Hacettepe University
Department of Industrial Engineering
Undergraduate Program
2023-2024 Fall

EMU 430 – Data Analytics
Week 7
November 17, 2023

Instructor: Erdi Dasdemir

edasdemir@hacettepe.edu.tr
www.erdidasdemir.com

Outline

Acknowledgment

I drew inspiration primarily from Dr. Rafael Irizarry's "Introduction to Data Science" Book

and "Data Science" course by HarvardX on edX for the slides this week.

Previously on

EMU430

Basics of ggplot2

- load the ggplot2

```
library(ggplot2)
```

- We can also load the ggplot2 package by loading `tidyverse` package.

```
library(tidyverse)
```

- `tidyverse` includes useful packages like `dplyr` in addition to `ggplot2`.

- **Components**

1. Data
2. Geometry components (scatter plot, bar plot, histograms, smooth densities, q-q plots, box plots)
3. Aesthetic mapping (x-axis: population size, y-axis: total number of murders, text: identify states, colors: denote four different regions
4. Scales: we generally scale ranges of the x and y-axis.
5. Labels, Title, Legend, Themes, etc. → defines style

- **Creating a new plot with Data Component**

```
library(tidyverse)
library(dslabs)
data(murders)

# first option
ggplot(data=murders)

# second option
murders %>% ggplot()
```

- **Creating a new plot with Data Component**

```
library(tidyverse)
library(dslabs)
data(murders)

# first option
ggplot(data=murders)

# second option
murders %>% ggplot()
```

Layers: Geometry and Aesthetic Mapping

- In general, a line of code in ggplot will look like this:

```
data %>% ggplot() + layer 1 + layer 2 + ... + layer n
```

- For `geom`, we need to provide **data** and **mapping**.

```
?geom_point()
```

data

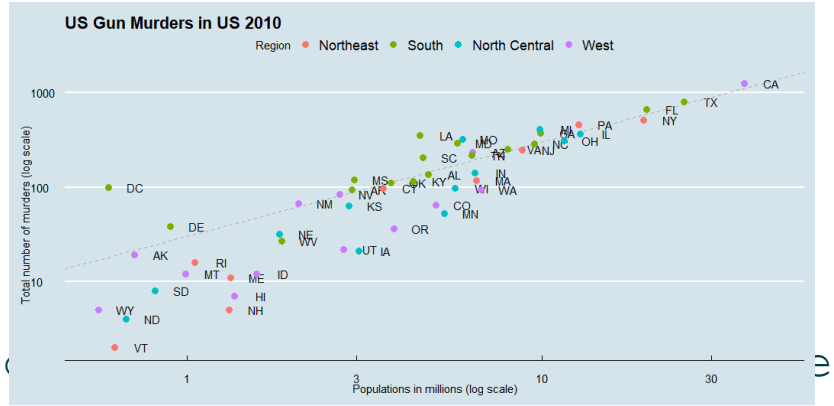
```
p <- murders %>% ggplot()
```

mapping

aes: this function connects data with what we see on the graph. we will use this frequently.

aesthetic mapping:

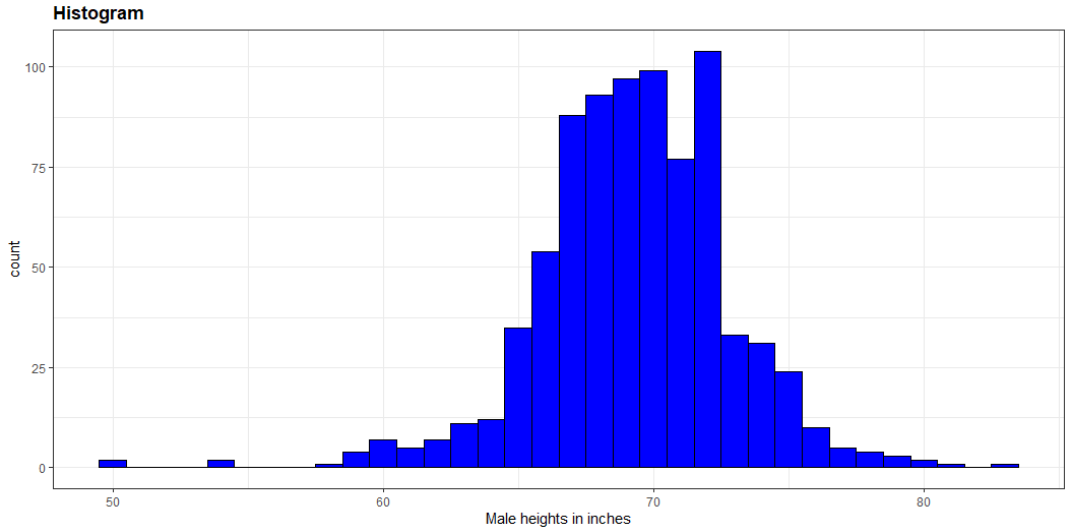
```
murders %>% ggplot() + geom_point(aes(x = population/10^6, y = total))
```




```
o library(ggthemes)
  library(ggrepel)

  ### first define the slope of the line
  r <- murders %>% summarize(rate = sum(total) / sum(po                    e

  ## now make the plot.
  murders %>% ggplot(aes(population/10^6, total, label = abb)) +
      geom_abline(intercept = log10(r), lty = 2, color = "darkgrey") +
      geom_point(aes(col = region), size = 3) +
      geom_text_repel() +
      scale_x_log10() +
      scale_y_log10() +
      xlab("Populations in millions (log scale)") +
      ylab("Total number of murders (log scale)") +
      ggtitle("US Gun Murders in US 2010") +
      scale_color_discrete(name = "Region") +
      theme_economist()
```

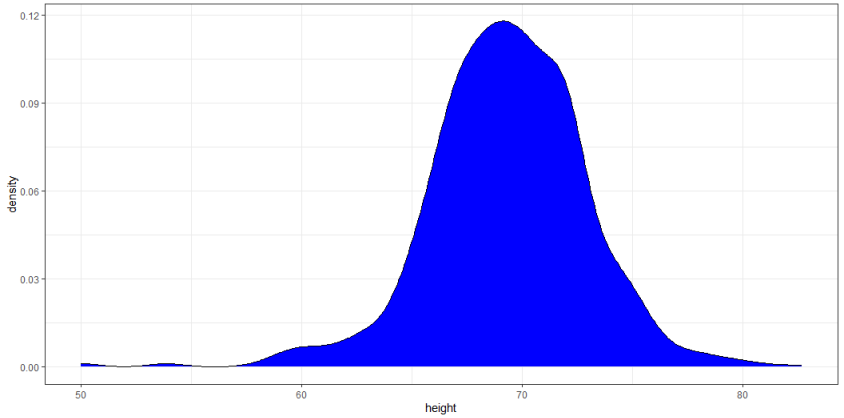
historgram

Let's make the histogram for the male heights.




```
p <- heights %>% filter(sex == "Male")

p <- p %>% ggplot(aes(x = height))
```

```
p + geom_histogram(binwidth = 1, fill = "blue", col = "black") + xlab("Male heights in inches") +

ggtitle("Histogram")
```




```
p + geom_density(fill="blue")
```

- `summarize`
- `group_by`
- `dot placeholder .$->` access resulting values
- `arrange` –> examine data after sorting

```
library(tidyverse)
library(dslabs)
data("heights")
# compute average and standard deviation for males
s <- heights %>% filter(sex == "Male") %>% summarize(average = mean(height), standard_deviation
= sd(height))
s
   average standard_deviation
1 69.31475           3.611024

# The resulting table stored in s is a data frame -> we can access the components with the
accessor dollar sign.
s$average
[1] 69.31475
s$standard_deviation
[1] 3.611024
```

- Most of dplyr functions always return data frames.What if we need numeric value?

```
data("murders")
us_murder_rate <- murders %>% summarize(rate = sum(total) / sum(population)*100000)

us_murder_rate %>% .$rate
[1] 3.034555

us_murder_rate$rate
[1] 3.034555
```

- We split data to groups, then compute summaries for each group

```
heights %>% group_by(sex)
# A tibble: 1,050 × 2
# Groups:   sex [2]
   sex    height
   <fct>   <dbl>
 1 Male       75
 2 Male       70
 3 Male       68
 4 Male       74
 5 Male       61
 6 Female     65
 7 Female     66
 8 Female     62
 9 Female     66
10 Male       67
# ⓘ 1,040 more rows
```

- This is a special data frame called group data frame.
- dplyr functions, particularly summarize, will behave differently when acting on this object.
- Conceptually, you can think this object as many tables with the same columns but not necessarily the same rows that are stacked into one object.

Mean and Standard Deviation

```
heights %>% group_by(sex) %>% summarize(average = mean(height), standard_deviation = sd(height))
```

```
# A tibble: 2 × 3
  sex    average standard_deviation
  <fct>    <dbl>              <dbl>
1 Female    64.9               3.76
2 Male      69.3               3.61
```

- sort table by different columns.
- we already know about the order and sort functions. dplyr has useful `arrange`

- **Order states by their population size:**

```
murders %>% arrange(population) %>% head()
```

	state	abb	region	population	total	murder_rate
1	Wyoming	WY	West	563626	5	0.8871131
2	District of Columbia	DC	South	601723	99	16.4527532
3	Vermont	VT	Northeast	625741	2	0.3196211
4	North Dakota	ND	North Central	672591	4	0.5947151
5	Alaska	AK	West	710231	19	2.6751860
6	South Dakota	SD	North Central	814180	8	0.9825837

World Health and Economics

Case Study

Case Study

- We will demonstrate how relatively simple `ggplot` and `dplyr` code can create insightful and aesthetically pleasing plots that help us better understand trends in world health and economics.

- We will use data from https://www.gapminder.org/

- Hans Rosling –> co-founder of the Gapminder Foundation, an organization dedicated to educating the public using data to dispel common myths about the so-called developing world.

- They use data to show how actual health and economic trends contradict the narratives originating from sensationalist media.

Gapminder Data Set

"Journalists and lobbyists tell dramatic stories. That's their job. They tell stories about extraordinary events and unusual people. The piles of dramatic stories pile up in people's minds into an overdramatic worldview and strong negative stress feelings"

"The world is getting worse!"
"It's we versus them!"
"Other people are strange!"
"The population just keeps growing!"
"Nobody cares!"

Gapminder Data Set

Can we confirm this by data? We will try to answer the following two questions:

1. Is it fair to say the world is divided into rich (Western nations) and poor (the developing world in Africa, Asia, and Latin America)?

2. Has income inequality across countries worsened during the last 40 years?

Gapminder Data Set

- The data set was put together by `dslabs` library, and it was created using a number of spreadsheets available from the Gapminder foundation.

```
library(dslabs)
library(tidyverse)
data(gapminder)
head(gapminder)
```

```
> head(gapminder)
                country year infant_mortality life_expectancy fertility population          gdp continent          region
1               Albania 1960           115.40           62.87      6.19    1636054           NA    Europe Southern Europe
2               Algeria 1960           148.20           47.50      7.65   11124892  13828152297    Africa Northern Africa
3                Angola 1960           208.00           35.98      7.32    5270844           NA    Africa   Middle Africa
4 Antigua and Barbuda 1960               NA           62.97      4.43      54681           NA  Americas       Caribbean
5             Argentina 1960            59.87           65.39      3.11   20619075 108322326649  Americas   South America
6               Armenia 1960               NA           66.86      4.55    1867396           NA      Asia    Western Asia
```

- We will test our knowledge regarding differences in child mortality across different countries.

Gapminder Case Study Quiz

Hans Rosling asked these question in his video “New Insights on Poverty.”

Q1. For each of the pairs of countries here, which country had the highest child mortality rate in 2005?

Q2. Which pair do you think are most similar?

Siri Lanka or Turkey

Poland or South Korea

Malaysia or Russia

Pakistan or Vietnam

Thailand or South Africa

- **Q1.** Typically selections are non-European countries, e.g. Sri Lanka, South Korea, Malaysia
- **Q2.** Countries part of the developing world, Pakistan, Vietnam, Thailand, and South Africa have similarly high mortality rates.

Compare Siri Lanka and Türkiye

```
gapminder %>% filter(year == 2015 & country %in% c("Sri Lanka", "Turkey")) %>%
select(country, infant_mortality)
```

```
    country infant_mortality
1 Sri Lanka              8.4
2    Turkey             11.6
```

Compare Pairs

country_1	infant_mortality	country_2	infant_mortality
Sri Lanka	8.4	Turkey	11.6
Poland	4.5	South Korea	2.9
Malaysia	6.0	Russia	8.2
Pakistan	65.8	Vietnam	17.3
Thailand	10.5	Sotuh Africa	33.6

- European countries have higher rates: Turkey and Poland have higher rates compared to their pairs.
- Countries from the developing world can have very different rates: Pakistan is very different from Vietnam. South Africa is very different from Thailand.
- Many people fail this quiz: This implies that we are more than ignorant, we are misinformed.

Our misconceptions stem from the preconceived notion that the world is divided into two groups:

1. the Western World (W. Europe and N. America) –> Long life spans and small families

2. Developing world (Africa, Asia and Latin America) –> Short life spans and large families

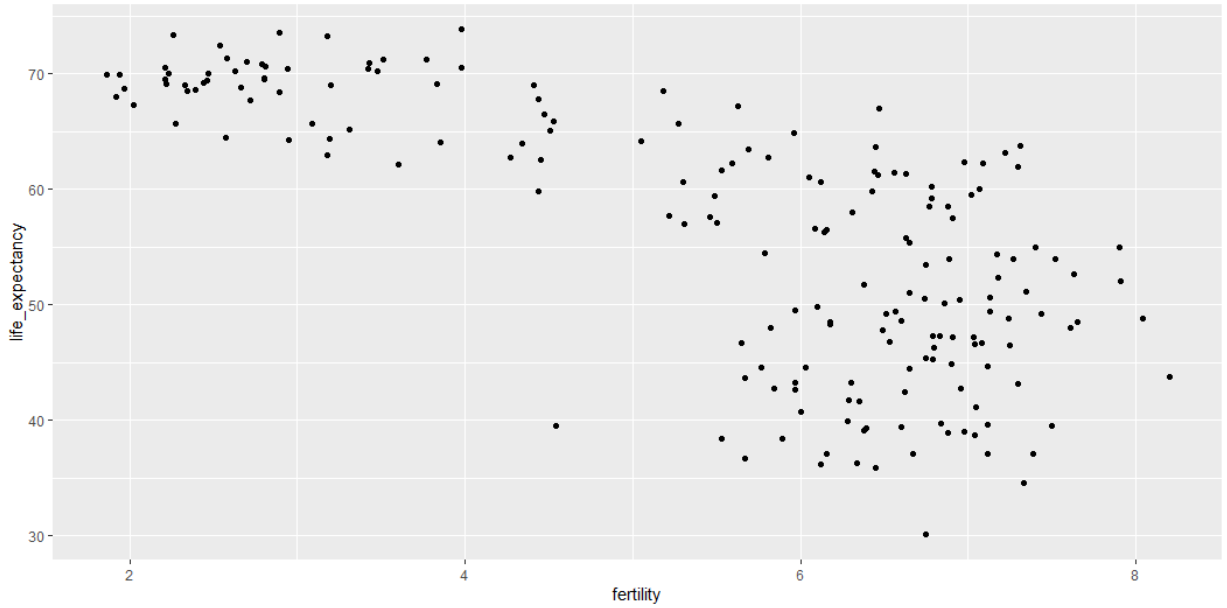
Does data support this?

We can draw a scatter plot of life expectancy versus fertility rates (average number of children per women):

Gapminder Life Expextanct and Fertility

Let's start with an older year, 1962.

```
gapminder %>% filter(year == 1962) %>% ggplot(aes(fertility, life_expectancy)) +
geom_point()
```



We observe two distinct categories.

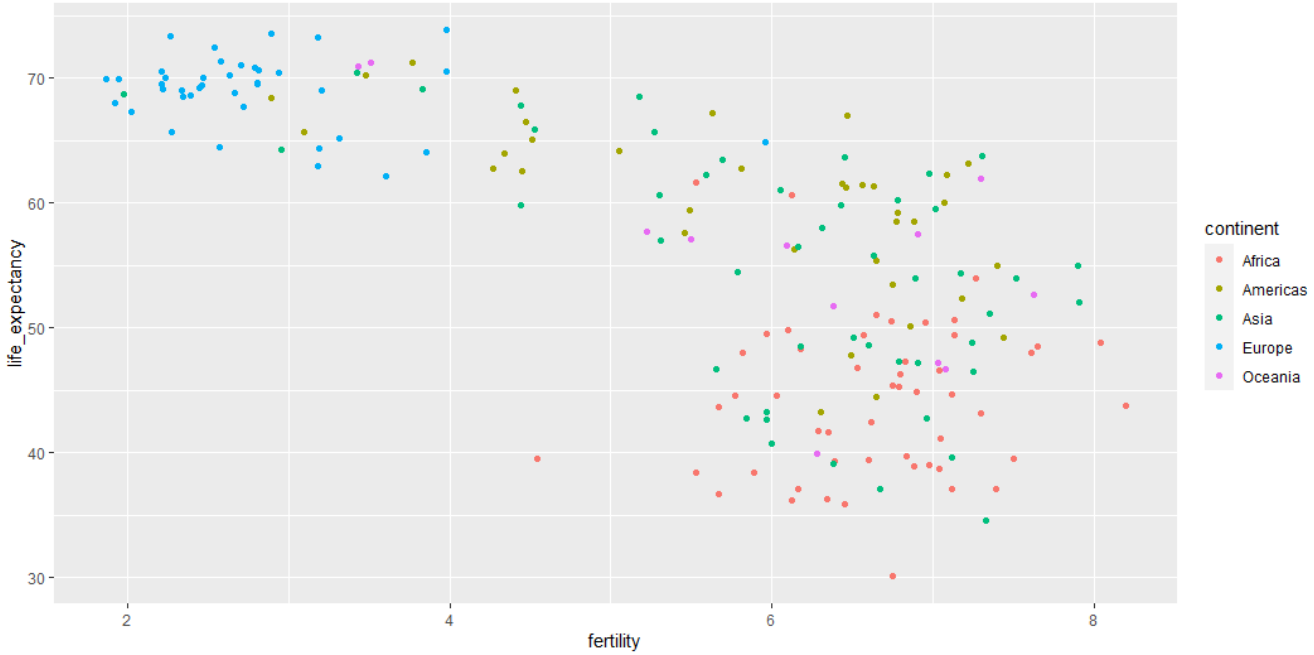
1. Life expectancy is around 70 years, with three or fewer children per family.

2. Life expectancy is lower than 65 years, more than five children per family.

To confirm indeed these countries are from the regions we expect, we can use color to represent continent.

Gapminder Life Expextanct and Fertility

```
gapminder %>% filter(year == 1962) %>% ggplot(aes(fertility, life_expectancy, color =
continent)) + geom_point()
```



in 1962, it is correct that the world is divided into two: West and developing countries, **but is this still the case 50 years later, for example, in 2012?**

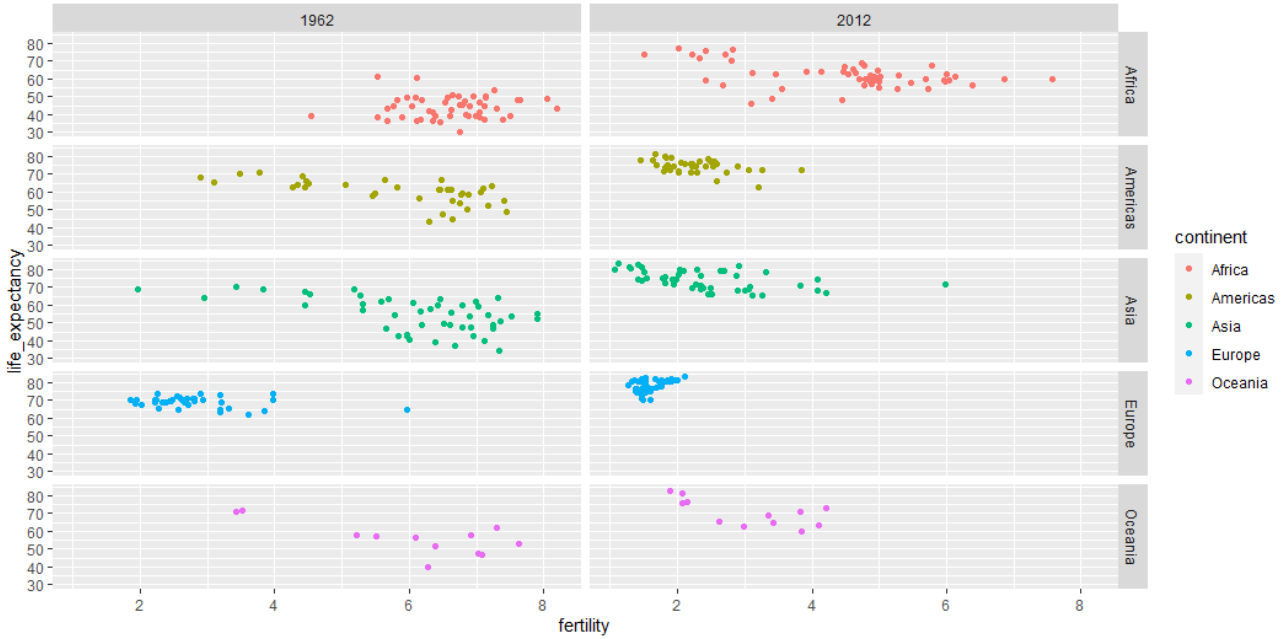
It would be nice to see two plots together side by side. –> **faceting**

Faceting

Faceting

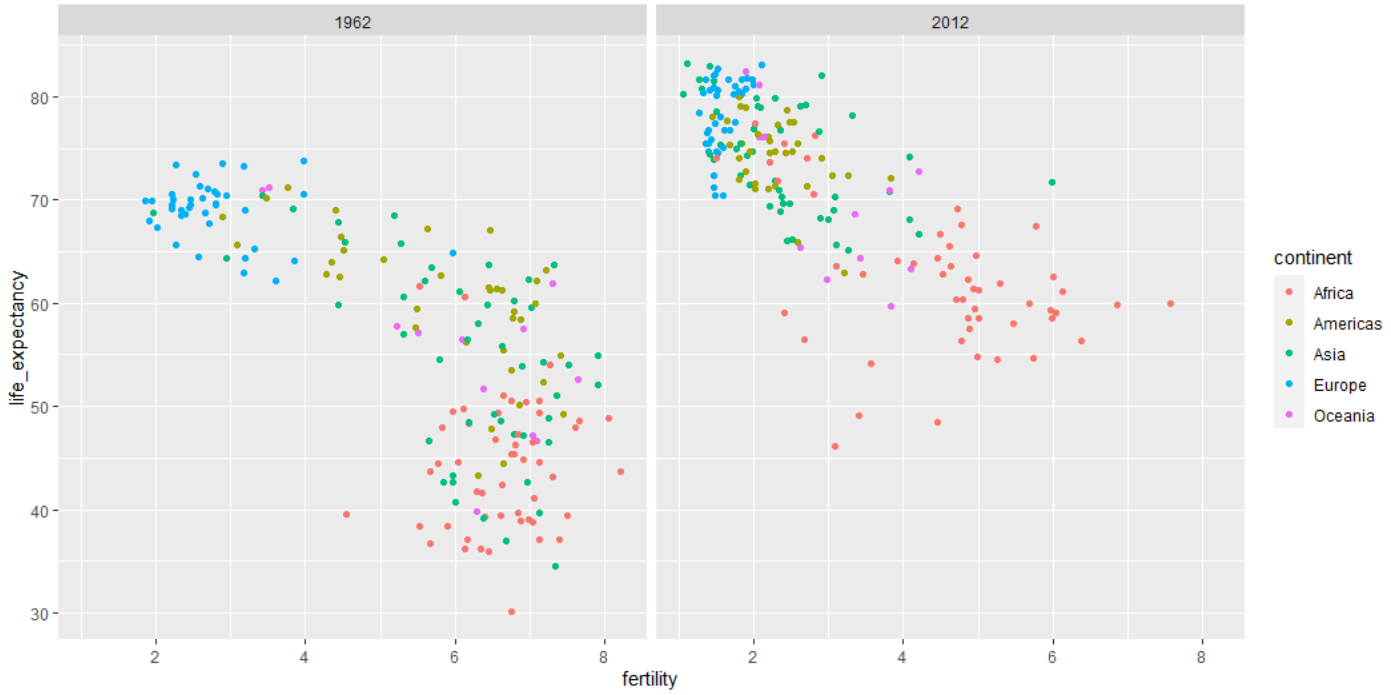
- We will facet by year using the `facet_grid()` function as an additional layer to our plot.
- `facet_grid()` lets us facet by up to 2 variables, columns to represent one variable and rows to represent the other.

```
gapminder %>% filter(year %in% c(1962, 2012)) %>% ggplot(aes(fertility, life_expectancy, color = continent)) + geom_point() + facet_grid(continent ~ year)
```



We can ignore continents and facet only using year.

```
gapminder %>% filter(year %in%
c(1962, 2012)) %>%
ggplot(aes(fertility,
life_expectancy, color = continent))
+ geom_point() + facet_grid(.~year)
```

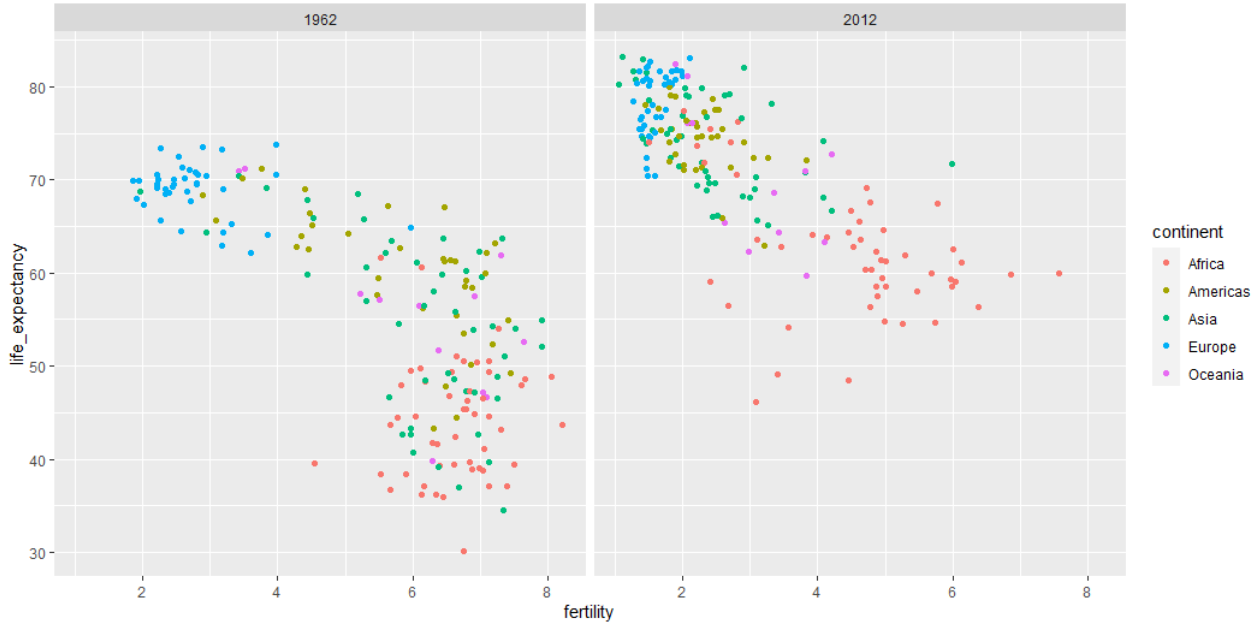


The majority of countries have moved from the developing world cluster to Western wold one.

In 2012, the Western versus developing worldview no longer makes sense.

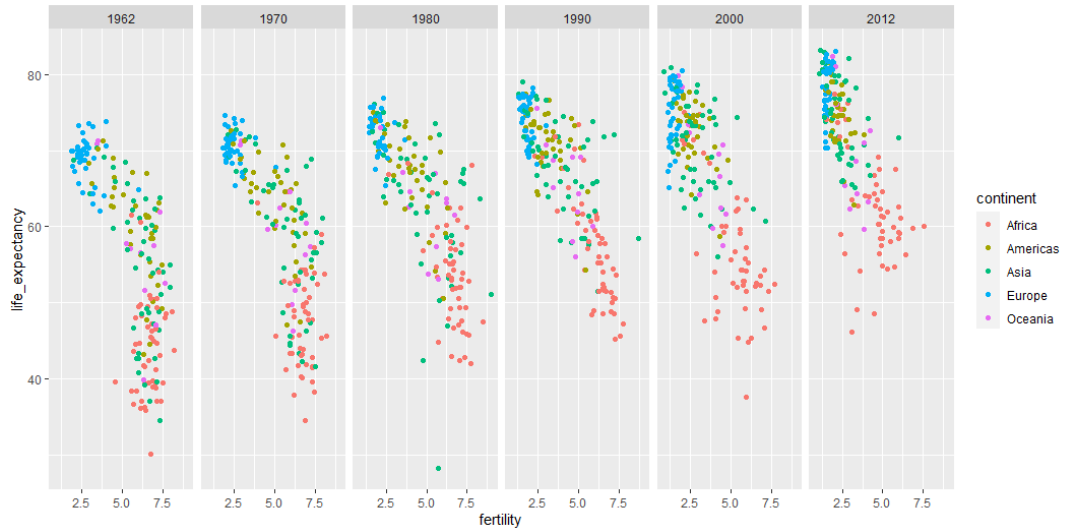
This is particularly clear when we compare Europe to Asia.

Asian countries made great improvements in the last 50 years.



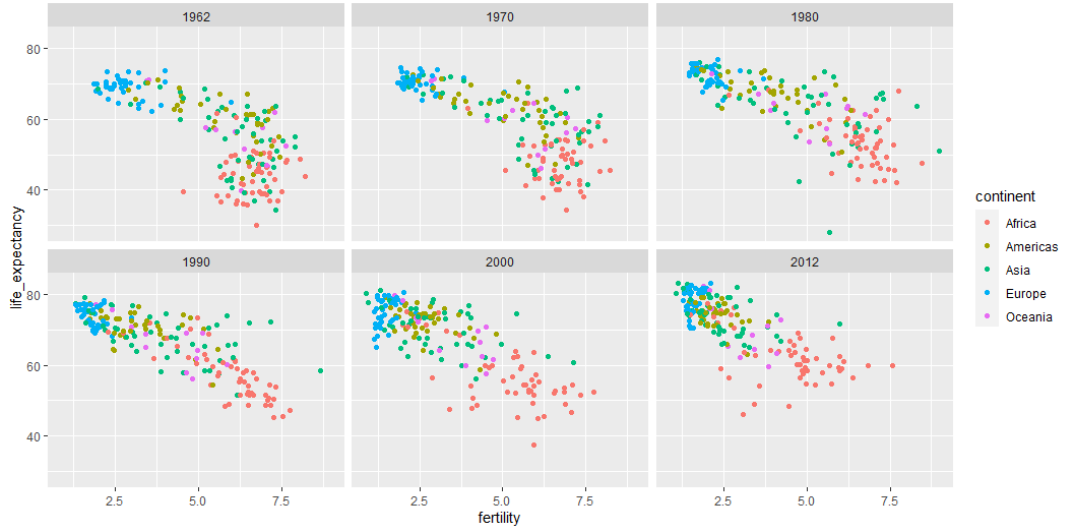
- To observe this improvement over multiple years, we can add more years to our plot.

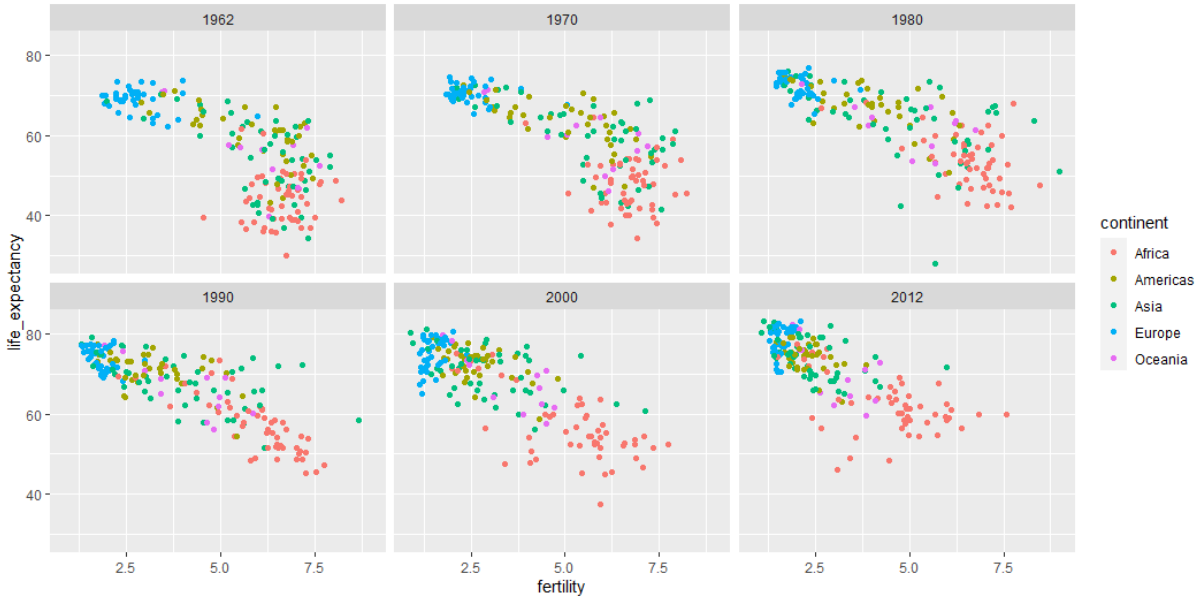
```
gapminder %>% filter(year %in% c(1962, 1970, 1980, 1990, 2000, 2012)) %>%
ggplot(aes(fertility, life_expectancy, color = continent)) + geom_point() + facet_grid(.~year)
```



- **facet_wrap() function**

```
gapminder %>% filter(year %in% c(1962, 1970, 1980, 1990, 2000, 2012)) %>%
ggplot(aes(fertility, life_expectancy, color = continent)) + geom_point() + facet_wrap(~year)
```



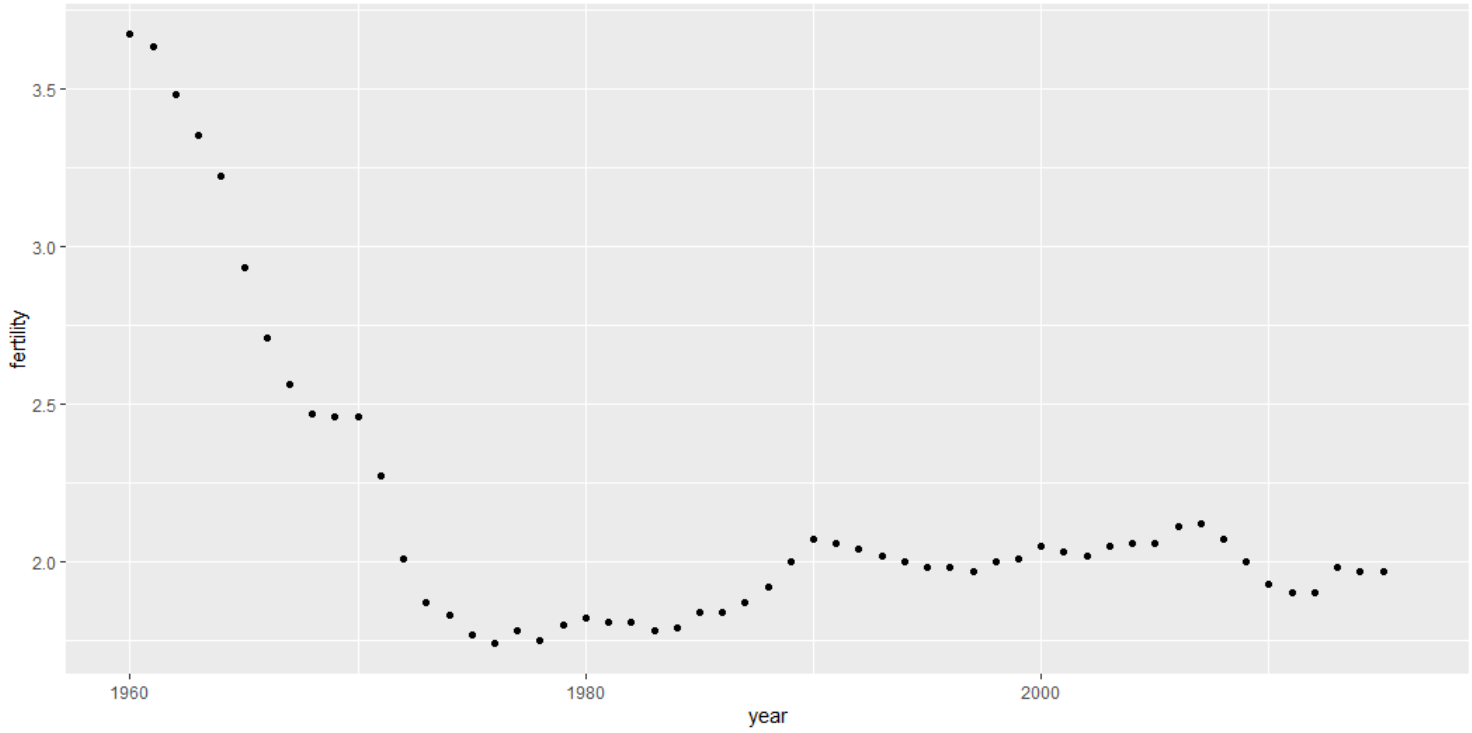


- Look at the Asian countries!
- Look at the ranges of axes. The range is determined by the data shown in all plots when the facet is used. Having the same range helps us for comparison.

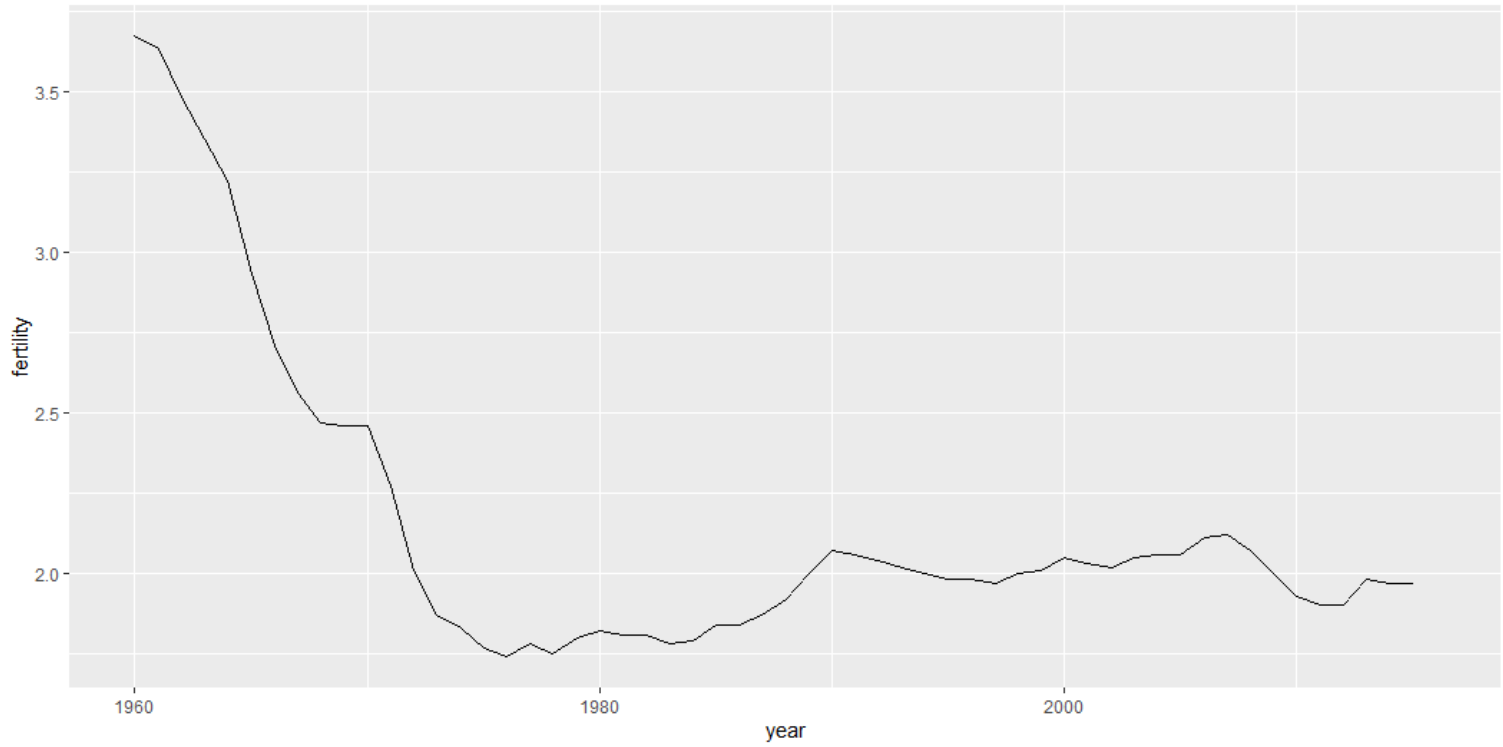
Time Series Plots

- We see that there is no distinct separation between the continents anymore.
- However, new questions arise.
- Which countries are improving more? Which ones are improving less? Was the improvement constant during the last 50 years, or was there more than an acceleration during a specific certain period?
- Time series plots have time on the x-axis, and the measure of interest on the y-axis.
- We can use a time series plot to investigate the fertility rates of a specific country.

```
o gapminder %>% filter(country == "United States") %>%
  ggplot(aes(x = year, y = fertility)) + geom_point()
```

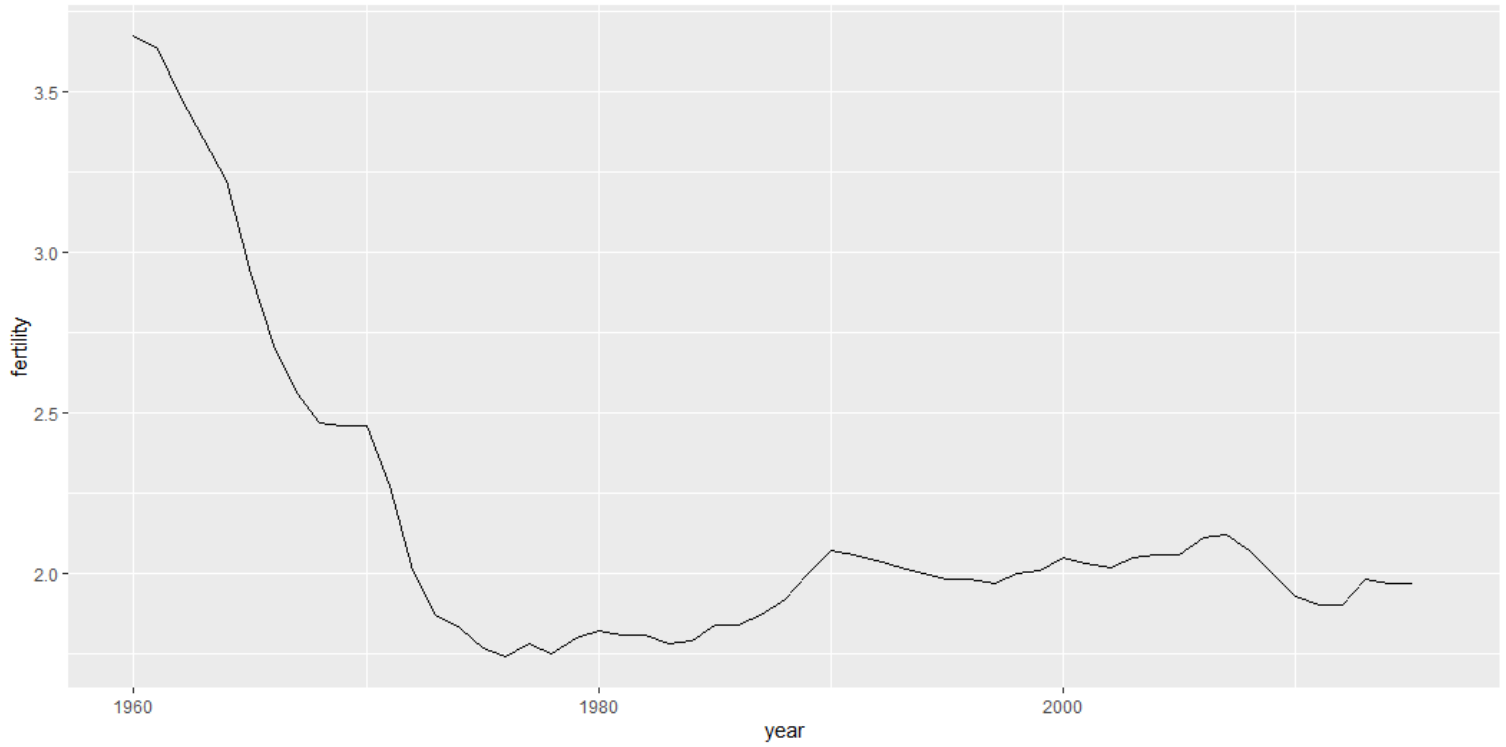


```
gapminder %>% filter(country == "United States") %>%
ggplot(aes(x=year, y=fertility)) + geom_line()
```



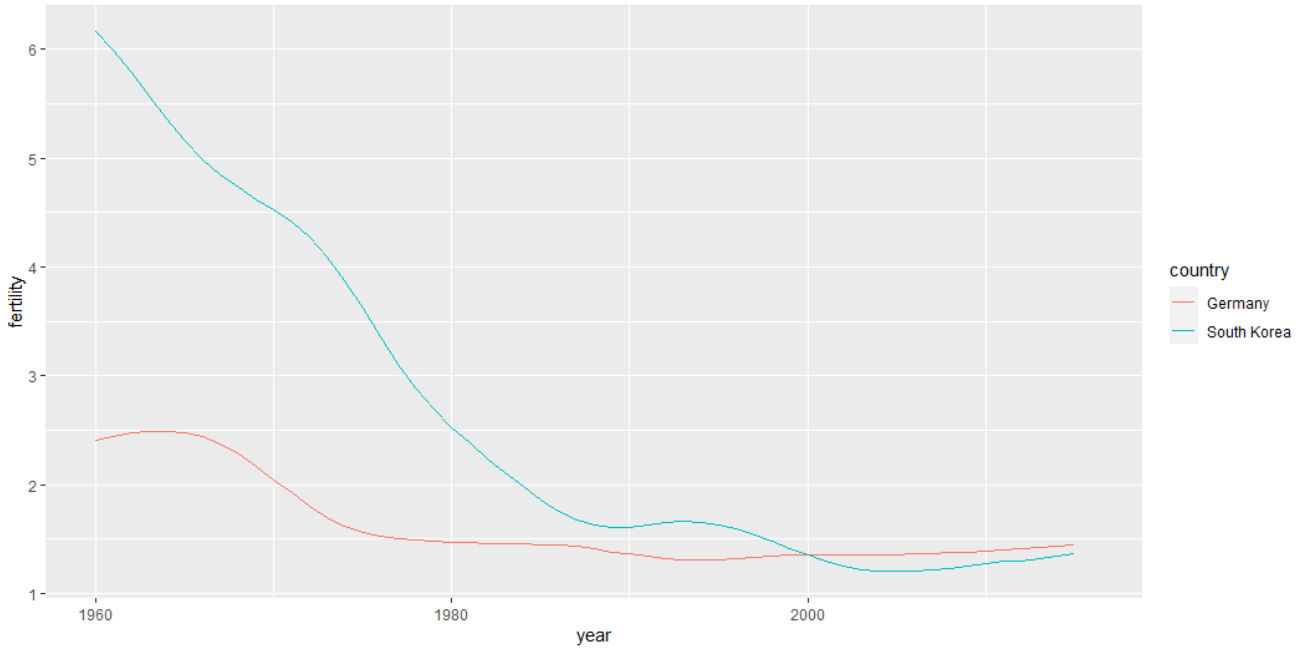
Compare two countries (South Korea and Germany) with a time series plot

```
countries <- c("South Korea", "Germany")

gapminder %>% filter(country %in% countries) %>% ggplot(aes(year,
fertility, group = country)) + geom_line()
```



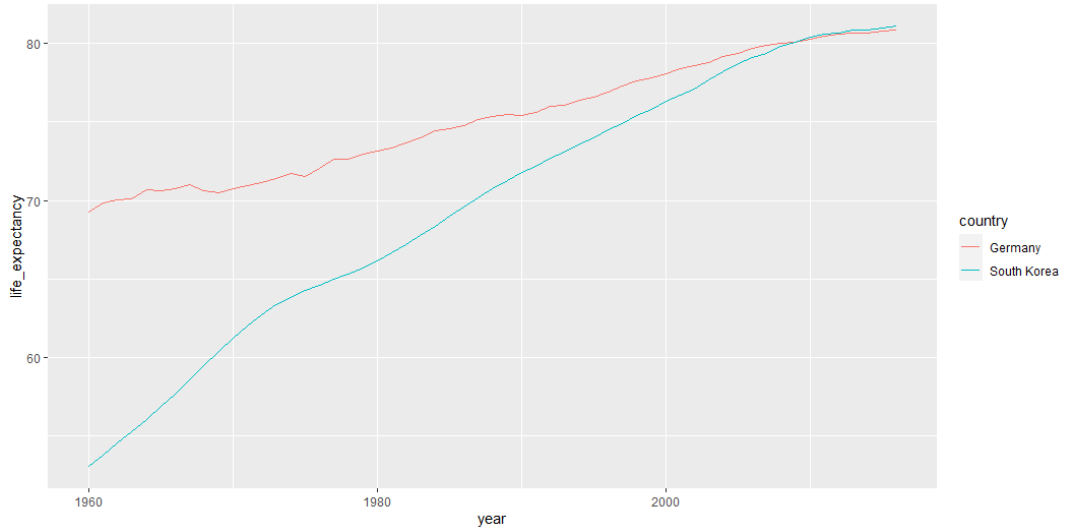
- We do not know which line belongs to which country.
- We can use colors.
- The good thing about using colors is that ggplot automatically groups data.

```
countries <- c("South Korea", "Germany")
gapminder %>% filter(country %in%
countries) %>% ggplot(aes(year,
fertility, color = country)) +
geom_line()
```



Let us look at life expectancy.

```
countries <- c("South Korea", "Germany")
gapminder %>% filter(country %in%
countries) %>% ggplot(aes(year,
life_expectancy, color = country)) +
geom_line()
```



Box Plots

Wealth Distribution

- Another common belief is that the wealth distribution across the world has become worse during the last decades:

  Rich countries become richer, poor countries become poorer. We will next look at this.

transformations

- We have `gdp` in our `gapminder` data table.
- GDP measures the market value of good and services produced by a country in given year.
- We can find GDP per person for per day.
- `gapminder <- gapminder %>% mutate(dollars_per_day = gdp/population/365)`

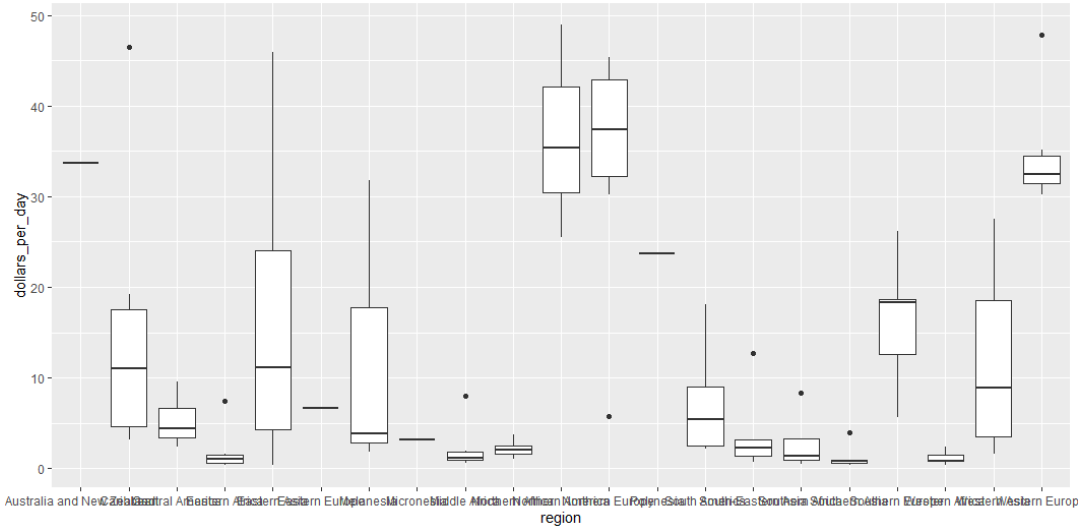
There are 22 regions and we cannot look at them individually using histograms or smooth densities.

```
length(levels(gapminder$region))
```

```
[1] 22
```

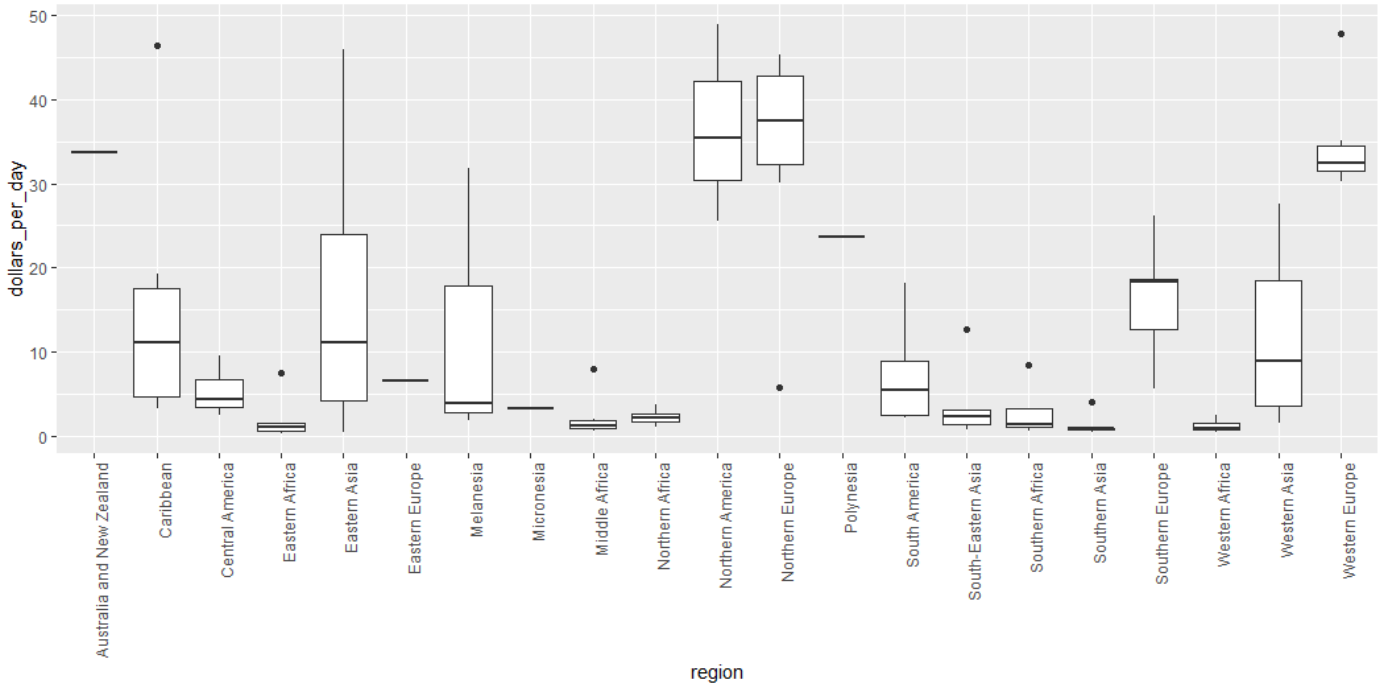
We can stack box plots next to each other.

```
past_year = 1970
p <- gapminder %>% filter(year == past_year & !is.na(gdp)) %>% ggplot(aes(region,
dollars_per_day))
p + geom_boxplot()
```



We cannot read x axis.

```
past_year = 1970
p <- gapminder %>% filter(year == past_year & !is.na(gdp)) %>% ggplot(aes(region,
dollars_per_day))

p + geom_boxplot() + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



We can see that there is indeed a west versus the rest.

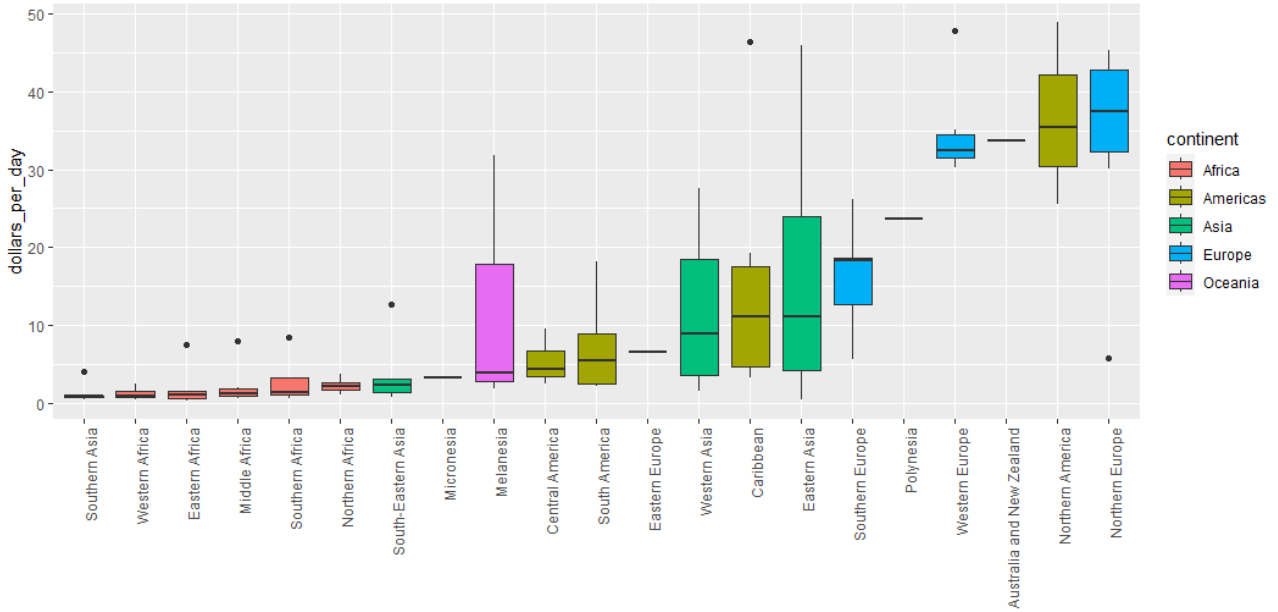
The order is alphabetically.

We can do something more meaningful.

We will use `reorder()` function.

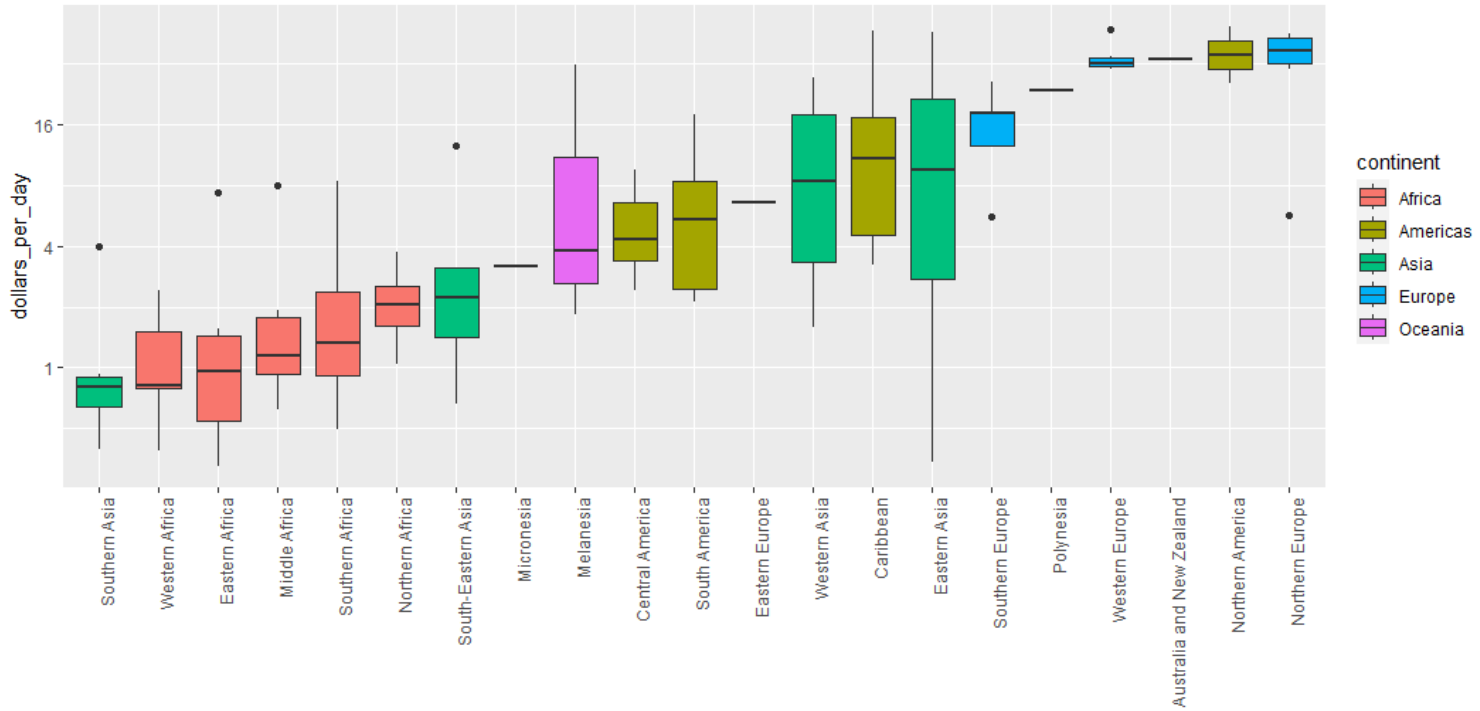
Reorder regions by their median income levels

```
past_year = 1970


p <- gapminder %>% filter(year ==
past_year & !is.na(gdp)) %>%
mutate(region = reorder(region,
dollars_per_day, FUN = median)) %>%
ggplot(aes(region, dollars_per_day,
fill = continent)) + geom_boxplot() +
theme(axis.text.x = element_text(angle
= 90, hjust = 1)) + xlab("")
p
```



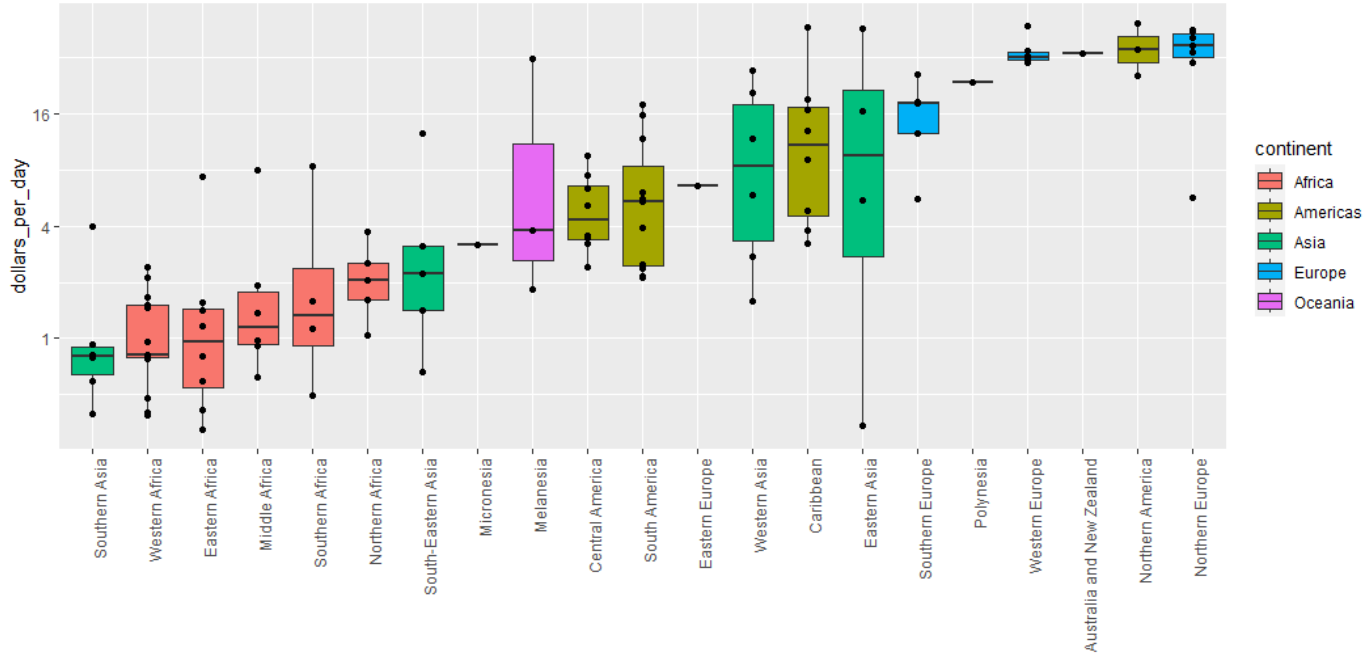
Change scale to log scale

```
p + scale_y_continuous(trans = "log2")
```



Show the data

```
p + scale_y_continuous(trans = "log2") + geom_point(show.legend = FALSE)
```



Comparing Distributions

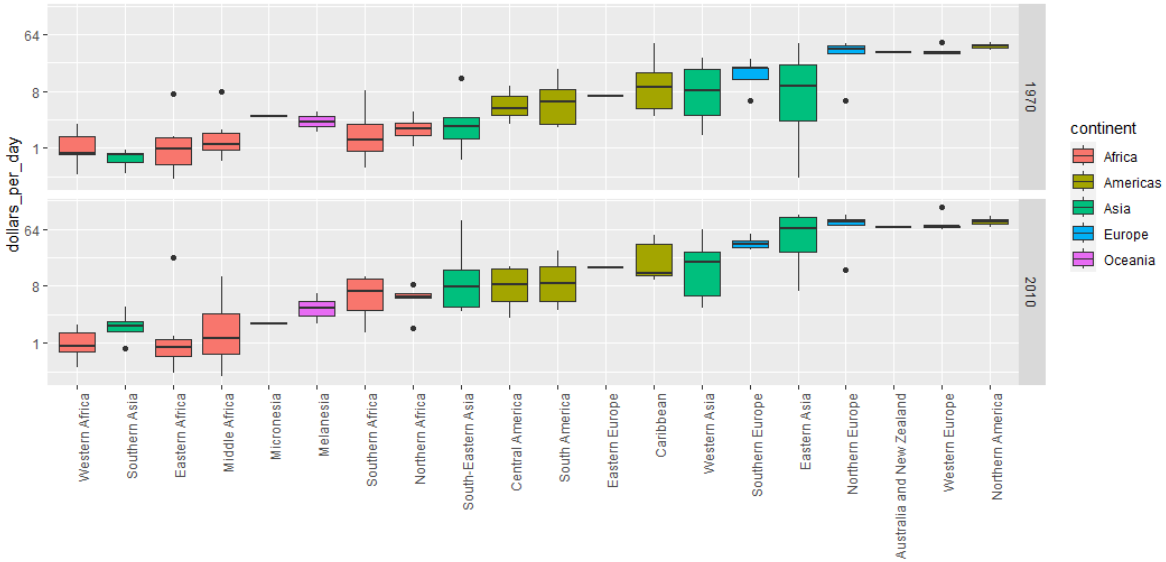
- Exploratory Data Analysis we have done so far has revealed two characteristics about average income distributions in 1970.
- Box plot showed that rich countries were mostly in Europe and Northern America.
- We also want to see the results for 2010 to make comparison.

In order to compare two years, we need to make sure that the list of countries are the same in these two years.

```
country_list_1 <- gapminder %>% filter(year == 1970 & !is.na(dollars_per_day))
%>% .$ country

country_list_2 <- gapminder %>% filter(year == 2010 & !is.na(dollars_per_day))
%>% .$ country

country_list <- intersect(country_list_1, country_list_2)
```

Now lets draw a box plot:

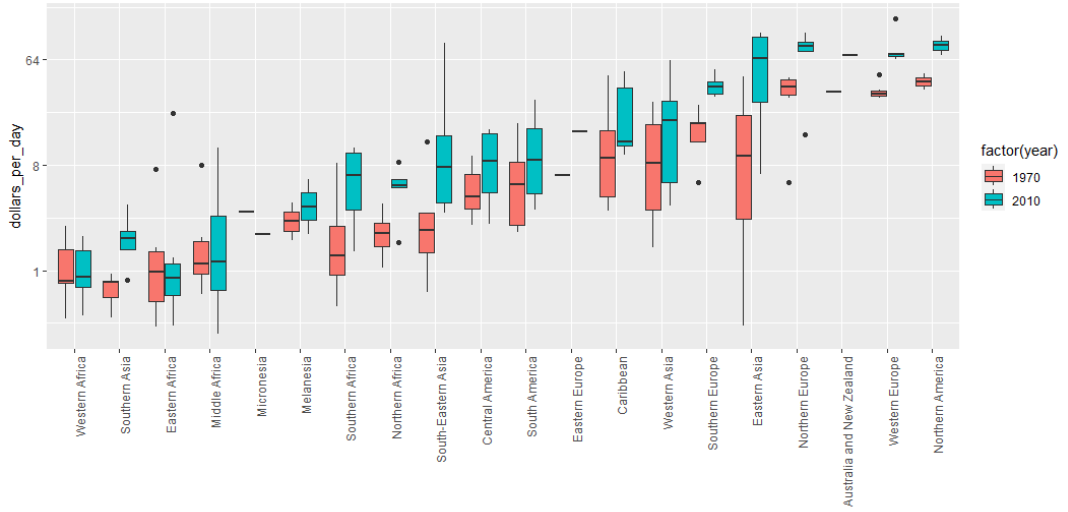
```
p <- gapminder %>% filter(year %in% c(1970,
2010) & country %in% country_list) %>%
mutate(region = reorder(region,
dollars_per_day, FUN = median)) %>% ggplot()
+ theme(axis.text.x = element_text(angle =
90, hjust = 1)) + xlab(" ") +
scale_y_continuous(trans = "log2")

p + geom_boxplot(aes(region, dollars_per_day,
fill = continent)) + facet_grid(year~.)
```



It is hard to interpret. We want box plots next to next each other.

Ease comparisons

```
p <- gapminder %>% filter(year %in% c(1970,
2010) & country %in% country_list) %>%
mutate(region = reorder(region,
dollars_per_day, FUN = median)) %>% ggplot()
+ theme(axis.text.x = element_text(angle =
90, hjust = 1)) + xlab(" ") +
scale_y_continuous(trans = "log2")

p + geom_boxplot(aes(region, dollars_per_day,
fill = factor(year)))
```




Any
Questions