```r
setwd("~/Desktop/IDS Project")

library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.5.2
```

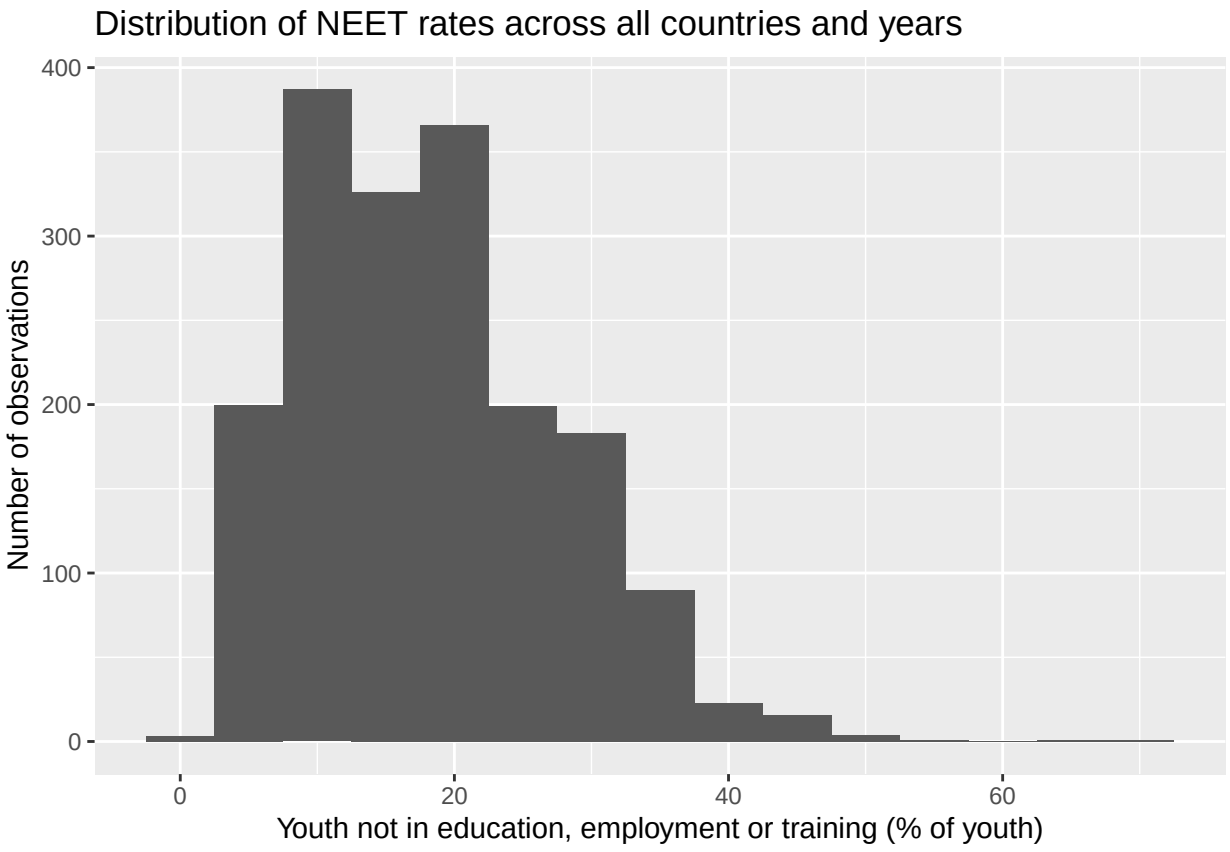
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.1     v stringr   1.5.2
## v ggplot2   4.0.1     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ 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 errors
```

```r
youth_raw <- read.csv("youth-not-in-education-employment-training.csv",
                      stringsAsFactors = FALSE)

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

youth <- youth_raw %>%
  rename(
    country = Entity,
    code    = Code,
    year    = Year
  )

names(youth)[4] <- "neet_pct"

youth <- youth %>%
  mutate(
    year     = as.integer(year),
    neet_pct = as.numeric(neet_pct)
  )

continents <- continents_raw %>%
  rename(
    country   = Entity,
    code      = Code,
    year      = Year,
    continent = Continent
  ) %>%
  filter(year == 2015) %>%
  select(country, continent) %>%
  distinct()

youth_cont <- youth %>%
  left_join(continents, by = "country") %>%
  filter(!is.na(continent),
         continent != "Antarctica",
```

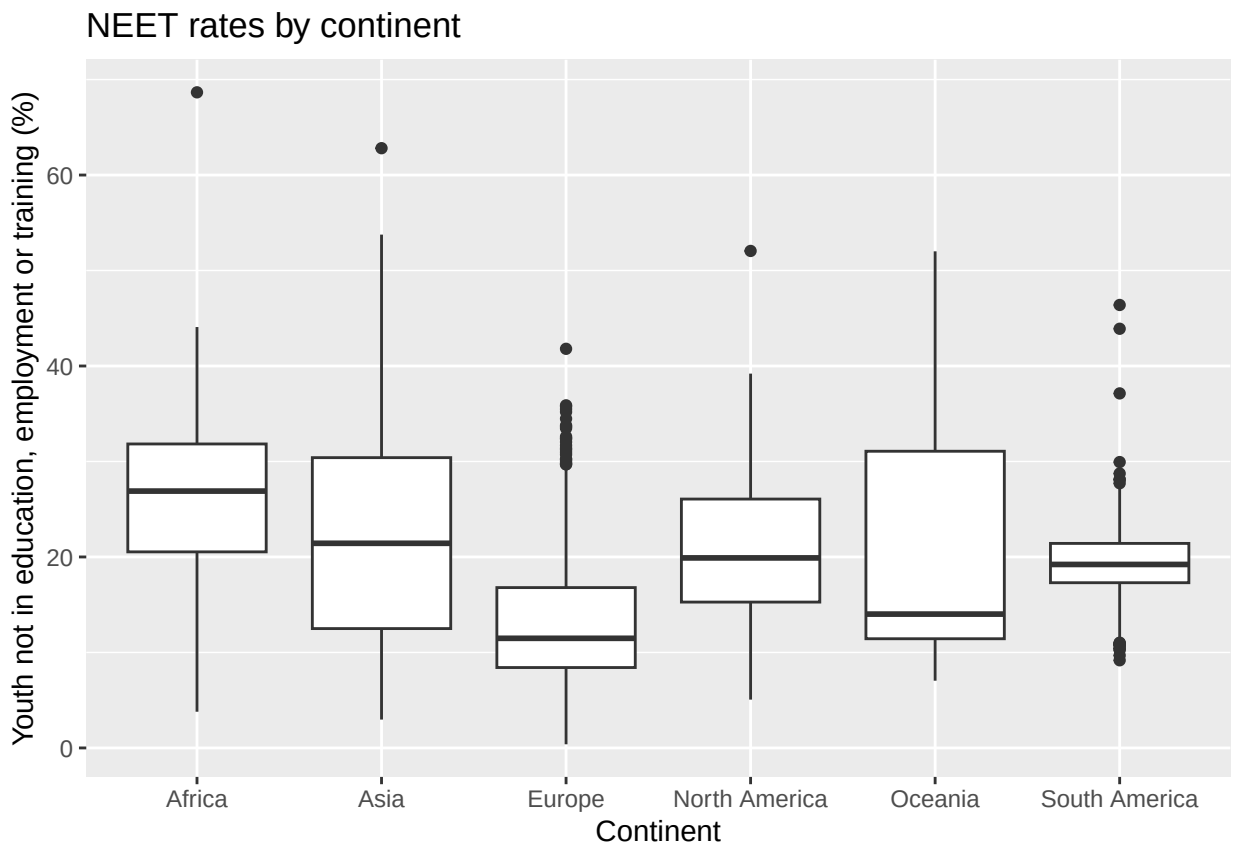
```
         !is.na(neet_pct))

summary(youth_cont$neet_pct)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.38   10.97   17.32   18.21   24.52   68.66
```

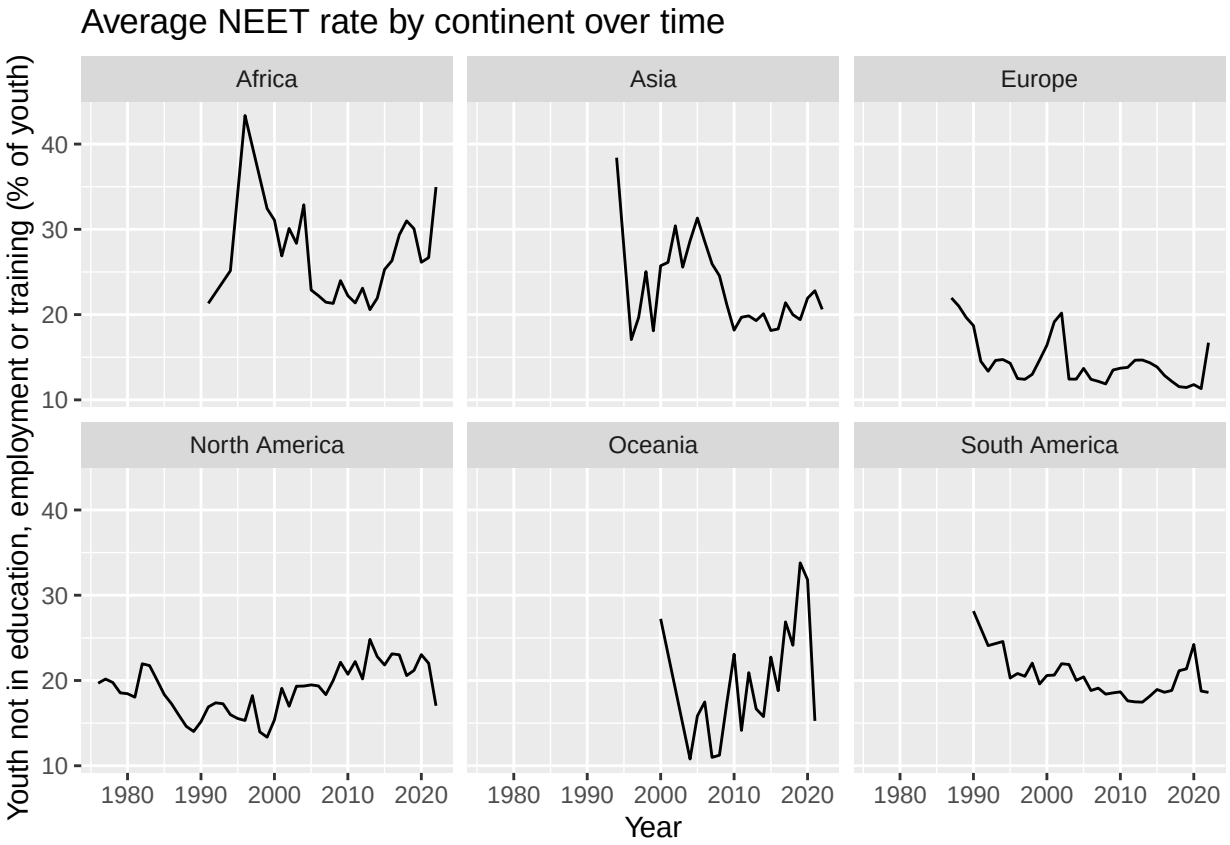
```
ggplot(youth_cont, aes(x = neet_pct)) +
  geom_histogram(binwidth = 5) +
  labs(
    title = "Distribution of NEET rates across all countries and years",
    x = "Youth not in education, employment or training (% of youth)",
    y = "Number of observations"
  )
```




```
ggplot(youth_cont, aes(x = continent, y = neet_pct)) +
  geom_boxplot() +
  labs(
    title = "NEET rates by continent",
    x = "Continent",
    y = "Youth not in education, employment or training (%)"
  )
```

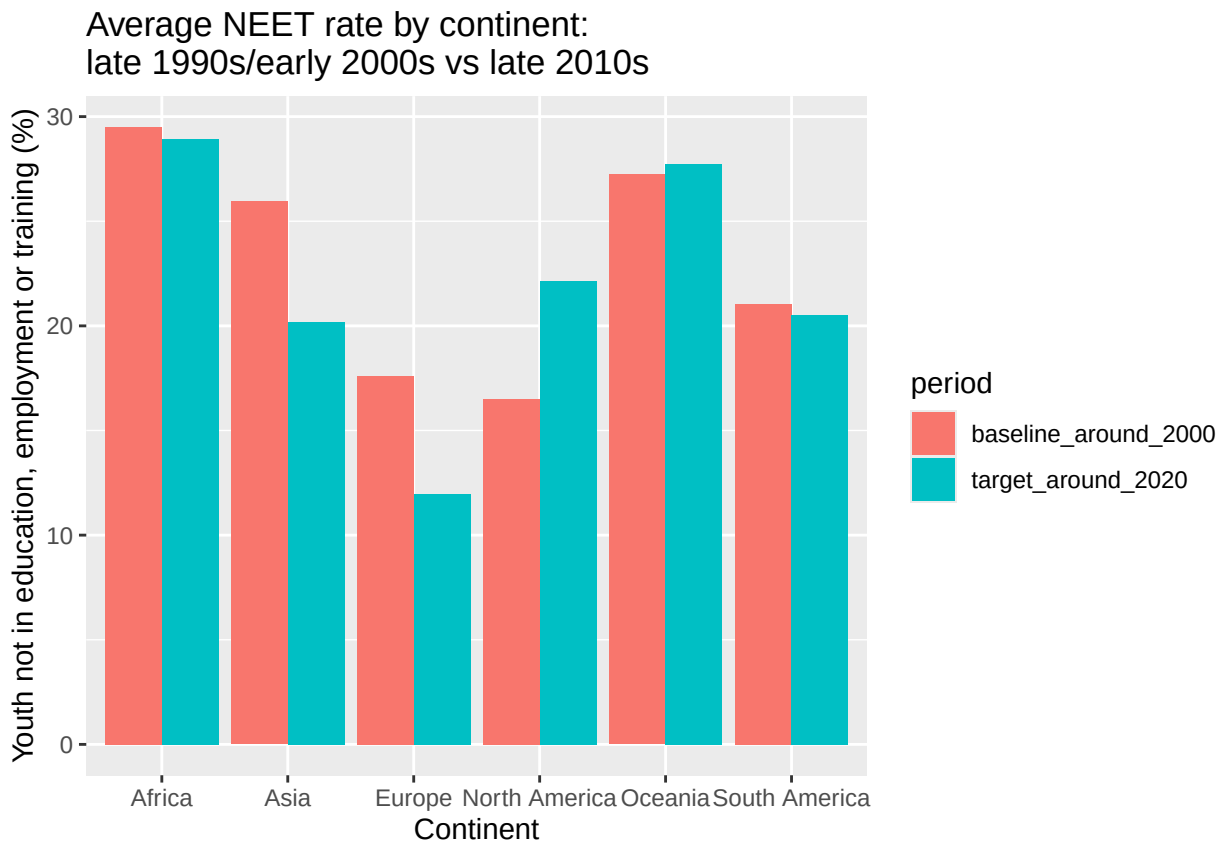



```
neet_cont_year <- youth_cont %>%
  group_by(continent, year) %>%
  summarise(
    mean_neet   = mean(neet_pct, na.rm = TRUE),
    n_countries = n(),
    .groups     = "drop"
  )

ggplot(neet_cont_year, aes(x = year, y = mean_neet)) +
  geom_line() +
  facet_wrap(~ continent) +
  labs(
    title = "Average NEET rate by continent over time",
    x = "Year",
    y = "Youth not in education, employment or training (% of youth)"
  )
```




```r
baseline_years <- 1998:2002
target_years   <- 2016:2020

neet_period_cont <- youth_cont %>%
  filter(year %in% c(baseline_years, target_years)) %>%
  mutate(
    period = case_when(
      year %in% baseline_years ~ "baseline_around_2000",
      year %in% target_years   ~ "target_around_2020"
    )
  ) %>%
  group_by(continent, period) %>%
  summarise(
    mean_neet = mean(neet_pct, na.rm = TRUE),
    .groups   = "drop"
  ) %>%
  pivot_wider(
    names_from  = period,
    values_from = mean_neet
  ) %>%
  mutate(
    abs_change  = target_around_2020 - baseline_around_2000,
    perc_change = (target_around_2020 - baseline_around_2000) /
      baseline_around_2000 * 100
  )
```

```r
neet_long_plot <- neet_period_cont %>%
  select(continent, baseline_around_2000, target_around_2020) %>%
  pivot_longer(
    cols      = c(baseline_around_2000, target_around_2020),
    names_to  = "period",
    values_to = "mean_neet"
  )

ggplot(neet_long_plot, aes(x = continent, y = mean_neet, fill = period)) +
  geom_col(position = "dodge") +
  labs(
    title = "Average NEET rate by continent:\nlate 1990s/early 2000s vs late 2010s",
    x = "Continent",
    y = "Youth not in education, employment or training (%)"
  )
```




```r
neet_country_period <- youth_cont %>%
  filter(year %in% c(baseline_years, target_years)) %>%
  mutate(
    period = case_when(
      year %in% baseline_years ~ "baseline_around_2000",
      year %in% target_years   ~ "target_around_2020"
    )
  ) %>%
  group_by(continent, country, period) %>%
```

```r
  summarise(
    mean_neet = mean(neet_pct, na.rm = TRUE),
    .groups   = "drop"
  ) %>%
  pivot_wider(
    names_from  = period,
    values_from = mean_neet
  ) %>%
  mutate(
    abs_change  = target_around_2020 - baseline_around_2000,
    perc_change = (target_around_2020 - baseline_around_2000) /
      baseline_around_2000 * 100
  )

top_improvers <- neet_country_period %>%
  arrange(abs_change) %>%
  group_by(continent) %>%
  slice_head(n = 3) %>%
  ungroup()

top_worseners <- neet_country_period %>%
  arrange(desc(abs_change)) %>%
  group_by(continent) %>%
  slice_head(n = 3) %>%
  ungroup()

neet_2019_2020 <- youth_cont %>%
  filter(year %in% c(2019, 2020)) %>%
  group_by(continent) %>%
  summarise(
    mean_neet_2019_2020 = mean(neet_pct, na.rm = TRUE),
    .groups             = "drop"
  )

summary_table <- neet_period_cont %>%
  left_join(neet_2019_2020, by = "continent") %>%
  arrange(target_around_2020)

summary_table
```

```
## # A tibble: 6 x 6
##   continent     baseline_around_2000 target_around_2020 abs_change perc_change
##   <chr>                        <dbl>              <dbl>      <dbl>       <dbl>
## 1 Europe                        17.6               12.0     -5.64       -32.0
## 2 Asia                          25.9               20.2     -5.74       -22.1
## 3 South America                 21.0               20.5     -0.526       -2.50
## 4 North America                 16.5               22.2      5.65        34.2
## 5 Oceania                       27.2               27.7      0.501        1.84
## 6 Africa                        29.5               28.9     -0.573       -1.94
## # i 1 more variable: mean_neet_2019_2020 <dbl>
```