SDS Review Ch8

2017100057 / 이영노

October 15, 2022

Chapter 8 : Data Visualization in Practice

Chapter8 Intro: 1) faceting, 2) time series plots, 3) ridge plots

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.7      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.0
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dslabs)
data(gapminder)
gapminder %>% as_tibble()
```

```
## # A tibble: 10,545 x 9
##    country   year infant_mortality life_expectancy fertility population      gdp
##    <fct>    <int>            <dbl>           <dbl>     <dbl>      <dbl>    <dbl>
##  1 Albania   1960            115.             62.9      6.19    1636054 NA
##  2 Algeria   1960            148.             47.5      7.65   11124892  1.38e10
##  3 Angola    1960            208              36.0      7.32    5270844 NA
##  4 Antigua~  1960             NA              63.0      4.43      54681 NA
##  5 Argenti~  1960             59.9            65.4      3.11   20619075  1.08e11
##  6 Armenia   1960             NA              66.9      4.55    1867396 NA
##  7 Aruba     1960             NA              65.7      4.82      54208 NA
##  8 Austral~  1960             20.3            70.9      3.45   10292328  9.67e10
##  9 Austria   1960             37.3            68.8      2.7     7065525  5.24e10
## 10 Azerbai~  1960             NA              61.3      5.57    3897889 NA
## # ... with 10,535 more rows, and 2 more variables: continent <fct>,
## #   region <fct>
```

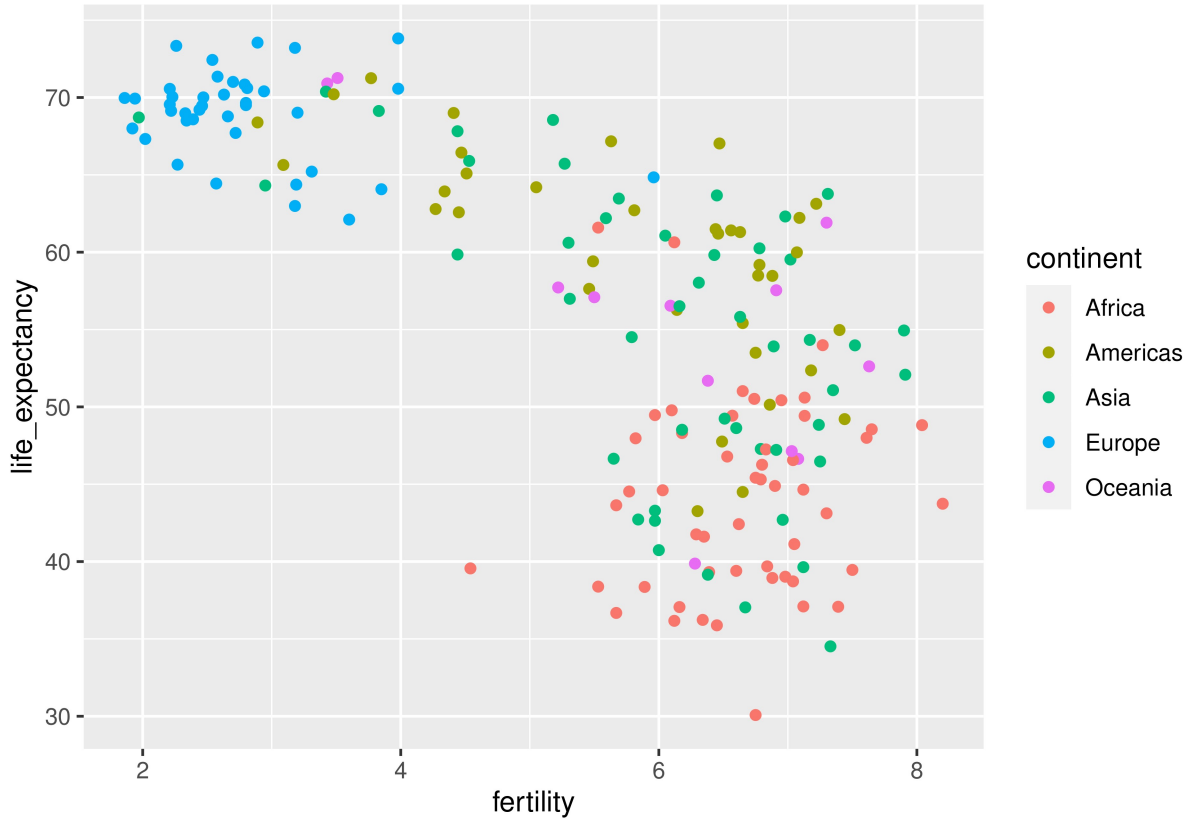
```
gapminder %>% filter(year==2015 & country %in% c("Sri Lanka","Turkey")) %>%
  select(country,infant_mortality)
```

```
##     country infant_mortality
```

```
## 1 Sri Lanka               8.4
## 2    Turkey              11.6
```

- example1 : color is setting the variable options

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



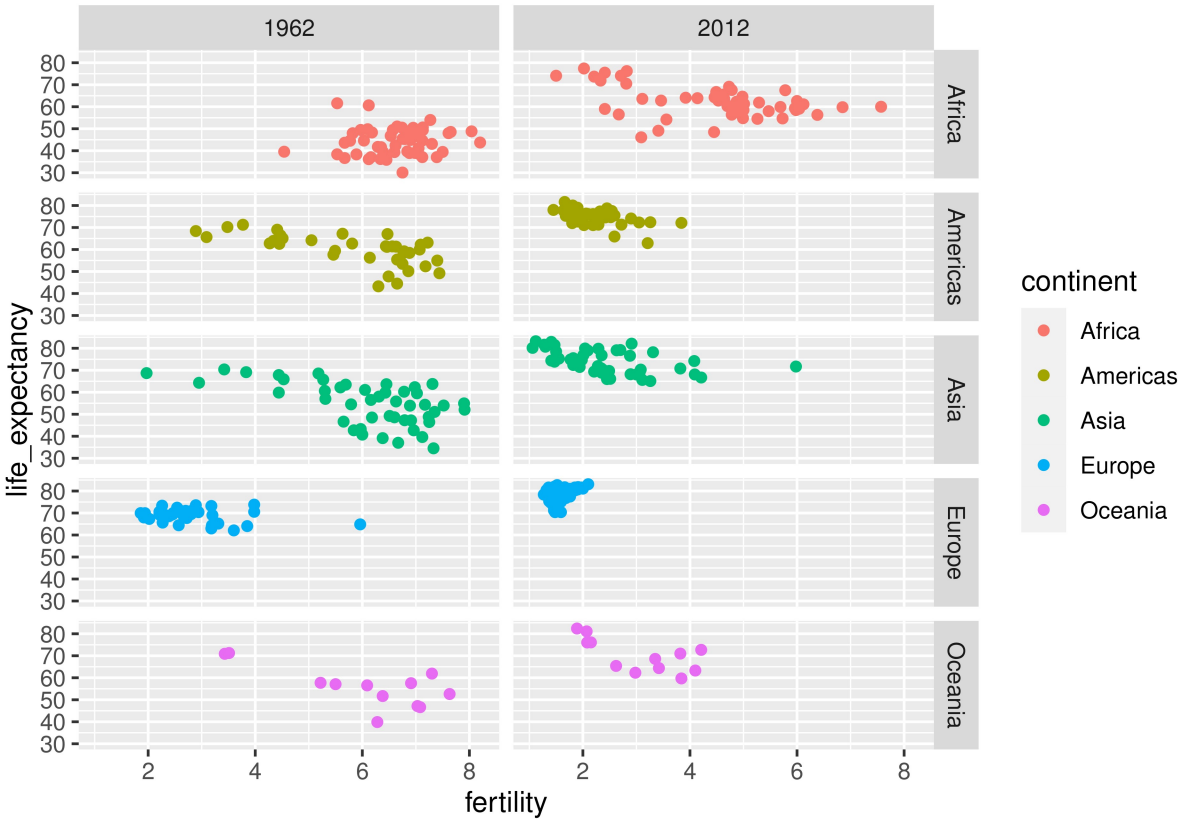
Chapter8.1 Faceting

Facet: **기준**. 어떤 기준으로 나누는 행위.

그렇다면 질문, **나누는 셋팅을 어떻게 설정함?**: variable mapping은 아님. 그러므로 **추가!**

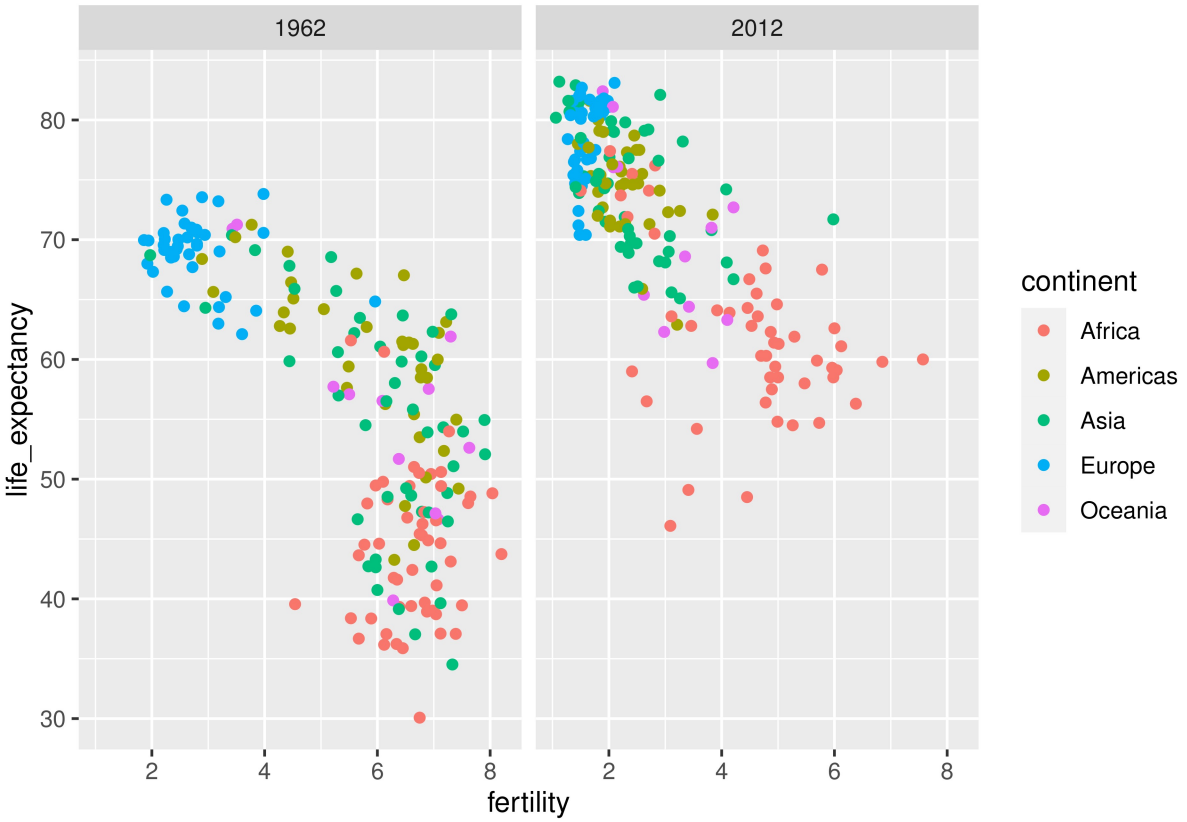
- example2-1: facet by continent & year

```
gapminder%>%filter(year %in% c(1962,2012))%>%
  ggplot(aes(fertility, life_expectancy, col=continent))+
  geom_point()+
  facet_grid(continent~year) # "X" ~ "Y"
```



* example2-2: facet by year

```
gapminder%>%filter(year %in% c(1962,2012))%>%
  ggplot(aes(fertility, life_expectancy, col=continent))+
  geom_point()+
  facet_grid(.~year) # dot operator means all the variables
```

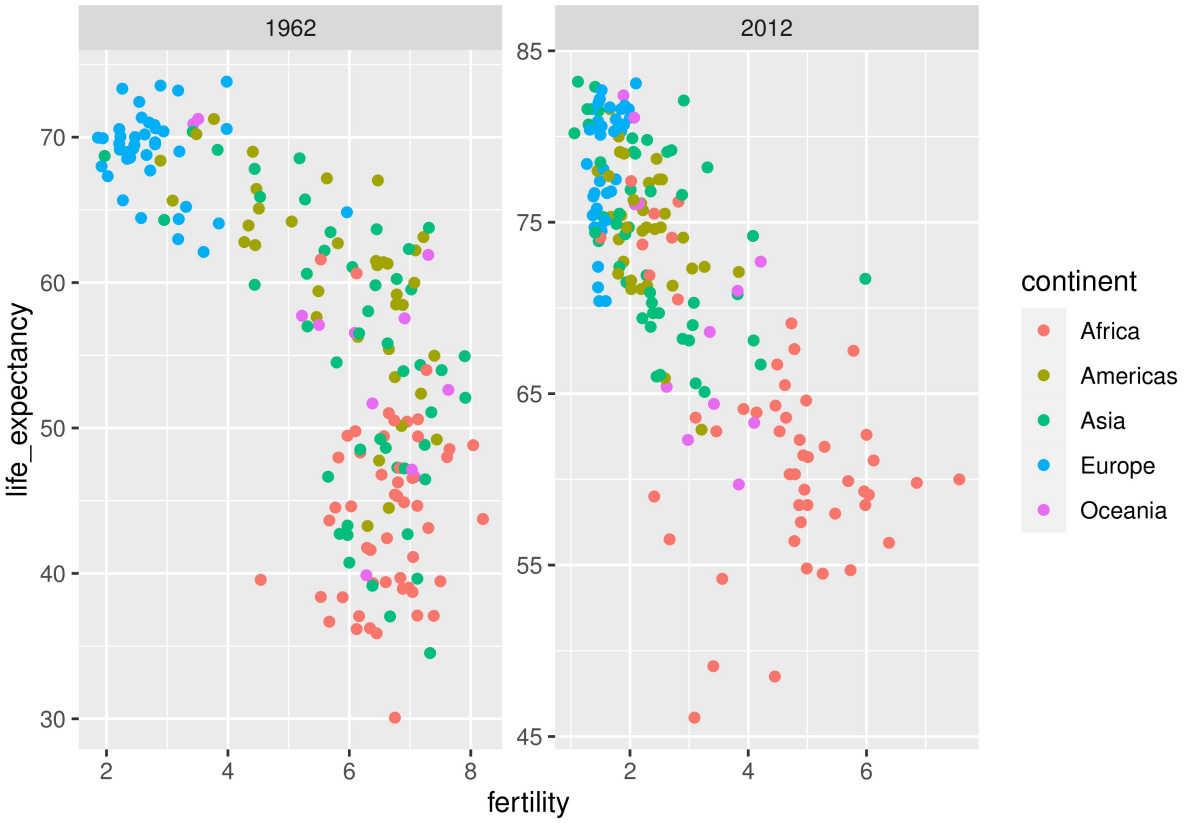


* example2-3: facet by VARIOUS years

```
years=c(1962,1980,1990,2000,2012)
continents=c("Europe","Asia")
gapminder%>%
  filter(year %in% years & continent %in% continents) %>%
  ggplot(aes(fertility,life_expectancy,col=continent))+
  geom_point()+
  facet_wrap(.~year)
```

* example2-4: scales but NOT RECOMMENDED

```
filter(gapminder,year%in%c(1962,2012))%>%
  ggplot(aes(fertility,life_expectancy,col=continent))+
  geom_point()+
  facet_wrap(.~year,scales="free")
```

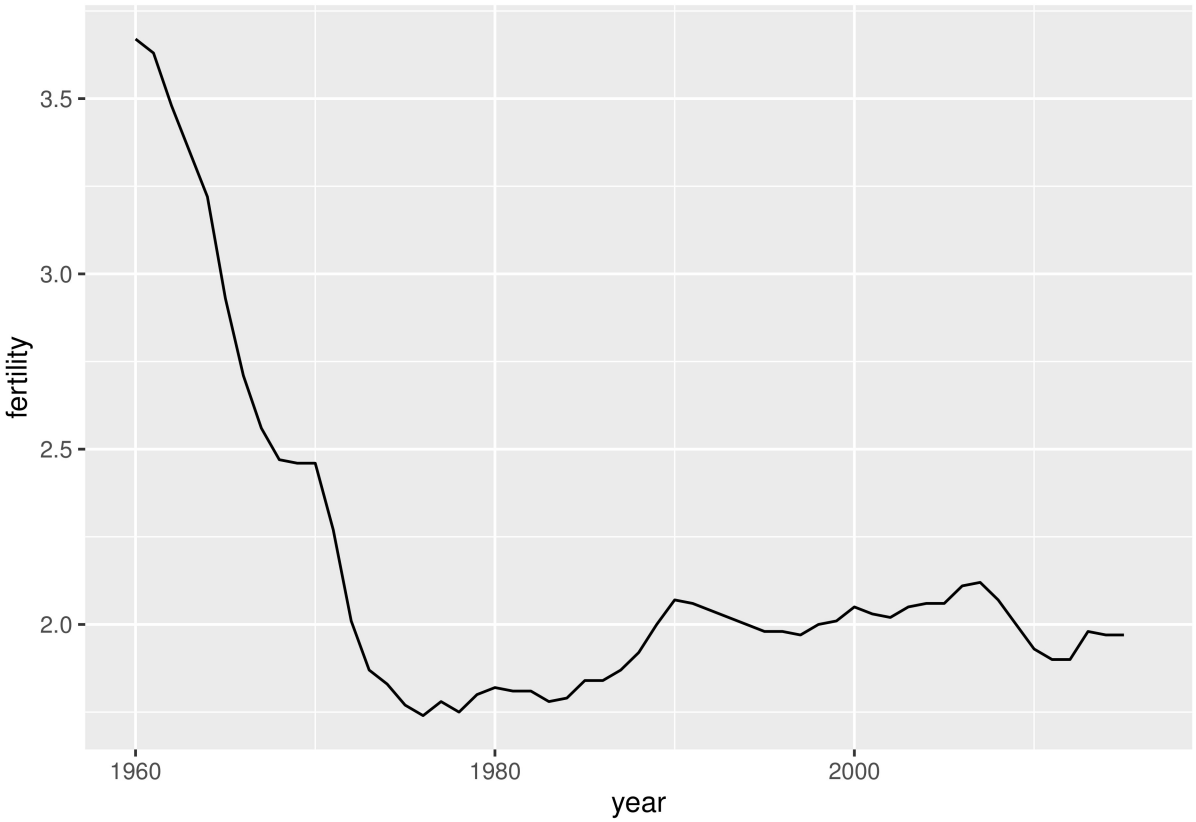


Chapter8.2 Time Series plots

- example3-1: basic time series plot

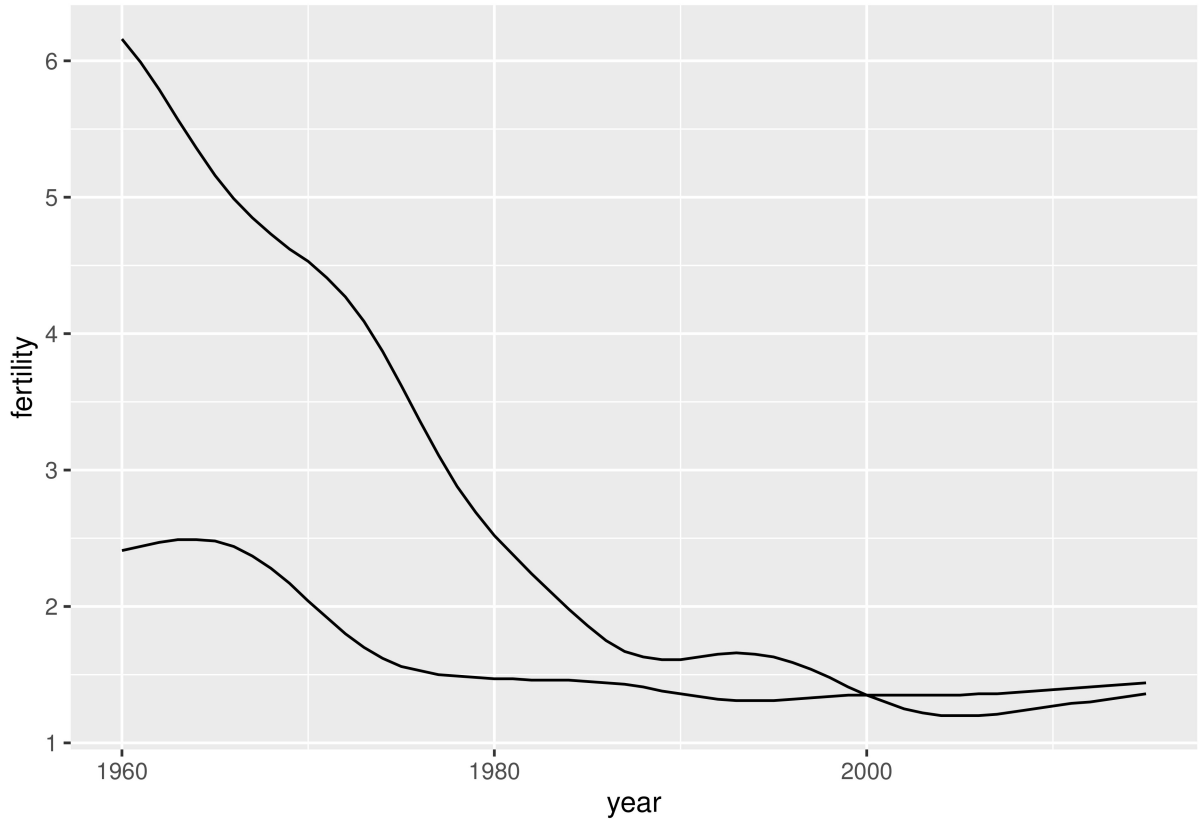
```
gapminder%>%filter(country=="United States")%>%
  ggplot(aes(year,fertility))+
  geom_line()
```

```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```



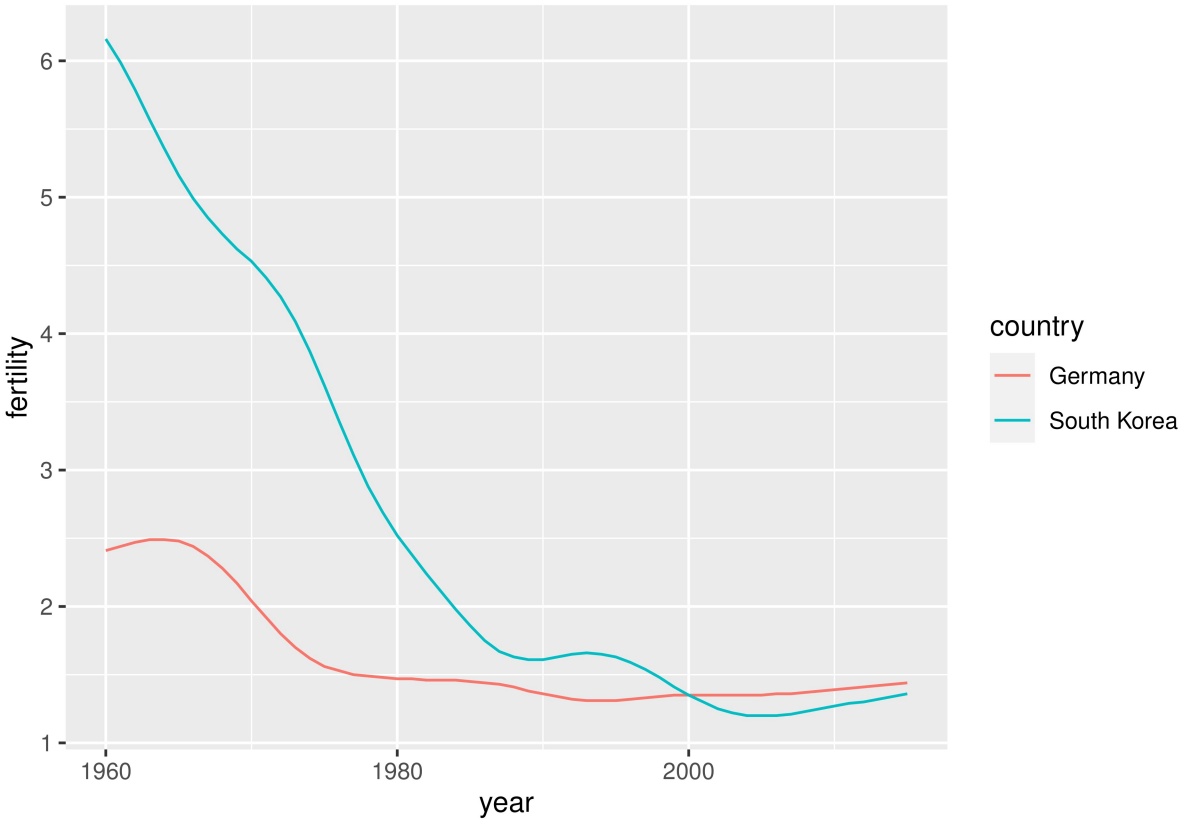
* example3-2: two objects time series plot each: **mapping variables**

```
countries <- c("South Korea","Germany")
gapminder %>% filter(country %in% countries & !is.na(fertility))%>%
  ggplot(aes(year,fertility,group=country)) +
  geom_line()
```




```
# group인자, col인자
```

```
countries <- c("South Korea","Germany")
gapminder %>% filter(country %in% countries & !is.na(fertility))%>%
  ggplot(aes(year,fertility,col=country))+
  geom_line()
```

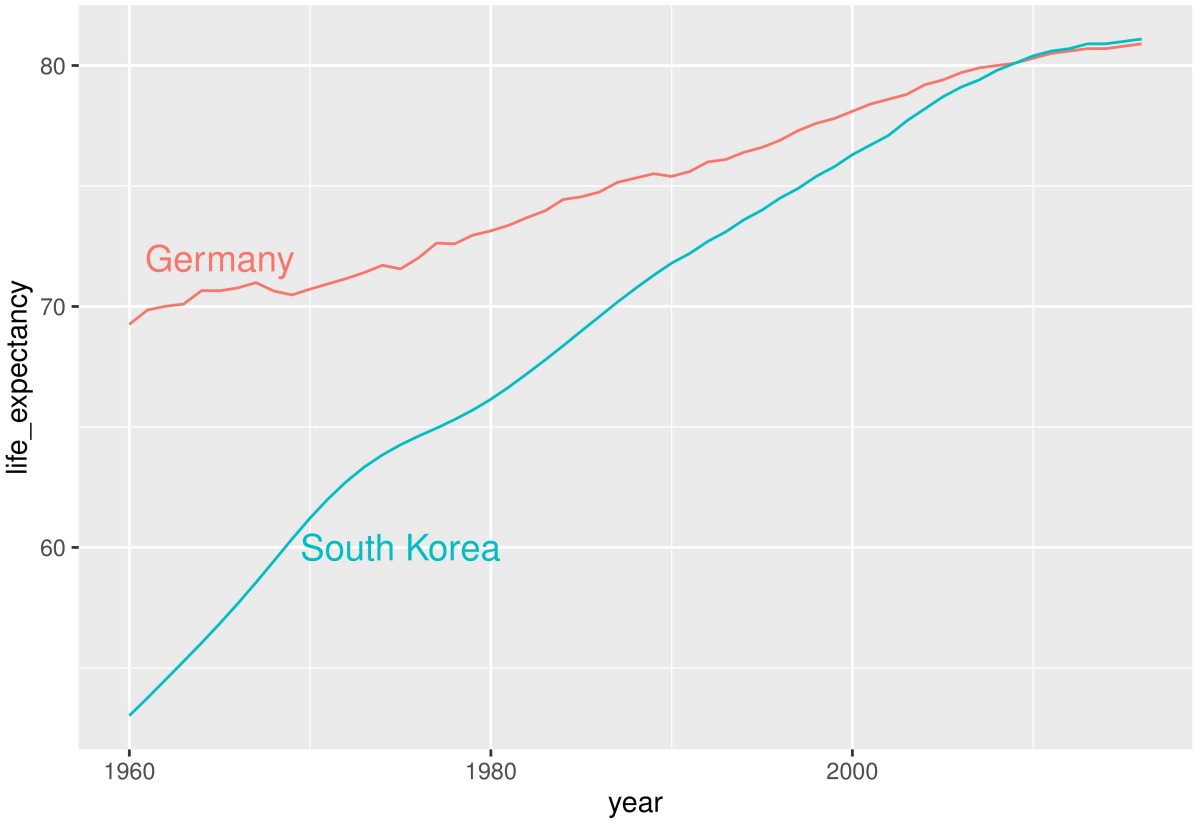


* example3-3: two objects putting labels by **LOCATION aes**

```
labels <-data.frame(country=countries,x=c(1975,1965),y=c(60,72))
labels
```

```
##       country    x  y
## 1 South Korea 1975 60
## 2     Germany 1965 72
```

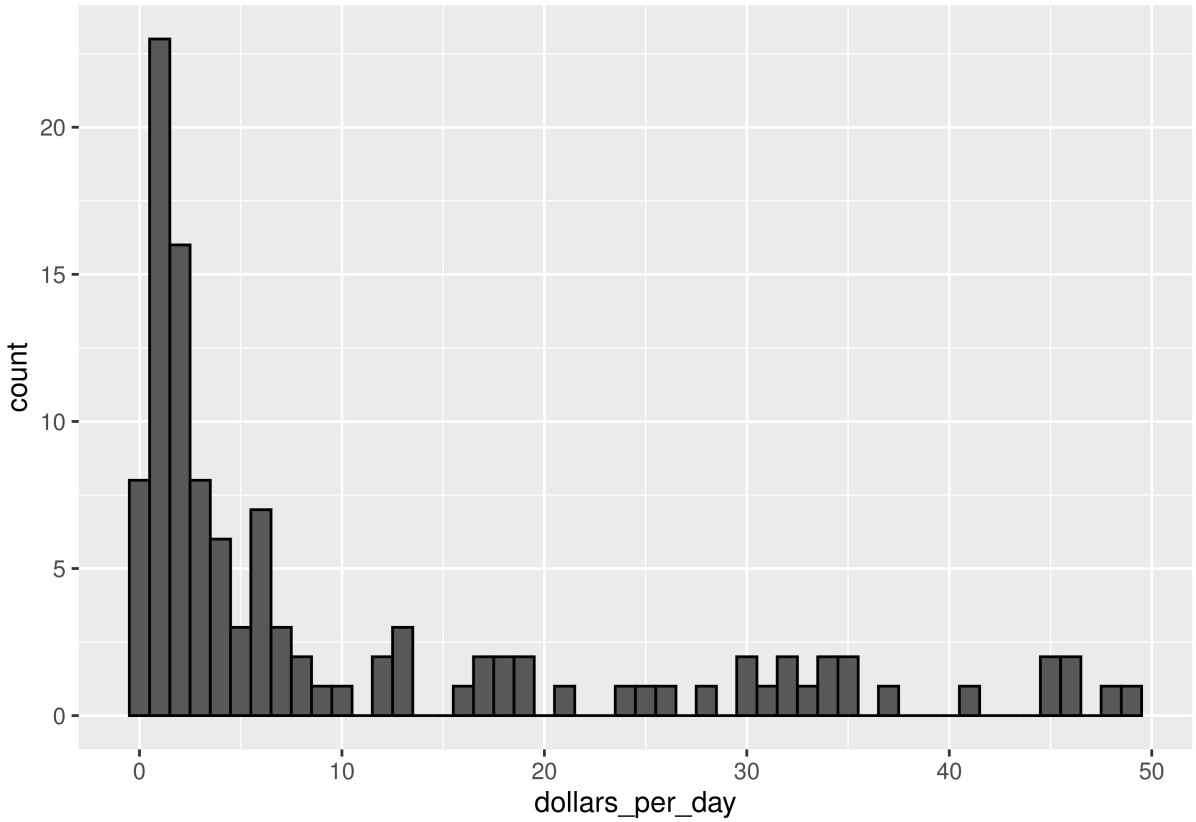
```
gapminder%>%
  filter(country %in% countries)%>%
  ggplot(aes(year,life_expectancy,col=country))+
  geom_line()+
  geom_text(data=labels,aes(x,y,label=country),size=5)+
  theme(legend.position="none")
```



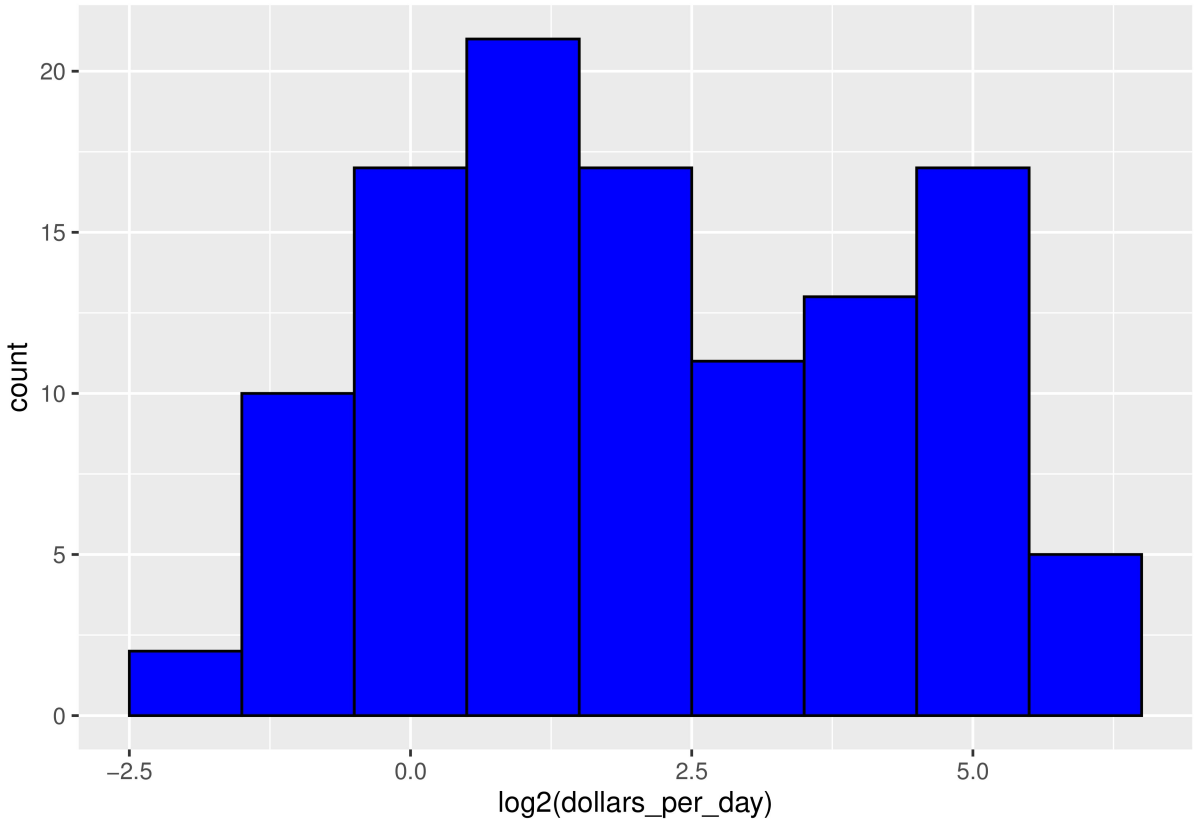
Chapter8.3 Scaling and Comparing Distributions

- example4: DATA TRANSFORMATIONS by scales **FOR INTERPRETATION**

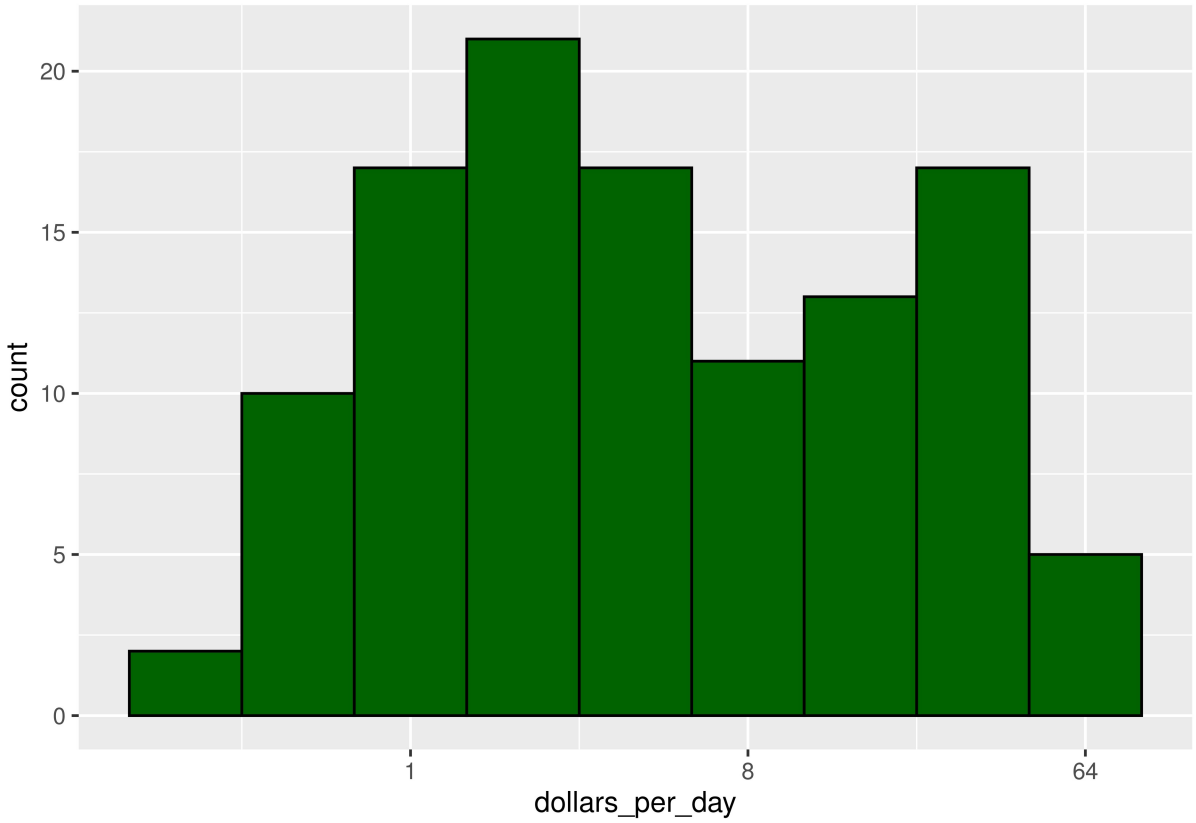
```
gapminder <- gapminder%>% mutate(dollars_per_day = gdp/population /365)
past_year <- 1970
gapminder %>%
  filter(year==past_year & !is.na(gdp))%>%
  ggplot(aes(dollars_per_day))+
  geom_histogram(binwidth=1,color="black")
```




```
# (1) Change the data
gapminder%>%
  filter(year==past_year & !is.na(gdp))%>%
  ggplot(aes(log2(dollars_per_day)))+
  geom_histogram(binwidth=1,fill="blue",color="black")
```



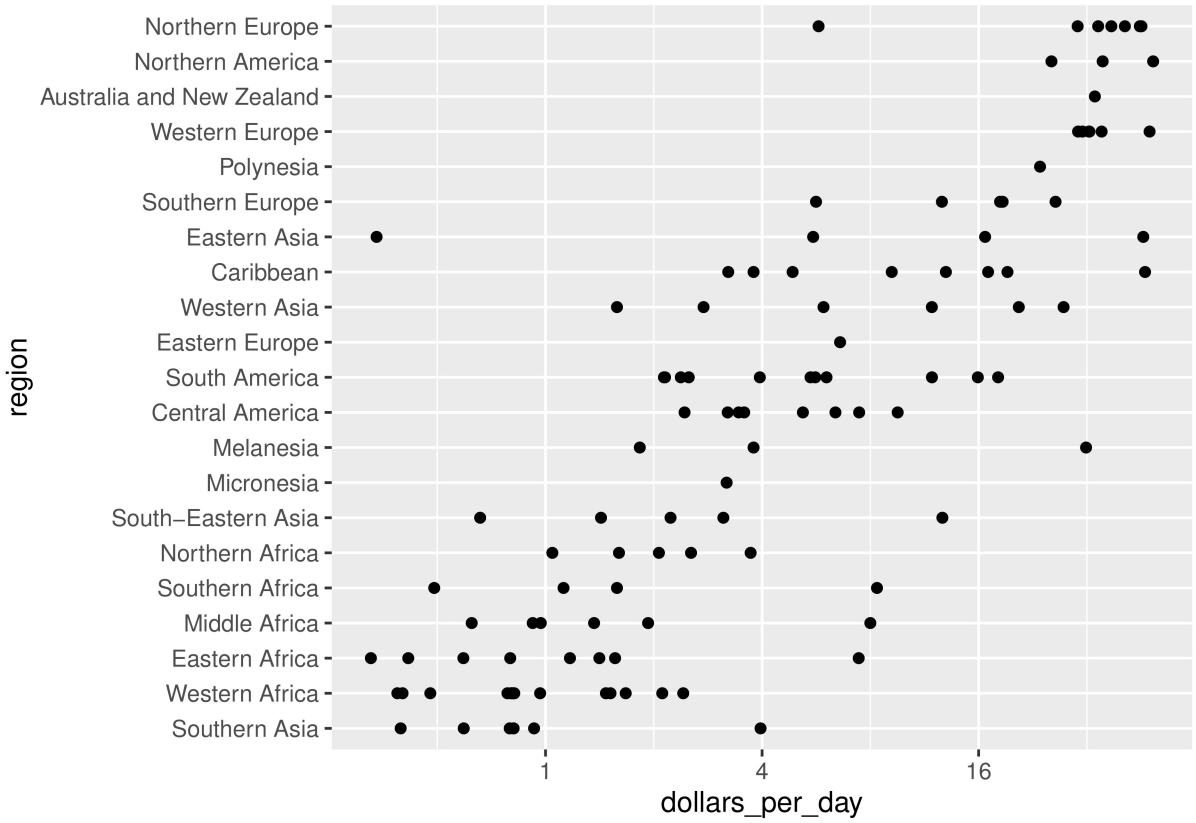

```
# 이때 x scale은 데이터 자체를 바꿨으므로 1,2,3,4,5 이렇게 나옴

# (2) Change the scale
gapminder%>%
  filter(year==past_year & !is.na(gdp))%>%
  ggplot(aes(dollars_per_day))+
  geom_histogram(binwidth=1,fill="darkgreen",color="black")+
  scale_x_continuous(trans="log2")
```




```
# 이때 x scale은 scale만 바꿨으므로 1,2,4,8,16 이렇게 나옴
```

- example5.1: Comparing multiple Distributions

```
# reorder(정렬할 데이터, 정렬기준)
gapminder%>%
  filter(year==past_year & !is.na(gdp)) %>%
  mutate(region=reorder(region,dollars_per_day,FUN=median))%>%
  ggplot(aes(dollars_per_day,region))+
  geom_point()+
  scale_x_continuous(trans="log2")
```




```
## if you have several values per level of your factor, you can specify
## which function to determine the order of your factor.
## factor is "region". And what determines order is region's dollars_per_day median.
gapminder%>%
  filter(year==past_year & !is.na(gdp)) %>%
  mutate(region=reorder(region,dollars_per_day))%>%
  select(region,dollars_per_day)%>%head(10)
```

```
##                       region dollars_per_day
## 1             Northern Africa       3.7172265
## 2               South America      18.1207496
## 3  Australia and New Zealand      33.6671656
## 4               Western Europe      30.2348264
## 5                    Caribbean      46.4254181
## 6                Southern Asia       0.7950843
## 7                    Caribbean      16.9973307
## 8               Western Europe      31.0956941
## 9             Central America       3.2095300
## 10              Western Africa       0.7837057
```

```
## reorder()함수는 그래프 표시에 사용되는 것이지, 일반 벡터 정렬에 사용되지 않는다.
```

- 중요! 순서정하는 함수들과 사용법

```r
# 순서를 정하는 기준, 사용법
a=c(10,40,30,20,50,60,70)
a
```

```
## [1] 10 40 30 20 50 60 70
```

```r
sort(a) # 원소의 결과값을 출력
```

```
## [1] 10 20 30 40 50 60 70
```

```r
order(a,decreasing=FALSE) # 원소의 순서를, 현재 원소별 위치로 설명
```

```
## [1] 1 4 3 2 5 6 7
```

```r
rank(a) # 원소의 순서를,
```

```
## [1] 1 4 3 2 5 6 7
```

```r
# arrange(a)
class(a)
```

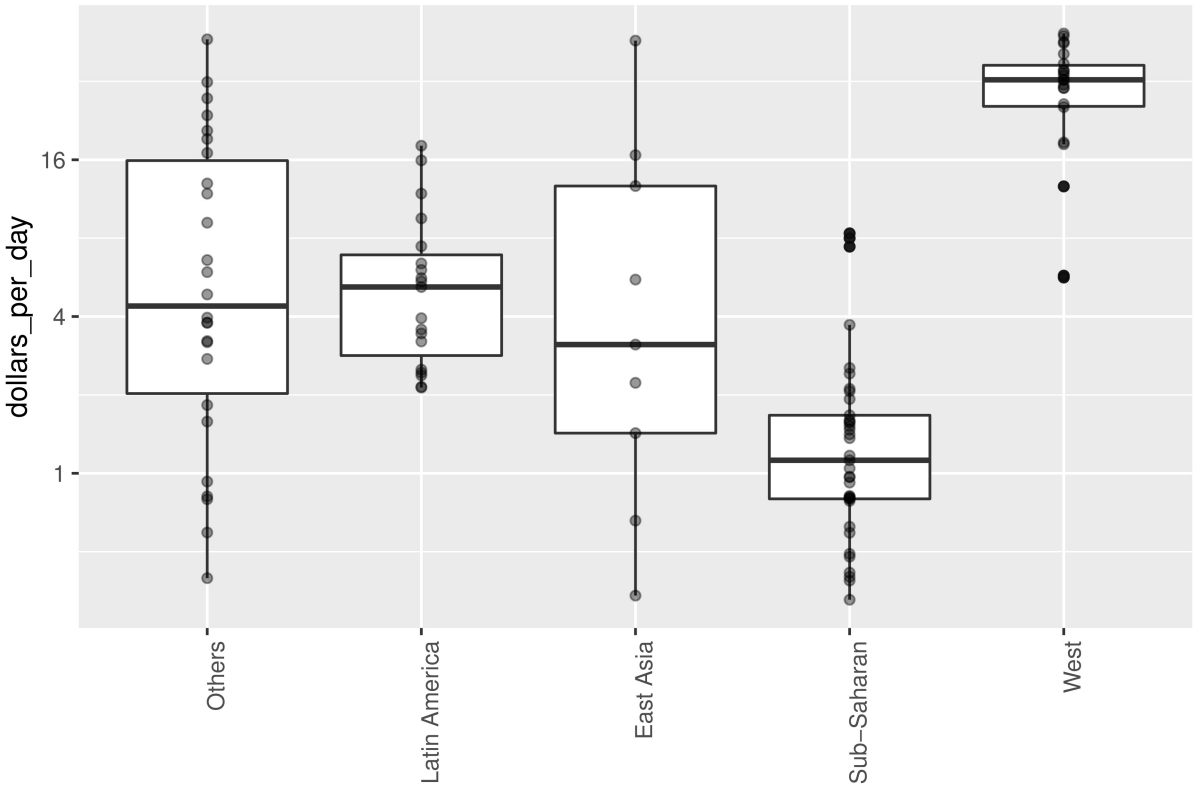
```
## [1] "numeric"
```

- example5.2: Comparing multiple Distributions by regions

```r
gapminder <- gapminder %>%
  mutate(group=case_when(
    region %in% c("Western Europe", "Northern Europe","Southern Europe",
"Northern America","Australia and New Zealand")  ~ "West",
    region %in% c("Eastern Asia","South-Eastern Asia") ~ "East Asia",
    region %in% c("Carribean","Central America","South America") ~ "Latin America",
    continent == "Africa" &
      region != "North Africa" ~ "Sub-Saharan",
    TRUE ~ "Others"))
gapminder <-gapminder%>%
  mutate(group=factor(group,
                      levels=c("Others", "Latin America",
"East Asia", "Sub-Saharan", "West")))
# factor함수: 범주형 자료일때, 원소들을 level로 만들어줌(class: factor)
```

- example5.3: Boxplot

```r
p <- gapminder %>%
  filter(year == past_year & !is.na(gdp))%>%
  ggplot(aes(group,dollars_per_day)) +
  geom_boxplot()+
  scale_y_continuous(trans="log2")+
  xlab("") +
  theme(axis.text.x = element_text(angle=90,hjust=1)) # rotate x 90, h adjust ->
p+geom_point(alpha=0.4)
```



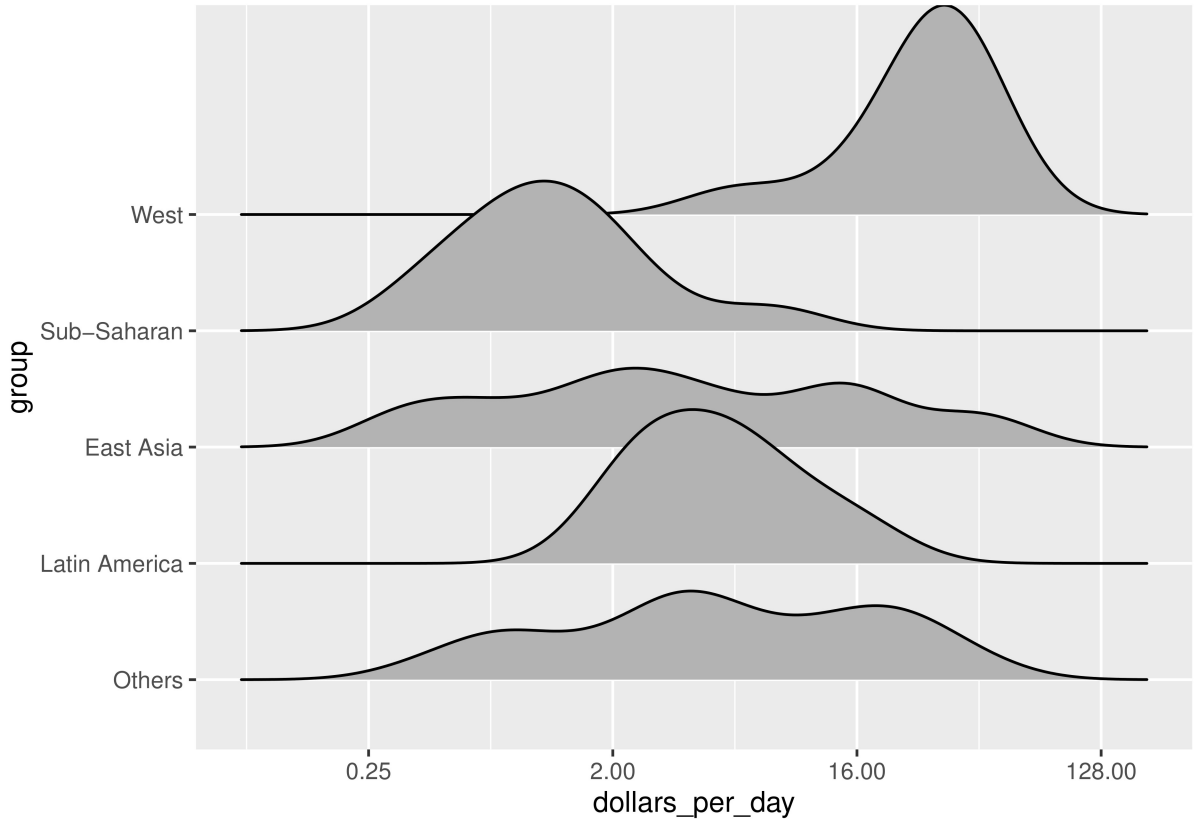
Chapter8.4 Ridge Plots

Boxplot의 요약으로 인한 Distribution 설명력 손실을 보완해줌 by Density

- example6: ridge plot of one year

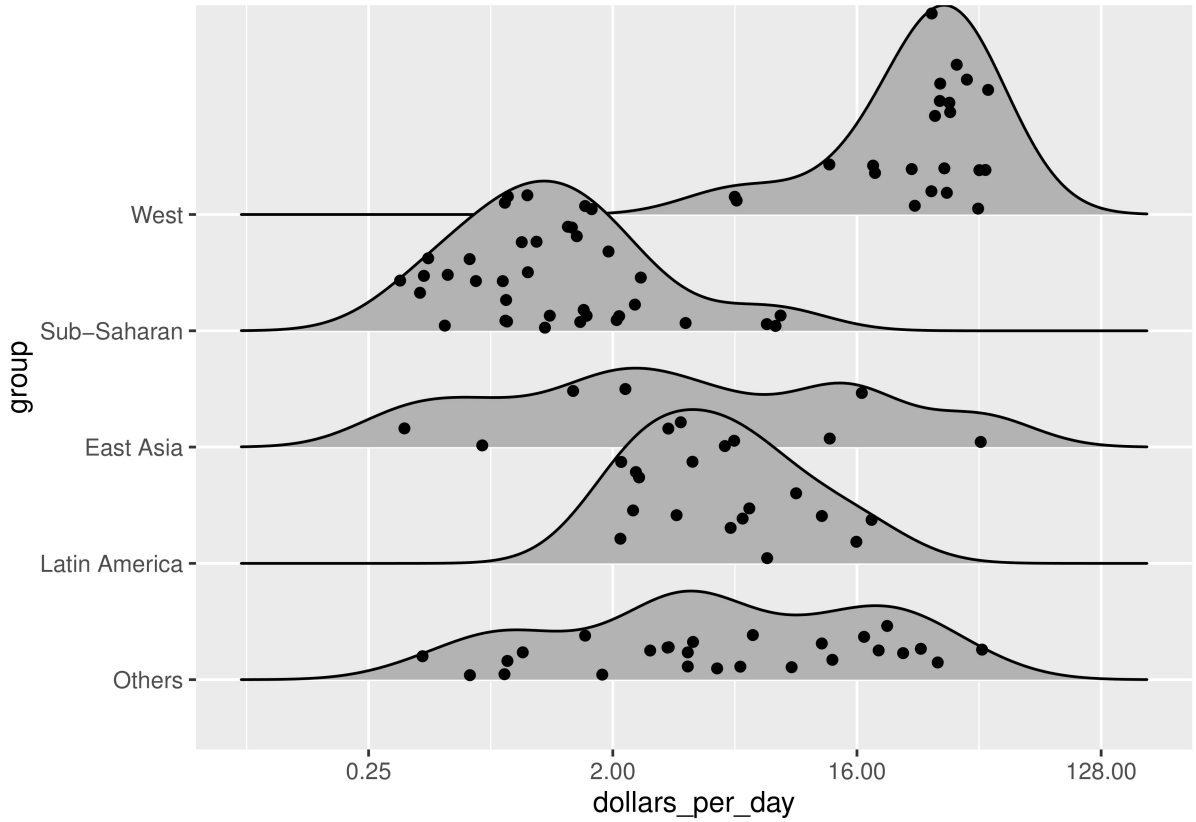
```
library(ggridges)
p <- gapminder%>%
  filter(year==past_year & !is.na(dollars_per_day))%>%
  ggplot(aes(dollars_per_day,group))+
  scale_x_continuous(trans="log2")
p+geom_density_ridges()
```

```
## Picking joint bandwidth of 0.65
```



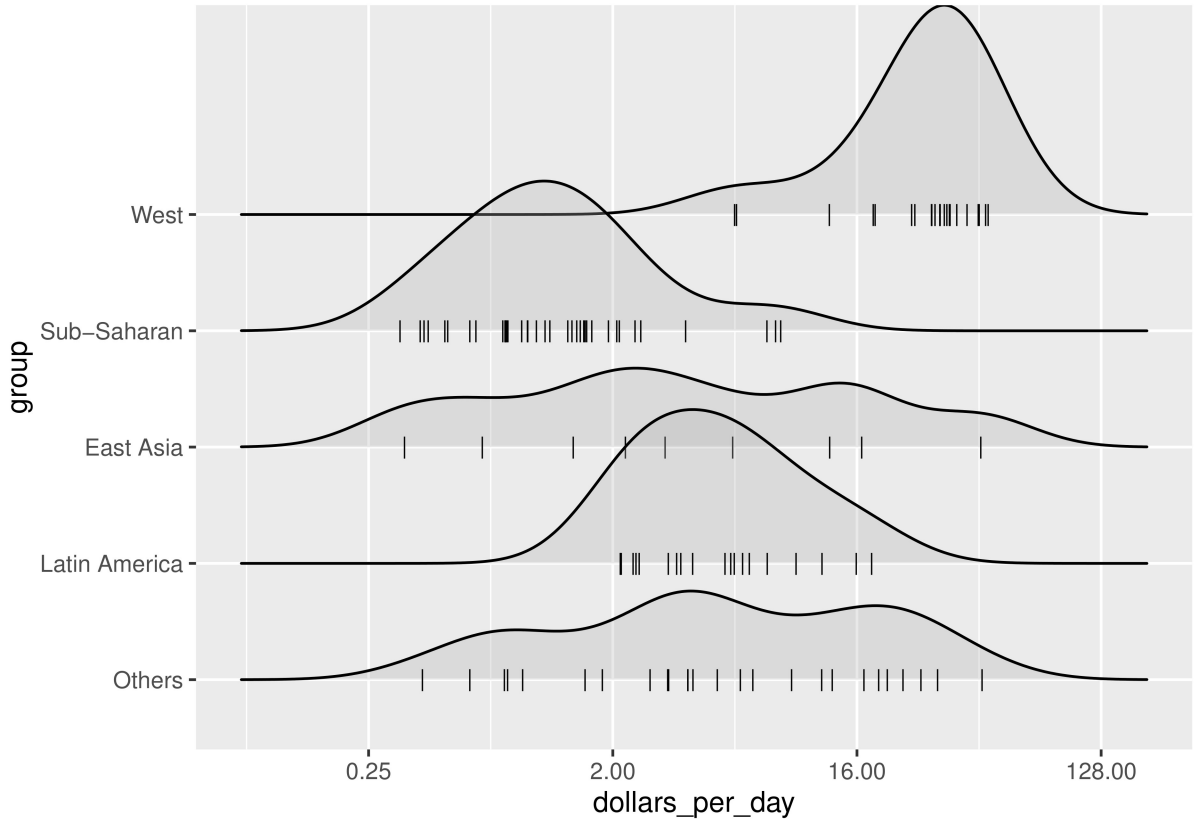

```
p+geom_density_ridges(jittered_points=TRUE) # jitter point는 random이라 신경nono
```

```
## Picking joint bandwidth of 0.65
```



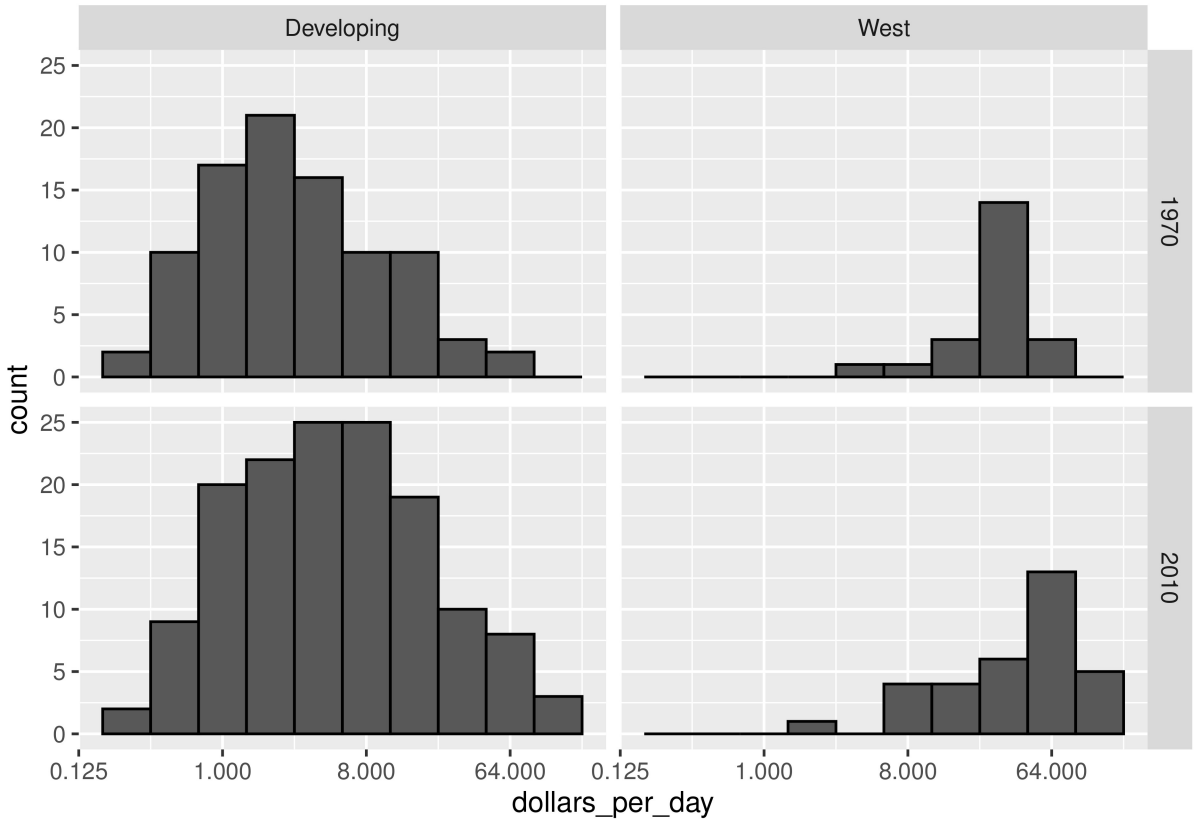

```
p+geom_density_ridges(jittered_points=TRUE,
                      position=position_points_jitter(height=0),
                      point_shape="|",point_size=3,
                      point_alpha=1,alpha=0.3)
```

```
## Picking joint bandwidth of 0.65
```



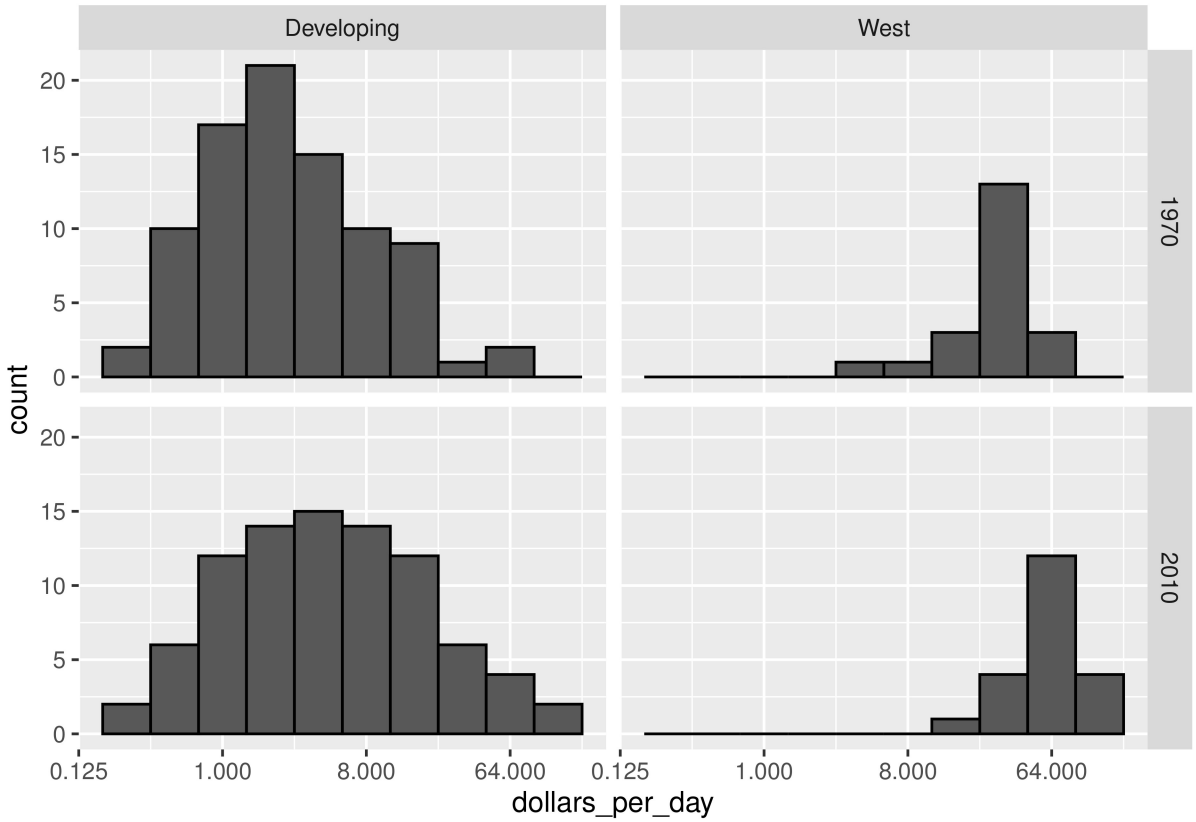
* example6-2: ridge plot of two years

```
past_year <- 1970
present_year <- 2010
years <- c(past_year,present_year)
## ifelse 사용해서 case_when처럼 만듬
## mutate=(westyn=case_when(group %in% "West" ~ "West", TRUE ~ "Developing"))
gapminder%>%
  filter(year %in% years & !is.na(gdp))%>%
  mutate(west=ifelse(group=="West","West","Developing")) %>%
  ggplot(aes(dollars_per_day))+
  geom_histogram(binwidth=1,color="black")+
  scale_x_continuous(trans="log2")+
  facet_grid(year~west)
```



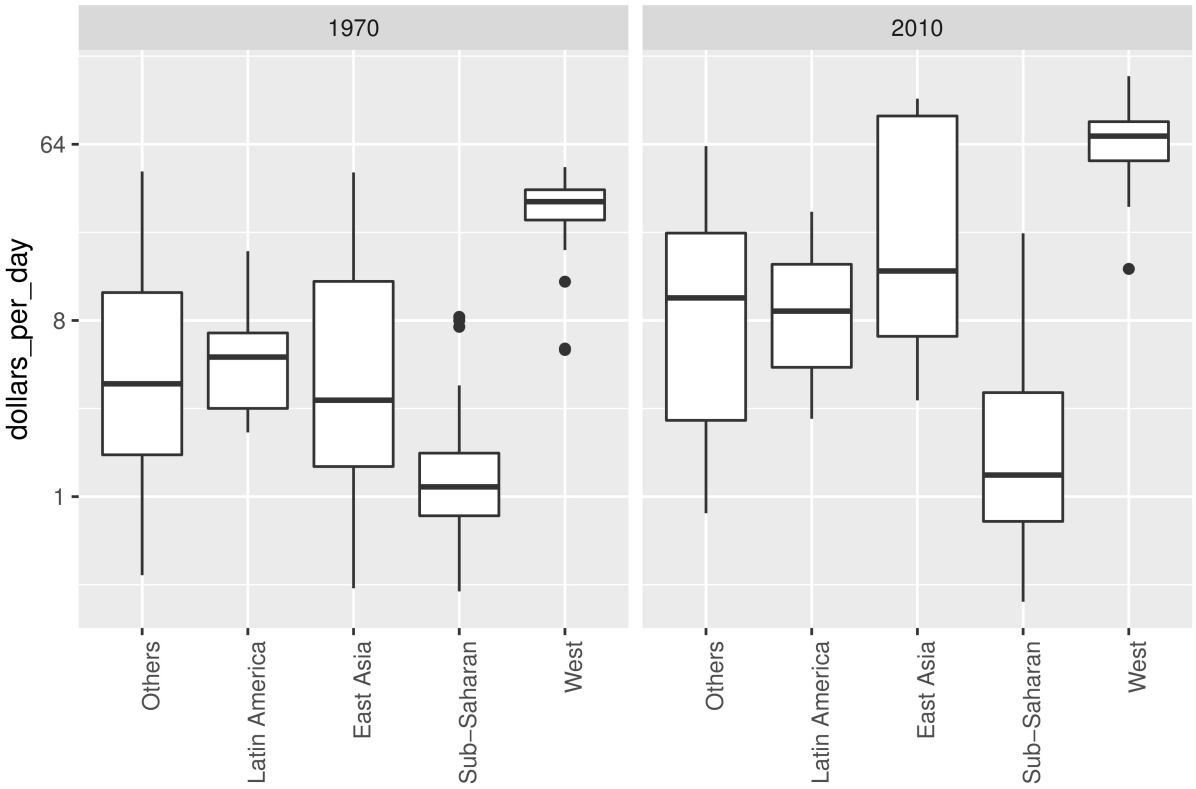
* example6-3: ridge plot of two years of only intersecting countries(accurate)

```
country_list1 <- gapminder %>%
  filter(year == past_year & !is.na(dollars_per_day))%>%
  pull(country)

country_list2 <- gapminder %>%
  filter(year == present_year & !is.na(dollars_per_day))%>%
  pull(country)

country_list = intersect(country_list1, country_list2) # 교집합 추출intersect

gapminder%>%
  filter(year %in% years & !is.na(gdp) & country %in% country_list)%>%
  mutate(west=ifelse(group=="West","West","Developing")) %>%
  ggplot(aes(dollars_per_day))+
  geom_histogram(binwidth=1,color="black")+
  scale_x_continuous(trans="log2")+
  facet_grid(year~west)
```



* example7.1 which groups improved the most between two years?

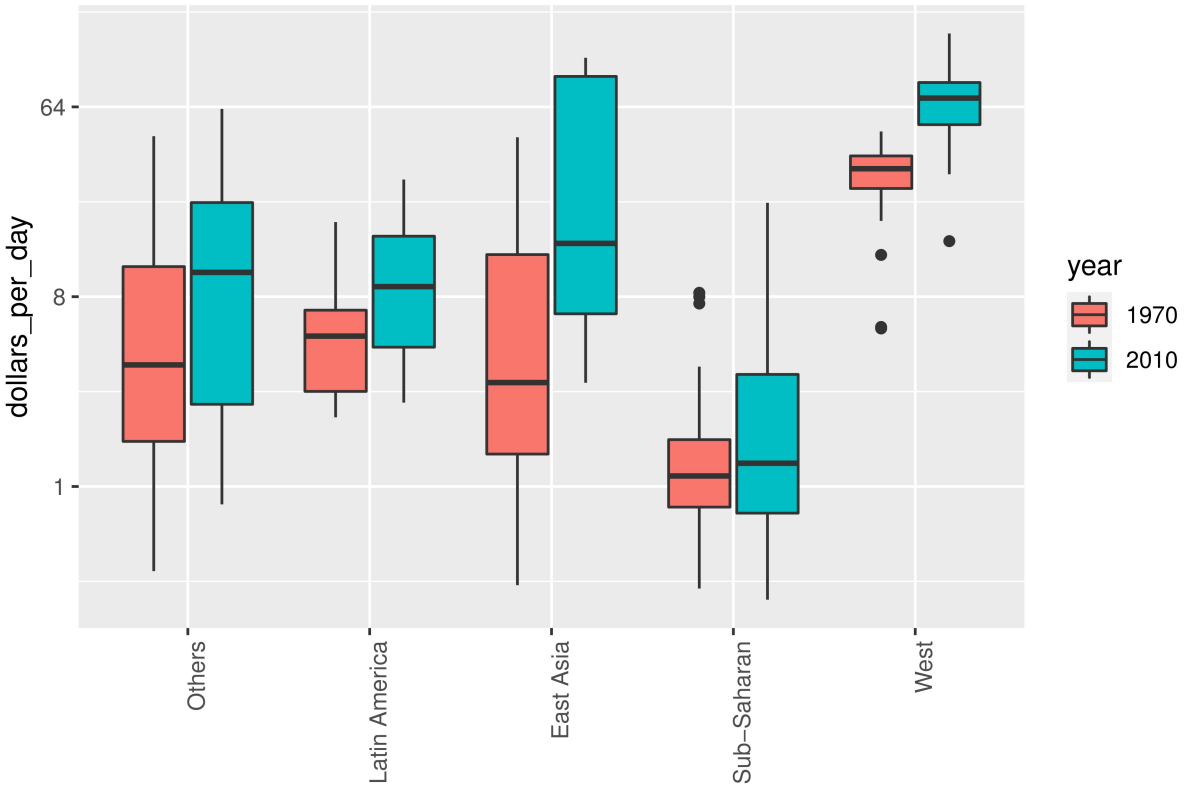
```
gapminder%>%
  filter(year %in% years & country %in% country_list) %>%
  ggplot(aes(group,dollars_per_day))+
  geom_boxplot() +
  theme(axis.text.x = element_text(angle=90,hjust=1))+
  scale_y_continuous(trans="log2")+
  xlab("") + facet_grid(.~year)
```



* example7.2 extension of 7.1 by `aes(fill=year)`

`fill=YEAR` should be the FACTOR VARIABLE

```
# fill = YEAR should be FACTOR YEAR
gapminder%>%
  filter(year %in% years & country %in% country_list) %>%
  mutate(year = factor(year))%>%
  ggplot(aes(group,dollars_per_day,fill=year))+
  geom_boxplot()+
  theme(axis.text.x = element_text(angle=90,hjust=1))+
  scale_y_continuous(trans="log2") + xlab("")
```


dollars_per_day
64
8
1
Others
Latin America
East Asia
Sub−Saharan
West
year
1970
2010