Data Science Project

Group K14

Cleaned and annotated R code used to generate the graphs for the Data Science Group Project.

Loading the packages

```
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(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.1     v readr     2.1.5
## v ggplot2   4.0.0     v stringr   1.5.2
## v lubridate 1.9.4     v tibble    3.3.0
## v purrr     1.1.0     v tidyr     1.3.1

## -- 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 errors
```

```
library(tidyr)
library(ggplot2)
```

Loading the datasets

```
continents <- read.csv("continents-according-to-our-world-in-data.csv")
gdp <- read.csv("gdp-per-capita-worldbank.csv")
NEET <- read.csv("youth-not-in-education-employment-training.csv")
```

Cleaning the data

Firstly, I shortened the column names for clarity

```
gdp <- gdp %>%
  rename(GDP = GDP.per.capita..PPP..constant.2017.international...)
NEET <- NEET %>%
  rename(share_NEET = Share.of.youth.not.in.education..employment.or.training..total....of.youth.popula
colnames(NEET)
```

```
## [1] "Entity"     "Code"       "Year"       "share_NEET"
```

I also found that in the continents data, all Years were recorded as 2015. This is not relevant so I used the select function to remove the Year column. I also removed any duplicate rows in the dataset.

```
continents_cleaned <- continents %>%
  select(Entity, Continent) %>%
  distinct()
```

The data from our chosen csv (Unemployment rate data) had to be cleaned and changed into 'long' format, as it had initially set each year as a column title. Since I needed to combine the unemployment rate data with continents I also needed to make sure that the column names matched. This involved removing the 'X' put in front of each year (e.g X1991) and changing 'Country.Name' to 'Entity' and 'Country.Code' to 'Code'.

```
Unemployment_Rate <- read.csv("unemployment_rate.csv")
names(Unemployment_Rate) <- gsub("^X", "", names(Unemployment_Rate))
Unemployment_Rate_Long <- Unemployment_Rate %>%
  gather(Year, Unemployment_Rate, "1991":"2024") %>%
  rename(Entity = Country.Name, Code = Country.Code) %>%
  mutate(Year = as.integer(Year), Unemployment_Rate = as.numeric(Unemployment_Rate))
```

Joining the datasets

Once the data was properly cleaned, I used the left_join function to combine it with the continents data, making sure to remove Antarctica as required. ## Dataset for target 1 and removing any Antarctica data:

```
continents_gdp <- gdp %>%
  left_join(continents_cleaned, by = "Entity") %>%
  filter(Continent != "Antarctica")
```

Dataset for target 2:

```
ontinents_NEET <- NEET %>%
  left_join(continents_cleaned, by = "Entity") %>%
  filter(!is.na(share_NEET), Continent != "Antarctica")
```

Dataset for additional csv

```
UnemploymentRate_Continents <- Unemployment_Rate_Long %>%
  left_join(continents_cleaned, by = "Entity") %>%
  filter(Continent != "Antarctica")
```

Approaching data analysis for target 1

Target 1 assesses per capita economic growth rates of the continents, whereas the datasets at the moment are just GDP per capita levels. Therefore, found the growth rate using percentage change from the previous year, with the formula current GDP - previous year GDP, divided by previous year GDP, multiplied by 100.

```
gdp_growth_rate <- continents_gdp %>%
  arrange(Entity, Year) %>%
  group_by(Entity) %>%
  mutate(growth_rate = (GDP - lag(GDP)) / lag(GDP) * 100)
```

This gave me a new dataset to use. I then wanted to find the average growth rate per continent per year. This is because otherwise, the growth rate for each individual country within each continent would be plotted, resulting in a messy, unclear graph. I also filtered out any missing values.

```
avg_continent_growth <- gdp_growth_rate %>%
  group_by(Continent, Year) %>%
  summarise(avg_growth = mean(growth_rate, na.rm = TRUE), .groups = "drop")
```

This is also done for the unemployment rate data.

```
Avg_unemployment_continent <- UnemploymentRate_Continents %>%
  group_by(Continent) %>%
  summarise(avg_unemployment = mean(Unemployment_Rate, na.rm = TRUE))
```

Graphs

Line graph of GDP per capita growth per continent

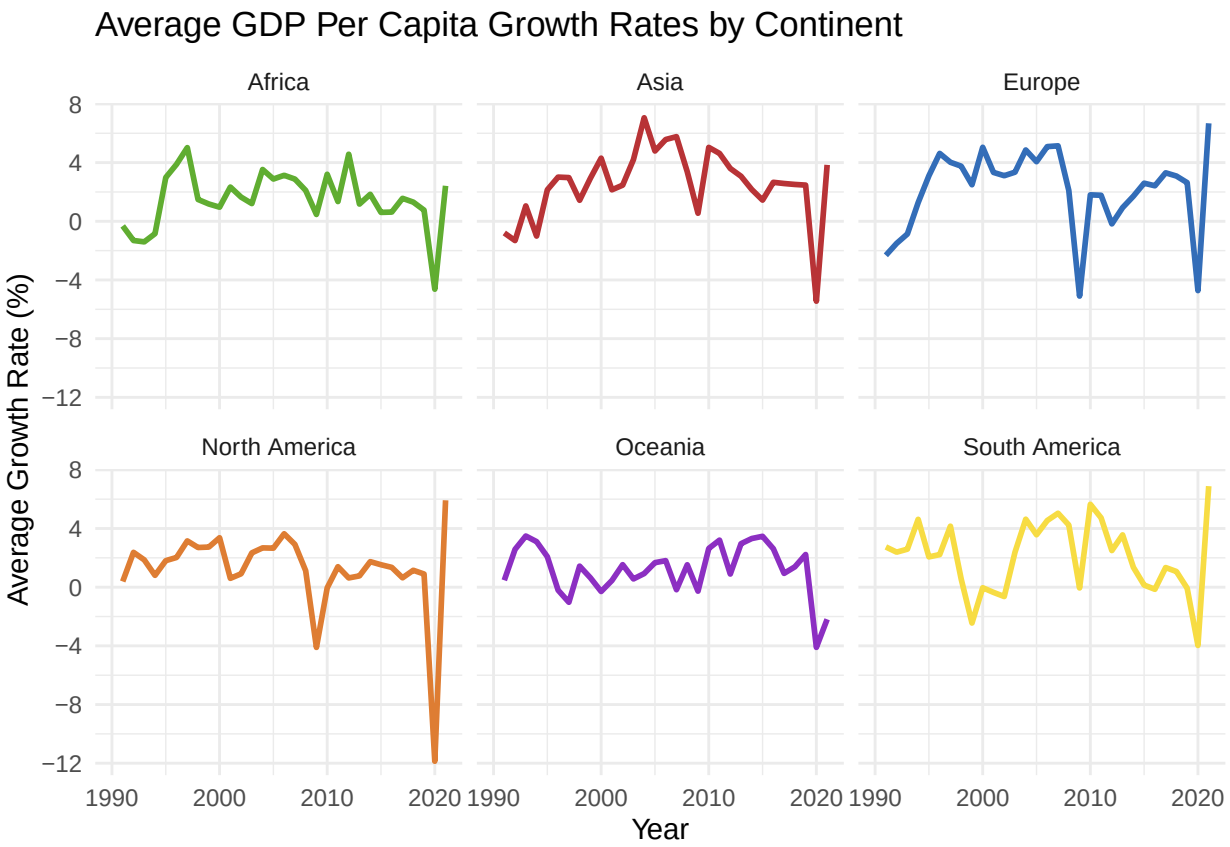
```
g2 <- avg_continent_growth %>%
  ggplot(aes(x = Year, y = avg_growth, colour = Continent, group = Continent)) +
  geom_line(size = 1) +
  facet_wrap(~Continent) +
  scale_colour_manual(values = c(
    "Africa" = "#60ad31",
```

```
    "Asia" = "#b83336",
    "Europe" = "#336db8",
    "North America" = "#de7d33",
    "South America" = "#f7dc43",
    "Oceania" = "#8c2fc2"
  )) +
  xlab("Year") +
  ylab("Average Growth Rate (%)") +
  ggtitle("Average GDP Per Capita Growth Rates by Continent") +
  theme_minimal() +
  theme(legend.position = "none")
```

```
## 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.
```

```
plot(g2)
```

```
## Warning: Removed 6 rows containing missing values or values outside the scale range
## (‘geom_line()‘).
```



Average unemployment rate of each continent over time bar chart

```
g6 <- ggplot(Avg_unemployment_continent, aes(x = Continent, y = avg_unemployment, fill = Continent)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(
    title = "Average Unemployment Rate by Continent",
    x = "Continent",
    y = "Average Unemployment Rate (%)") +
  scale_fill_manual(values = c(
    "Africa" = "#60ad31",
    "Asia" = "#b83336",
    "Europe" = "#336db8",
    "North America" = "#de7d33",
    "South America" = "#f7dc43",
    "Oceania" = "#8c2fc2"
  )) +
  theme(legend.position = "none")
plot(g6)
```