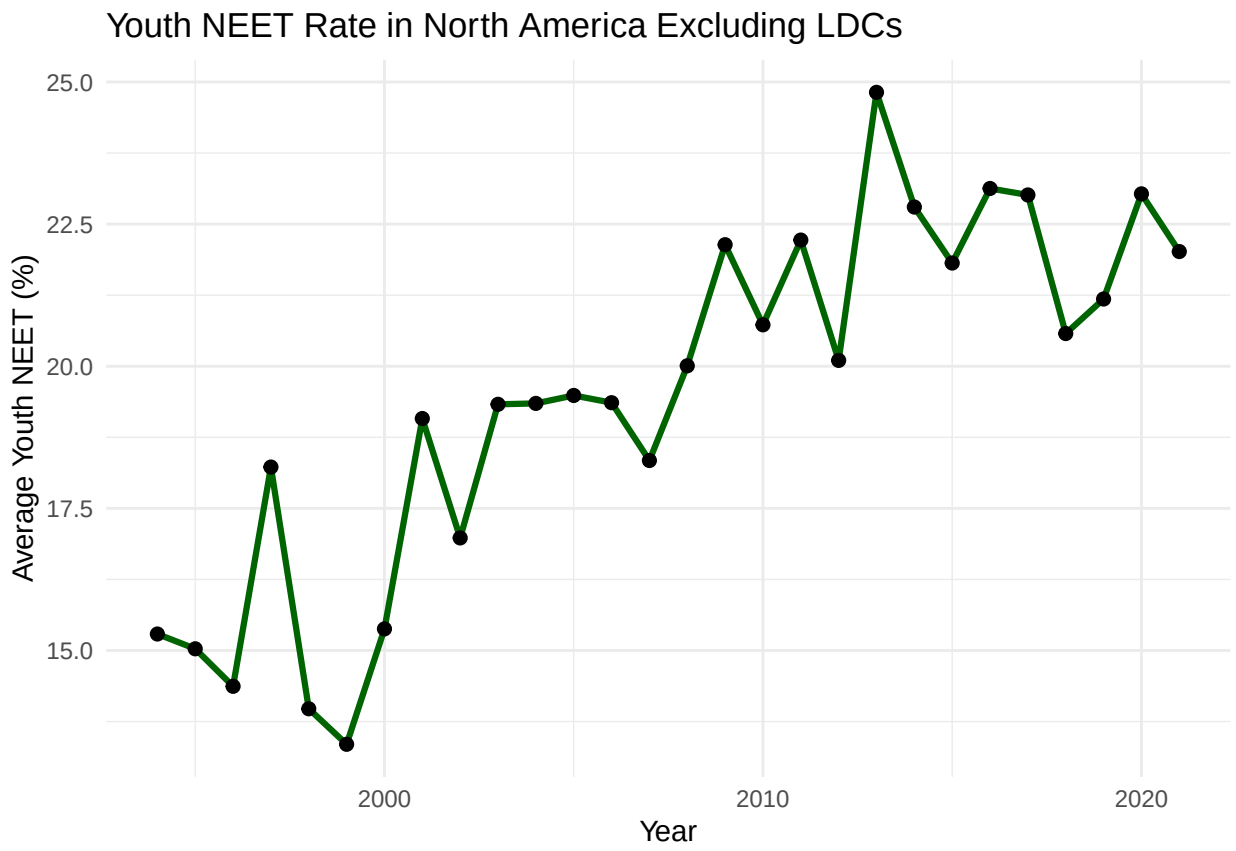

```
plot_neet_by_continent(
  df = data,
  continent_name = "Asia",
  line_colour = "pink",
  plot_title = "Youth NEET Rate in Asia Excluding LDCs"
)
```



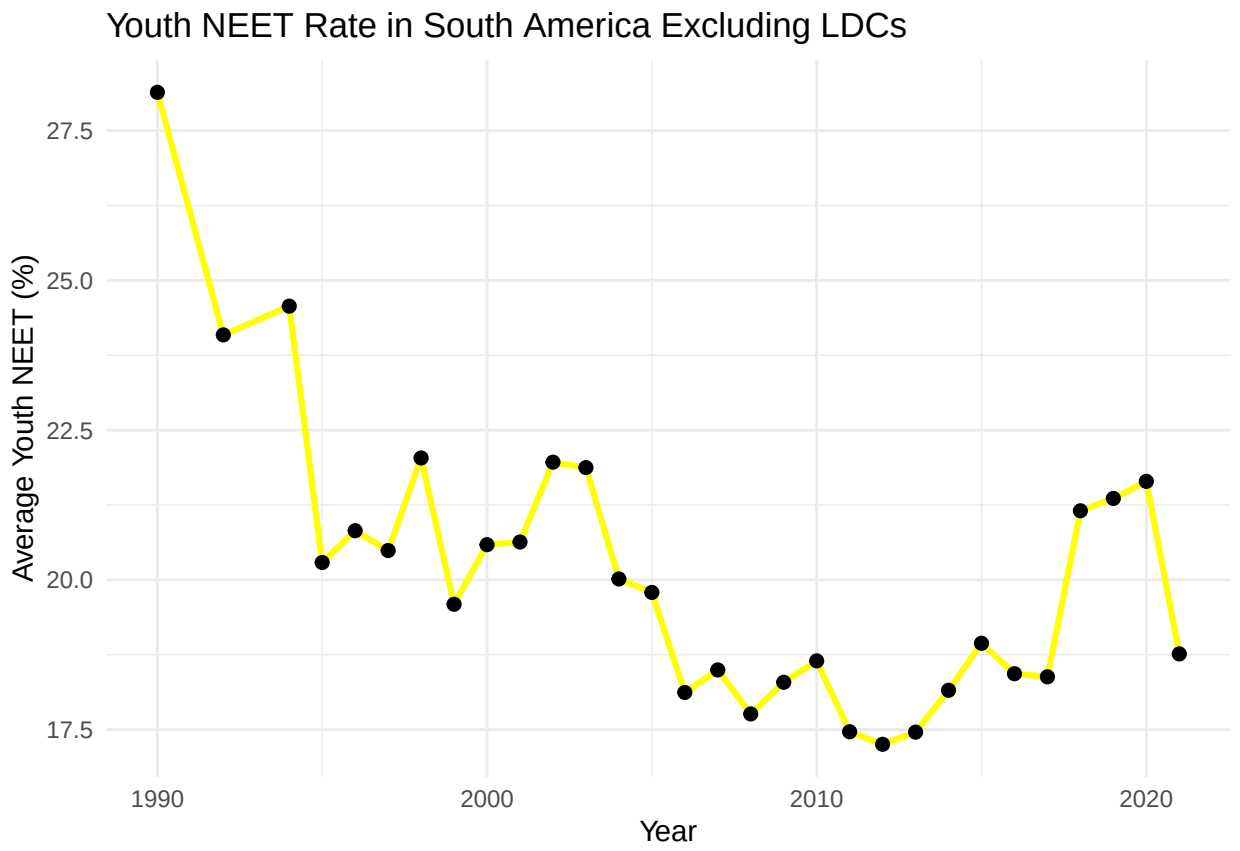

```
plot_neet_by_continent(
  df = data,
  continent_name = "Oceania",
  line_colour = "purple",
  plot_title = "Youth NEET Rate in Oceania Excluding LDCs"
)
```



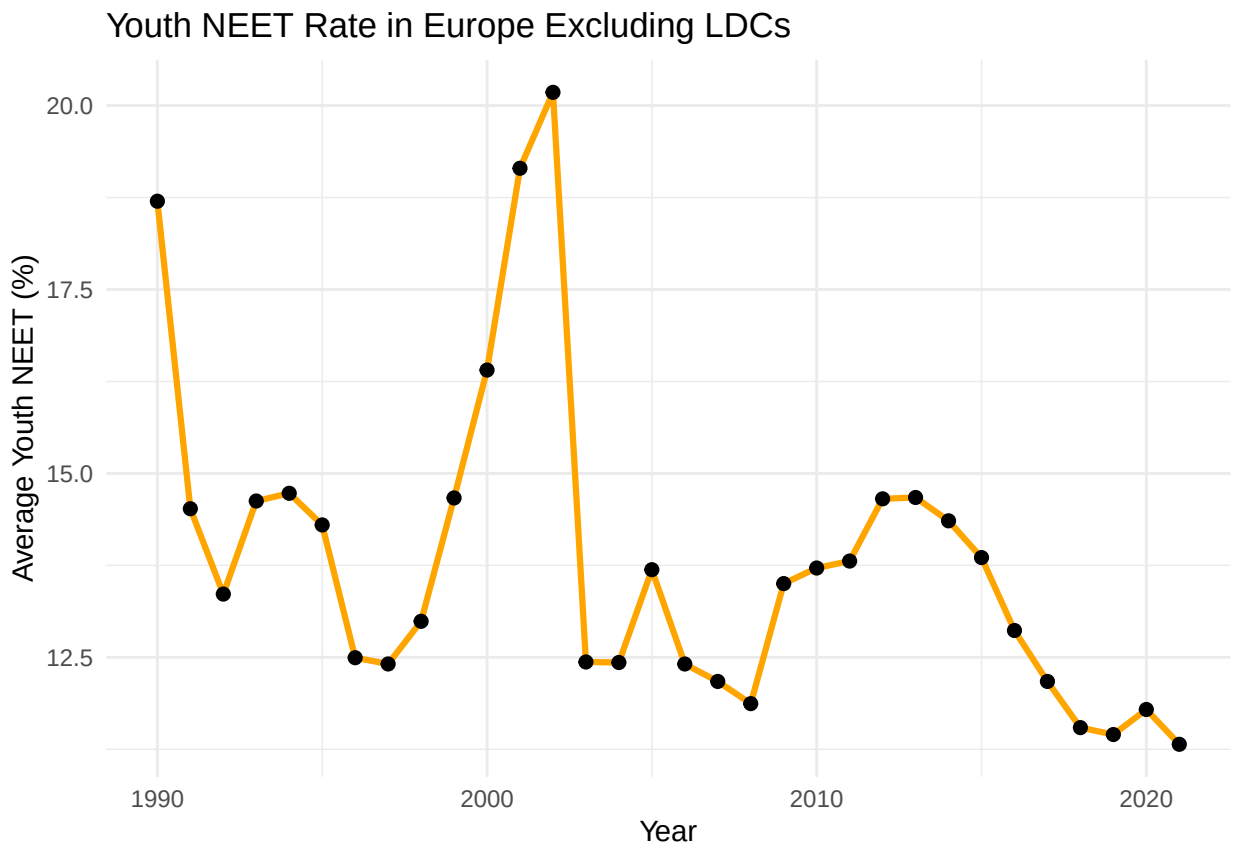

```
plot_neet_by_continent(
  df = data,
  continent_name = "North America",
  line_colour = "darkgreen",
  plot_title = "Youth NEET Rate in North America Excluding LDCs"
)
```




```
plot_neet_by_continent(
  df = data,
  continent_name = "South America",
  line_colour = "yellow",
  plot_title = "Youth NEET Rate in South America Excluding LDCs"
)
```

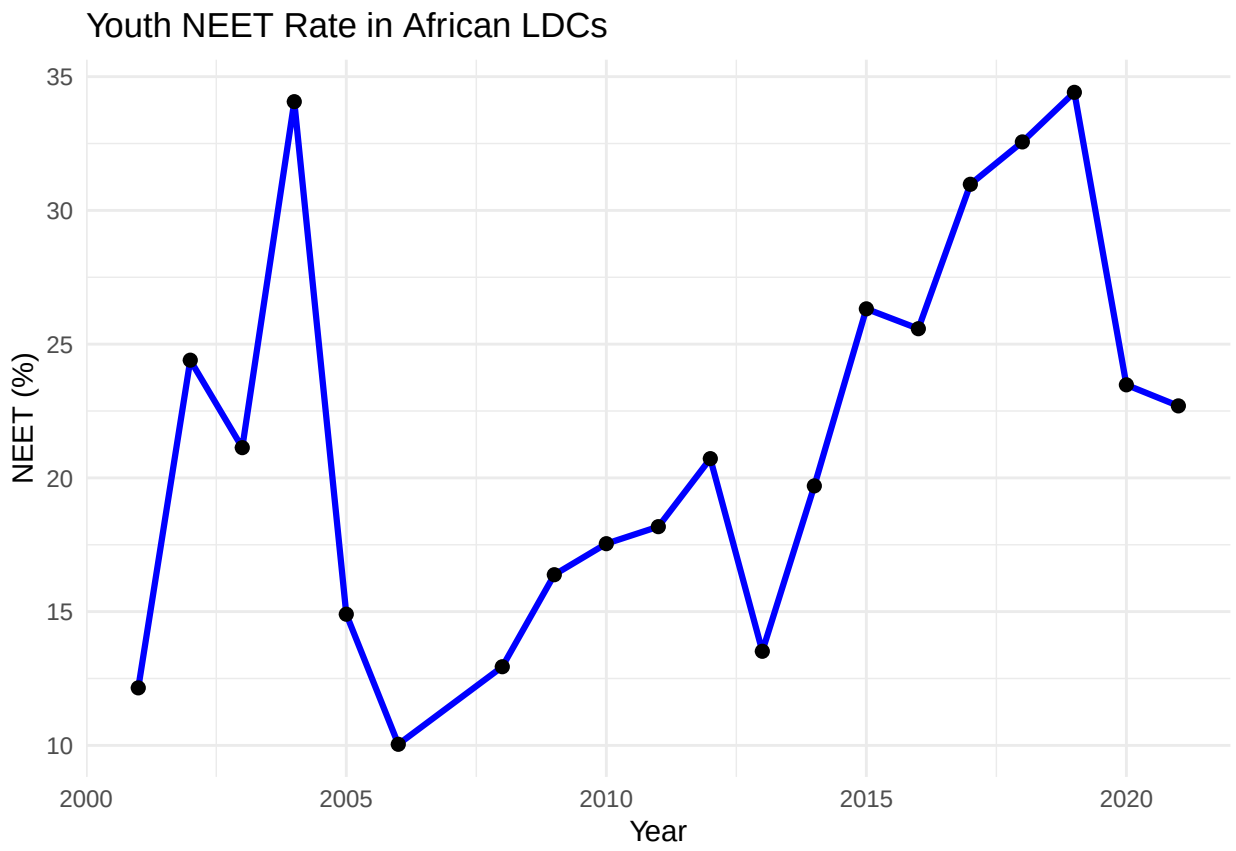



```
plot_neet_by_continent(
  df = data,
  continent_name = "Europe",
  line_colour = "orange",
  plot_title = "Youth NEET Rate in Europe Excluding LDCs"
)
```

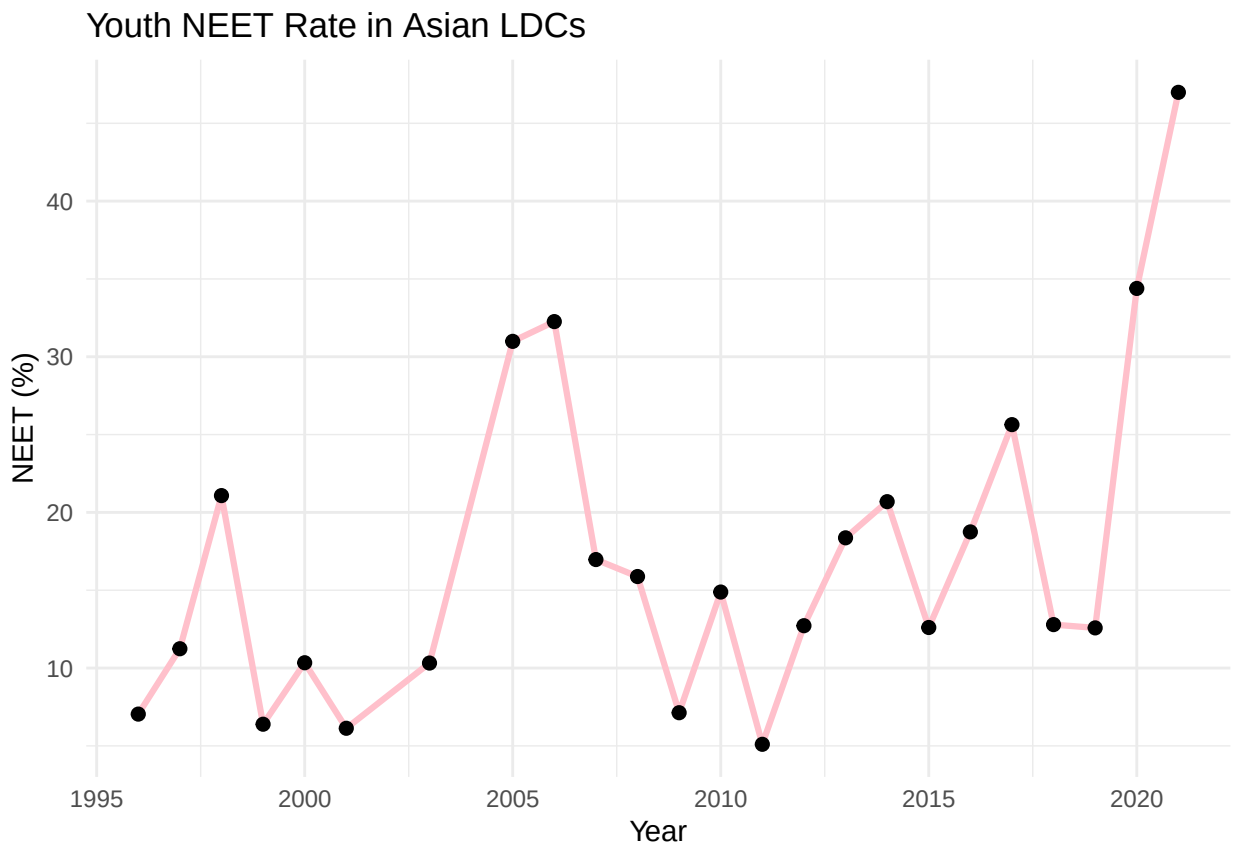


Finally, a focus on continents and their LDCs exclusively.

```
library(tidyverse)
library(ggplot2)

ldc <- read.csv("ldc_full_data.csv",
                stringsAsFactors = FALSE)

ldc_africa <- ldc %>%
  filter(continent == "Africa" & !is.na(youth_NIEET)) %>%
  group_by(year) %>%
  summarise(avg_neet = mean(youth_NIEET, na.rm = TRUE))

ggplot(ldc_africa, aes(x = year, y = avg_neet)) +
  geom_line(color = "blue", size = 1.1) +
  geom_point(size = 2) +
  theme_minimal() +
  labs(title = "Youth NEET Rate in African LDCs",
       x = "Year",
       y = "NEET (%)")
```

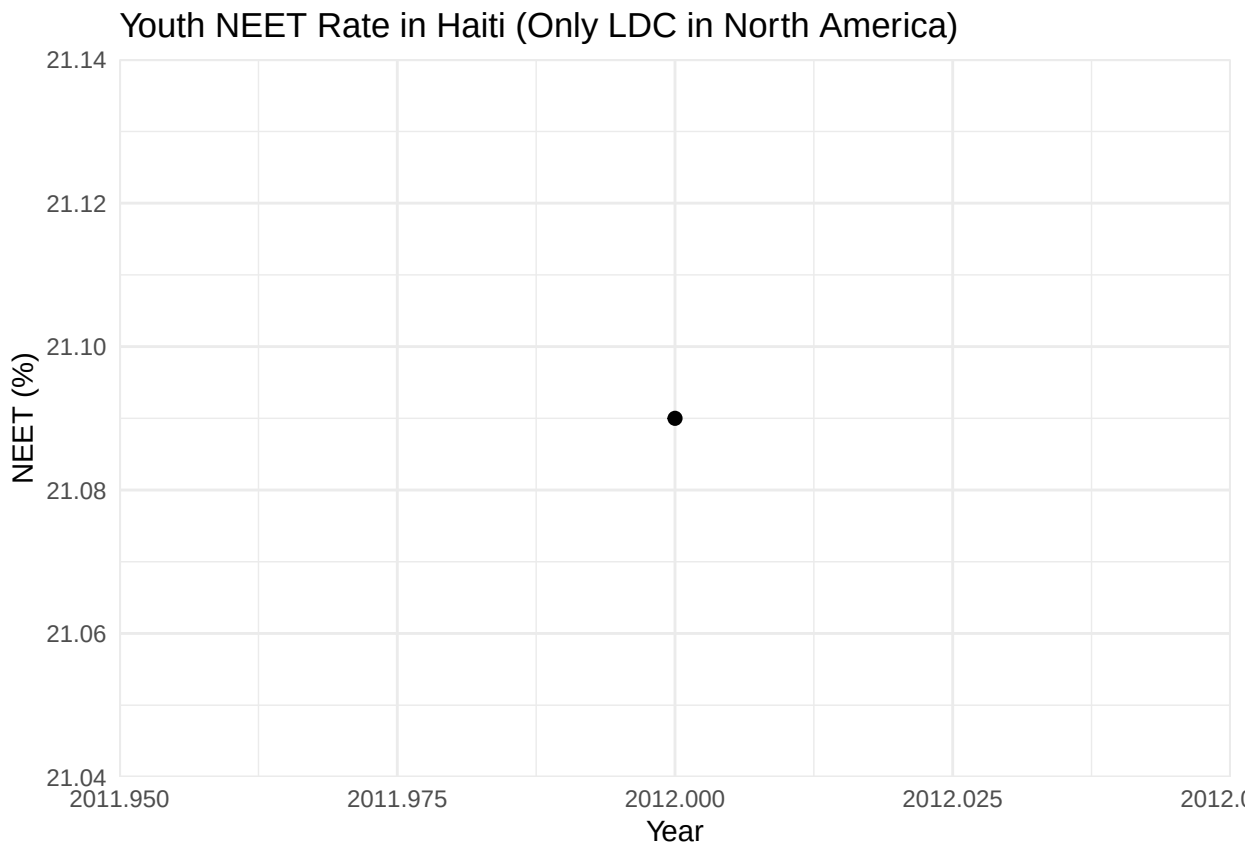



```
library(tidyverse)
library(ggplot2)

ldc <- read.csv("ldc_full_data.csv",
                stringsAsFactors = FALSE)

ldc_asia <- ldc %>%
  filter(continent == "Asia" & !is.na(youth_NIEET)) %>%
  group_by(year) %>%
  summarise(avg_neet = mean(youth_NIEET, na.rm = TRUE))

ggplot(ldc_asia, aes(x = year, y = avg_neet)) +
  geom_line(color = "pink", size = 1.1) +
  geom_point(size = 2) +
  theme_minimal() +
  labs(title = "Youth NEET Rate in Asian LDCs",
       x = "Year",
       y = "NEET (%)")
```

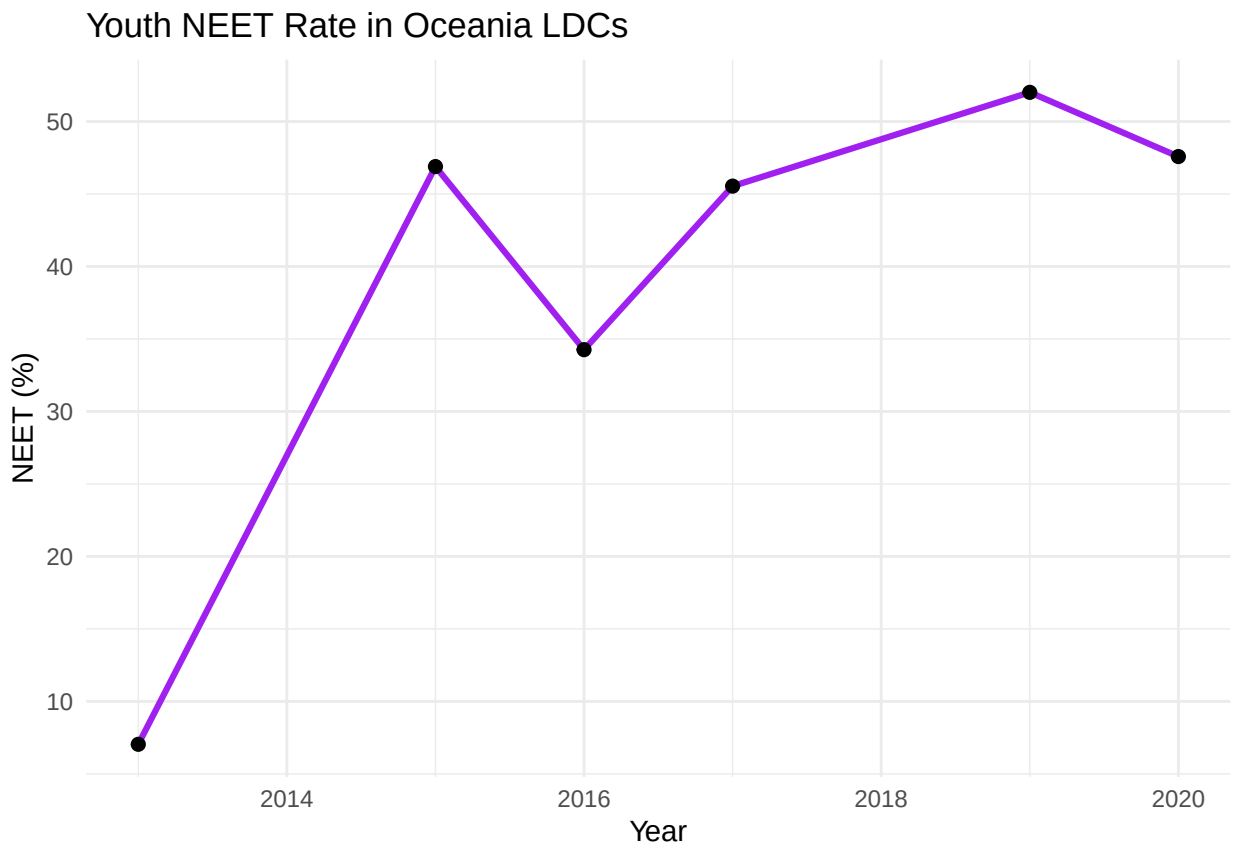



```
library(tidyverse)
library(ggplot2)

ldc <- read.csv("ldc_full_data.csv",
                stringsAsFactors = FALSE)

ldc_north <- ldc %>%
  filter(continent == "North America" &
           !is.na(youth_NIEET))

ggplot(ldc_north, aes(x = year, y = youth_NIEET)) +
  geom_line(color = "darkgreen", size = 1.1) +
  geom_point(size = 2) +
  theme_minimal() +
  labs(title = "Youth NEET Rate in Haiti (Only LDC in North America)",
       x = "Year",
       y = "NEET (%)")
```

```
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
```




```
library(tidyverse)
library(ggplot2)

ldc <- read.csv("ldc_full_data.csv",
                stringsAsFactors = FALSE)

ldc_oceania <- ldc %>%
  filter(continent == "Oceania" & !is.na(youth_NIEET)) %>%
  group_by(year) %>%
  summarise(avg_neet = mean(youth_NIEET, na.rm = TRUE))

ggplot(ldc_oceania, aes(x = year, y = avg_neet)) +
  geom_line(color = "purple", size = 1.1) +
  geom_point(size = 2) +
  theme_minimal() +
  labs(title = "Youth NEET Rate in Oceania LDCs",
       x = "Year",
       y = "NEET (%)")
```



That is all the code for question two. For more information, see the associated presentation and/or report. Thank you,

C10.