

data_vis_practice

juho

2021 11 4

```
data("gapminder")
```

```
gapminder=gapminder %>% as_tibble()  
str(gapminder)
```

```
## tibble [10,545 x 9] (S3: tbl_df/tbl/data.frame)  
## $ country      : Factor w/ 185 levels "Albania","Algeria",...: 1 2 3 4 5 6 7 8 9 10 ...  
## $ year         : int [1:10545] 1960 1960 1960 1960 1960 1960 1960 1960 1960 1960 ...  
## $ infant_mortality: num [1:10545] 115.4 148.2 208 NA 59.9 ...  
## $ life_expectancy: num [1:10545] 62.9 47.5 36 63 65.4 ...  
## $ fertility     : num [1:10545] 6.19 7.65 7.32 4.43 3.11 4.55 4.82 3.45 2.7 5.57 ...  
## $ population    : num [1:10545] 1636054 11124892 5270844 54681 20619075 ...  
## $ gdp           : num [1:10545] NA 1.38e+10 NA NA 1.08e+11 ...  
## $ continent     : Factor w/ 5 levels "Africa","Americas",...: 4 1 1 2 2 3 2 5 4 3 ...  
## $ region        : Factor w/ 22 levels "Australia and New Zealand",...: 19 11 10 2 15 21 2 1 22 21
```

```
summary(gapminder)
```

```
##           country      year infant_mortality life_expectancy  
## Albania      : 57   Min.   :1960   Min.    : 1.50   Min.    :13.20  
## Algeria      : 57   1st Qu.:1974   1st Qu.: 16.00   1st Qu.:57.50  
## Angola       : 57   Median :1988   Median : 41.50   Median :67.54  
## Antigua and Barbuda: 57   Mean    :1988   Mean    : 55.31   Mean    :64.81  
## Argentina    : 57   3rd Qu.:2002   3rd Qu.: 85.10   3rd Qu.:73.00  
## Armenia      : 57   Max.    :2016   Max.    :276.90   Max.    :83.90  
## (Other)      :10203      NA's    :1453  
## fertility    population      gdp      continent  
## Min.    :0.840   Min.    :3.124e+04   Min.    :4.040e+07   Africa :2907  
## 1st Qu.:2.200   1st Qu.:1.333e+06   1st Qu.:1.846e+09   Americas:2052  
## Median :3.750   Median :5.009e+06   Median :7.794e+09   Asia    :2679  
## Mean    :4.084   Mean    :2.701e+07   Mean    :1.480e+11   Europe  :2223  
## 3rd Qu.:6.000   3rd Qu.:1.523e+07   3rd Qu.:5.540e+10   Oceania : 684  
## Max.    :9.220   Max.    :1.376e+09   Max.    :1.174e+13  
## NA's    :187     NA's    :185      NA's    :2972  
##           region  
## Western Asia :1026  
## Eastern Africa : 912  
## Western Africa : 912  
## Caribbean     : 741  
## South America : 684  
## Southern Europe: 684  
## (Other)       :5586
```

Han Rosling's quiz(1)

1

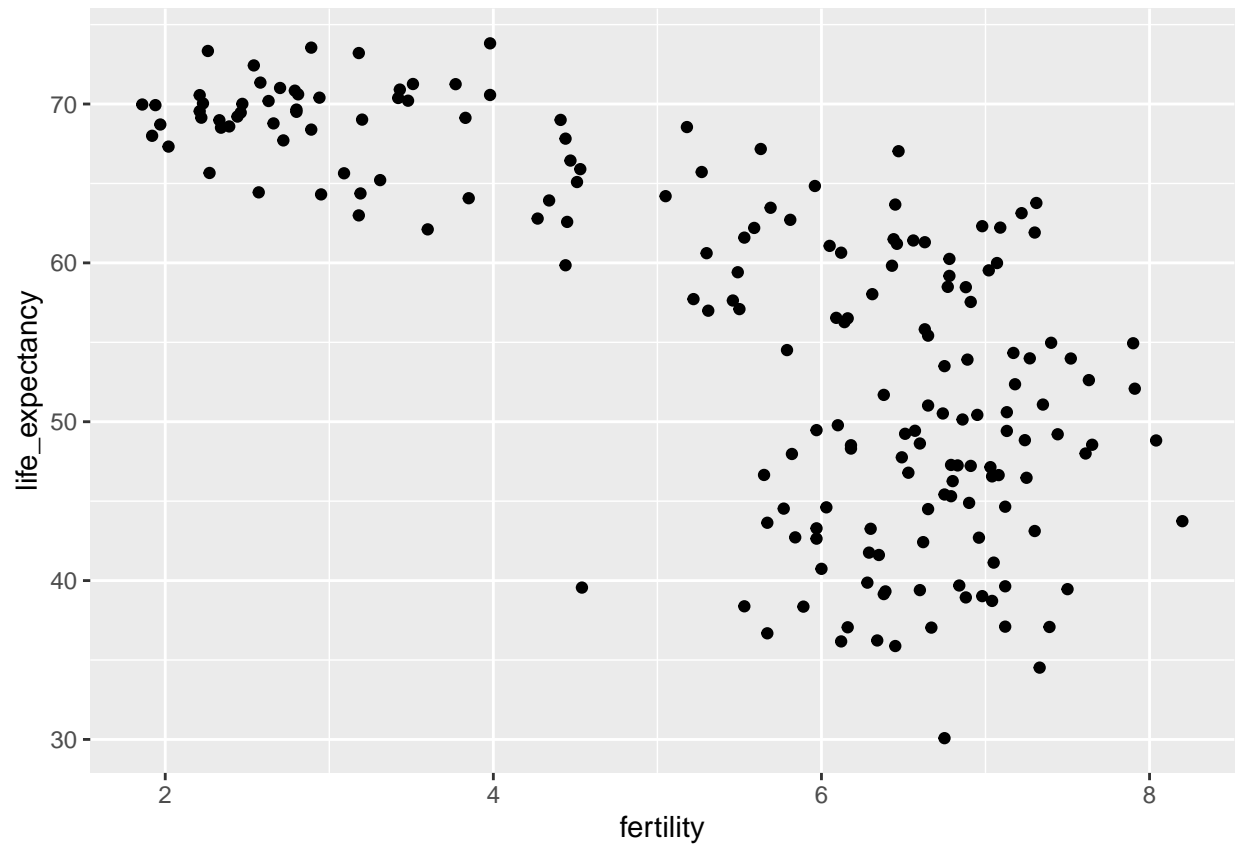
```
a=gapminder %>%  
  filter(year==2015) %>%  
  filter(country %in% c("Sri Lanka","Turkey")) %>%  
  select(country,infant_mortality)  
a
```

```
## # A tibble: 2 x 2  
##   country infant_mortality  
##   <fct>      <dbl>  
## 1 Sri Lanka      8.4  
## 2 Turkey        11.6
```

2

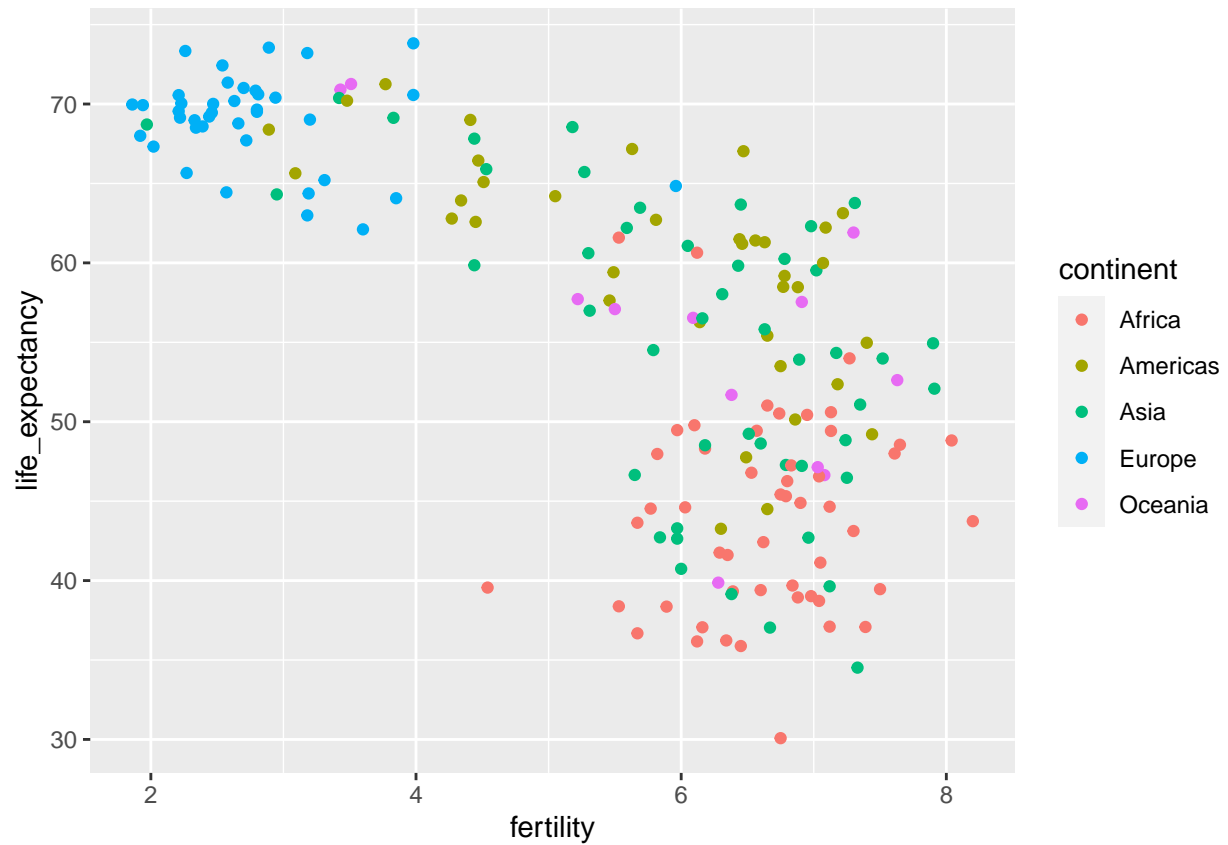
()

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



1. 5 . 2. . 3. .(or)

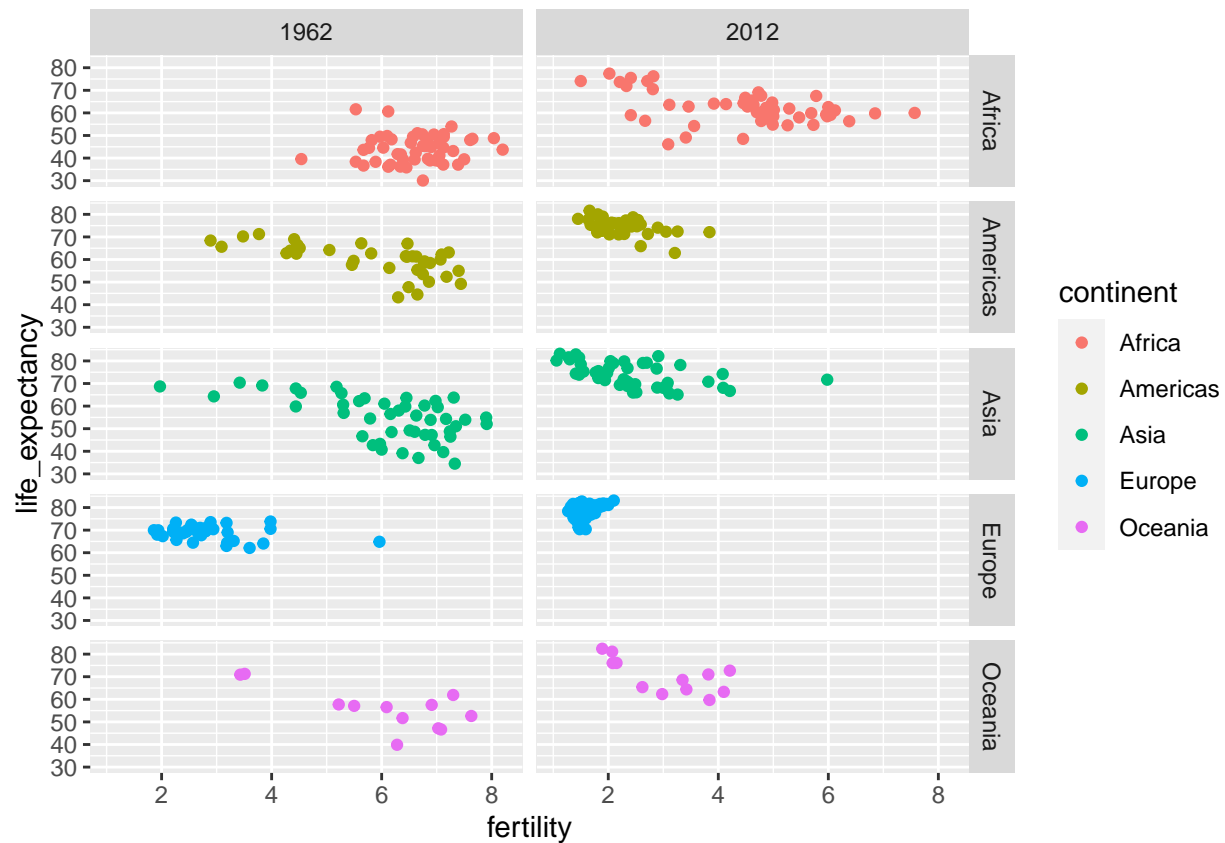
```
gapminder %>%
  filter(year == 1962) %>%
  ggplot(aes(fertility, life_expectancy, color=continent)) +
  geom_point()
```



facet_grid

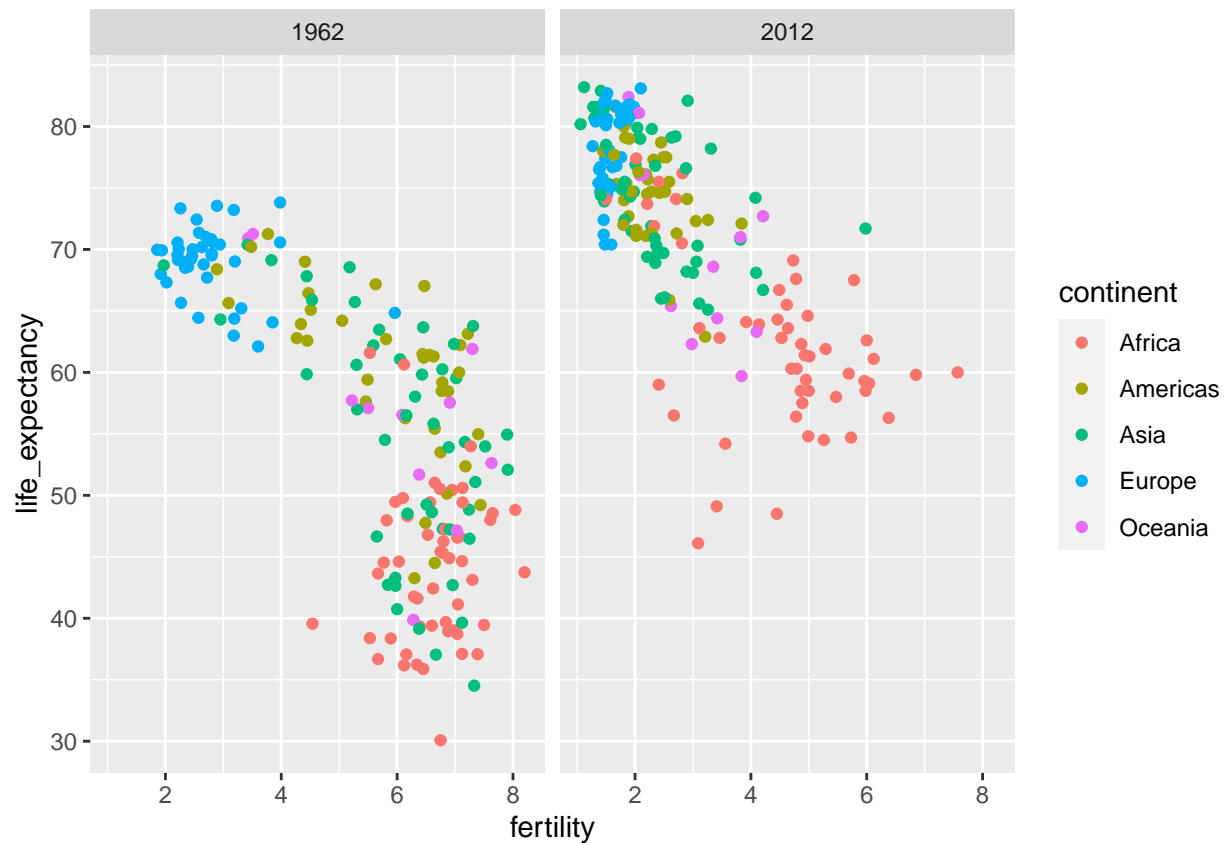
option ##### faceting 2012 1962

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



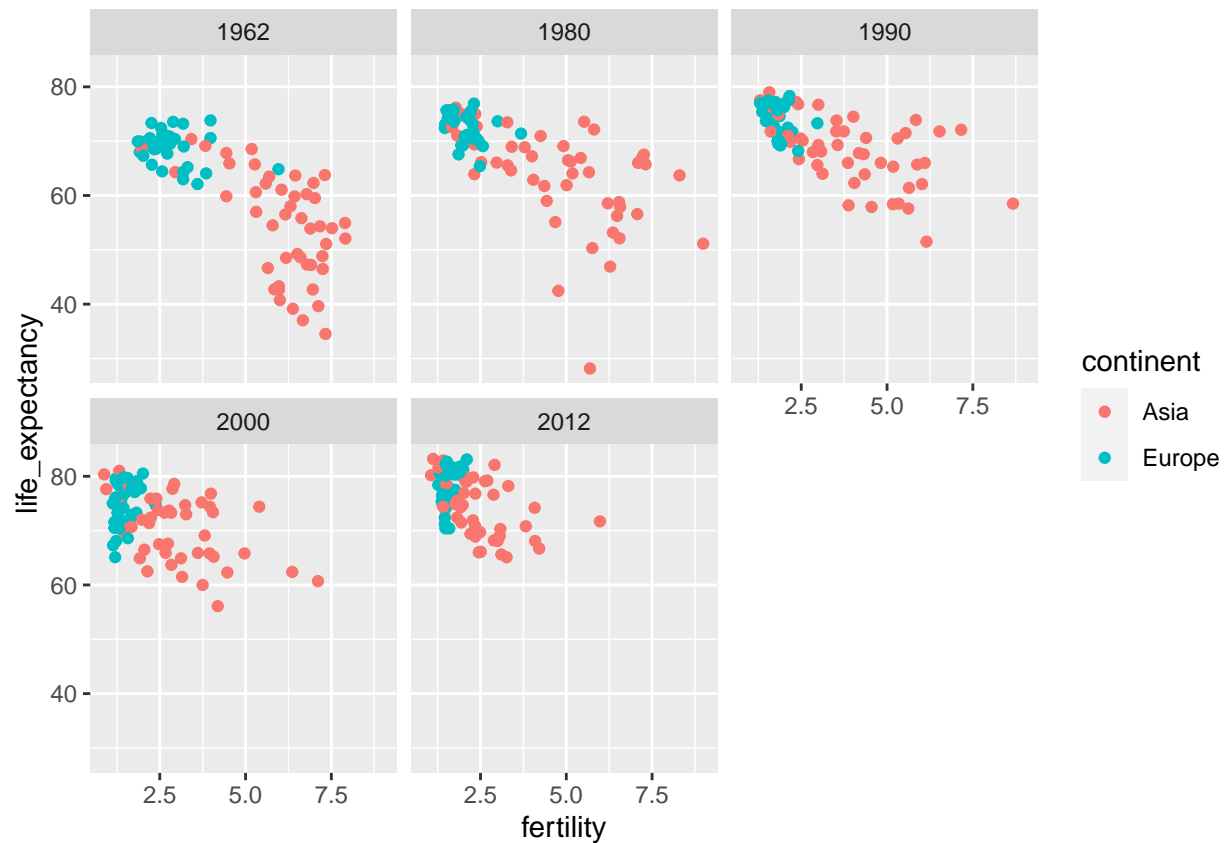
=> . ### continent x facet

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

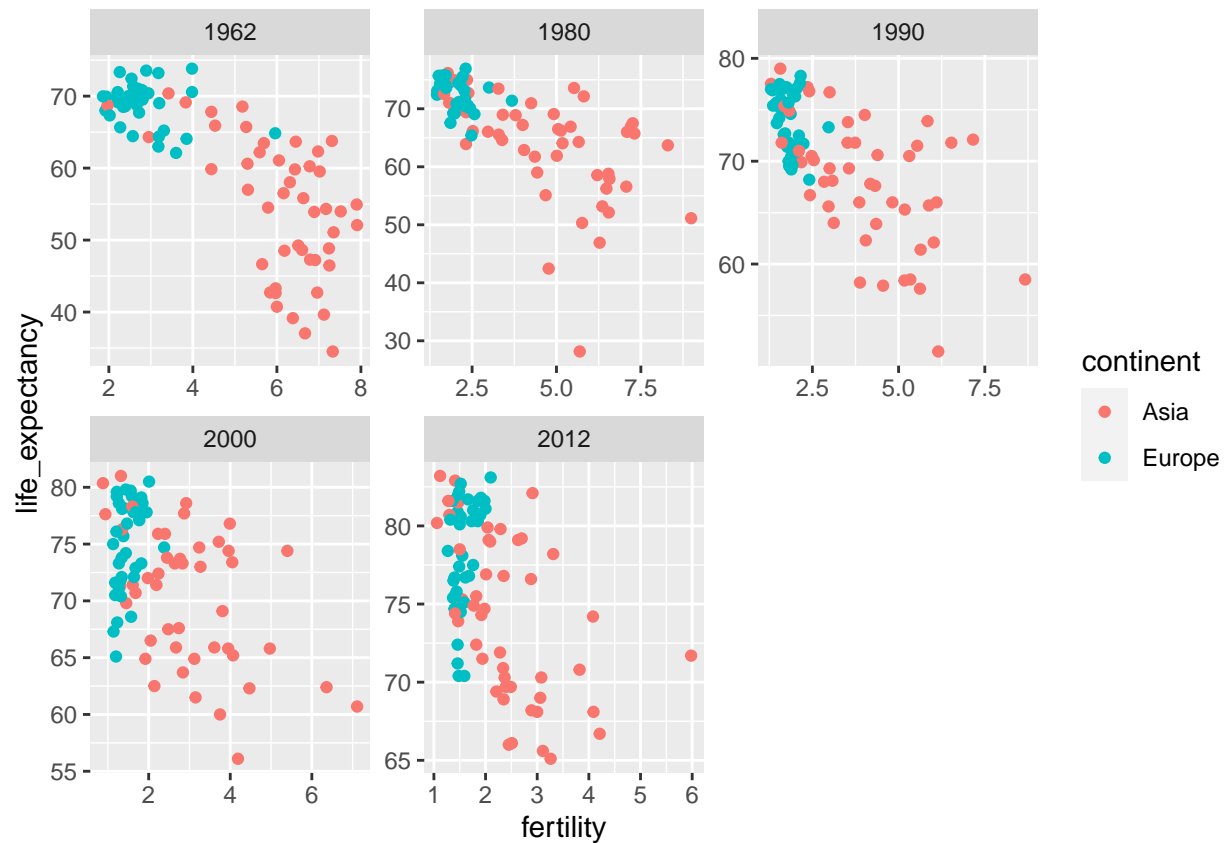


2012 . ### facet_wrap facet_grid scales="free" x,y
 x,y free_x or free_y => but data point

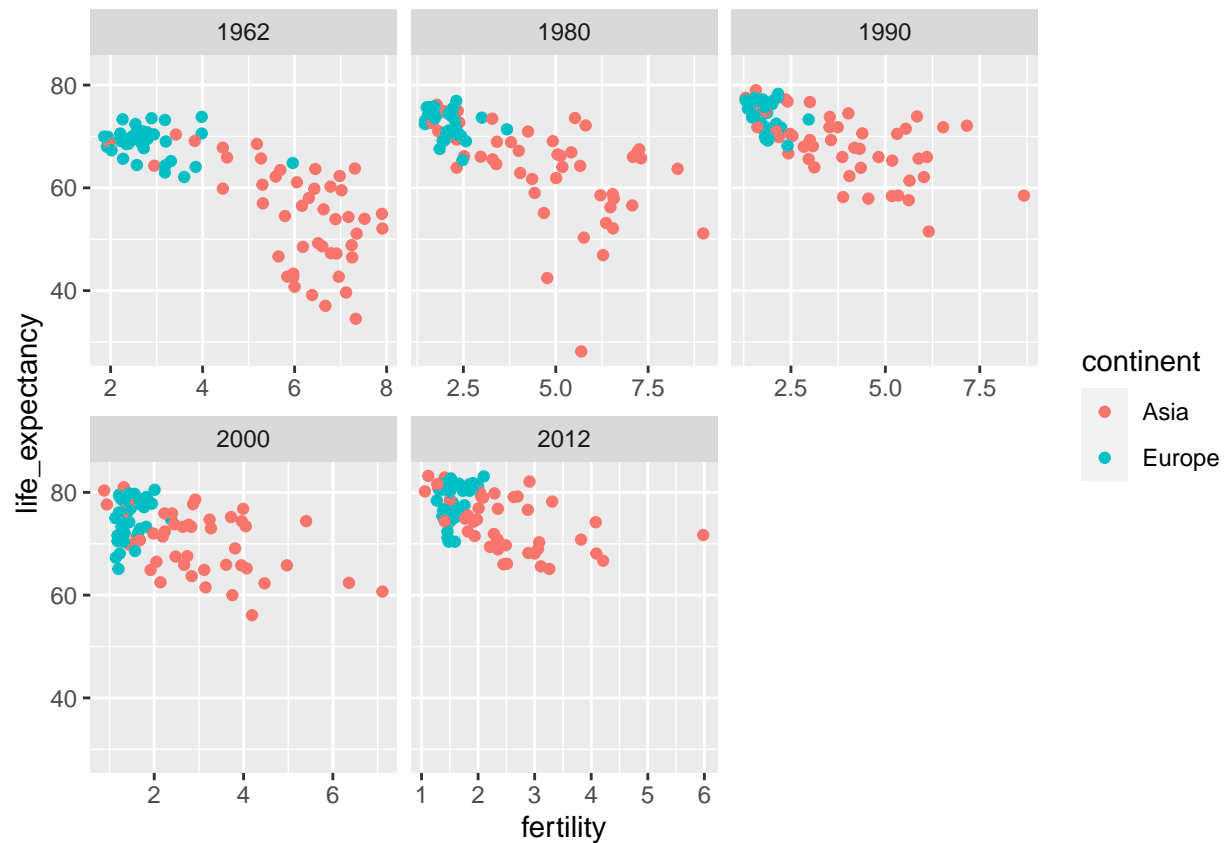
```
years = c(1962,1980,1990,2000,2012)
continents = c("Europe","Asia")
gapminder %>%
  filter(continent %in% continents) %>%
  filter(year %in% years) %>%
  ggplot(aes(fertility,life_expectancy,col=continent))+
  geom_point()+
  facet_wrap(~year)
```



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



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

```
con = gapminder %>%
  filter(continent %in% c("Asia", "Europe")) %>%
  group_by(continent) %>%
  summarize(min = min(fertility, na.rm=T)) %>%
  select(min)
e_min = as.double(con[2,])
country = gapminder %>%
  filter(continent %in% c("Asia")) %>%
  filter(year == 2012) %>%
  filter(fertility < e_min) %>%
  arrange(desc(life_expectancy))
country$fertility < 3
```

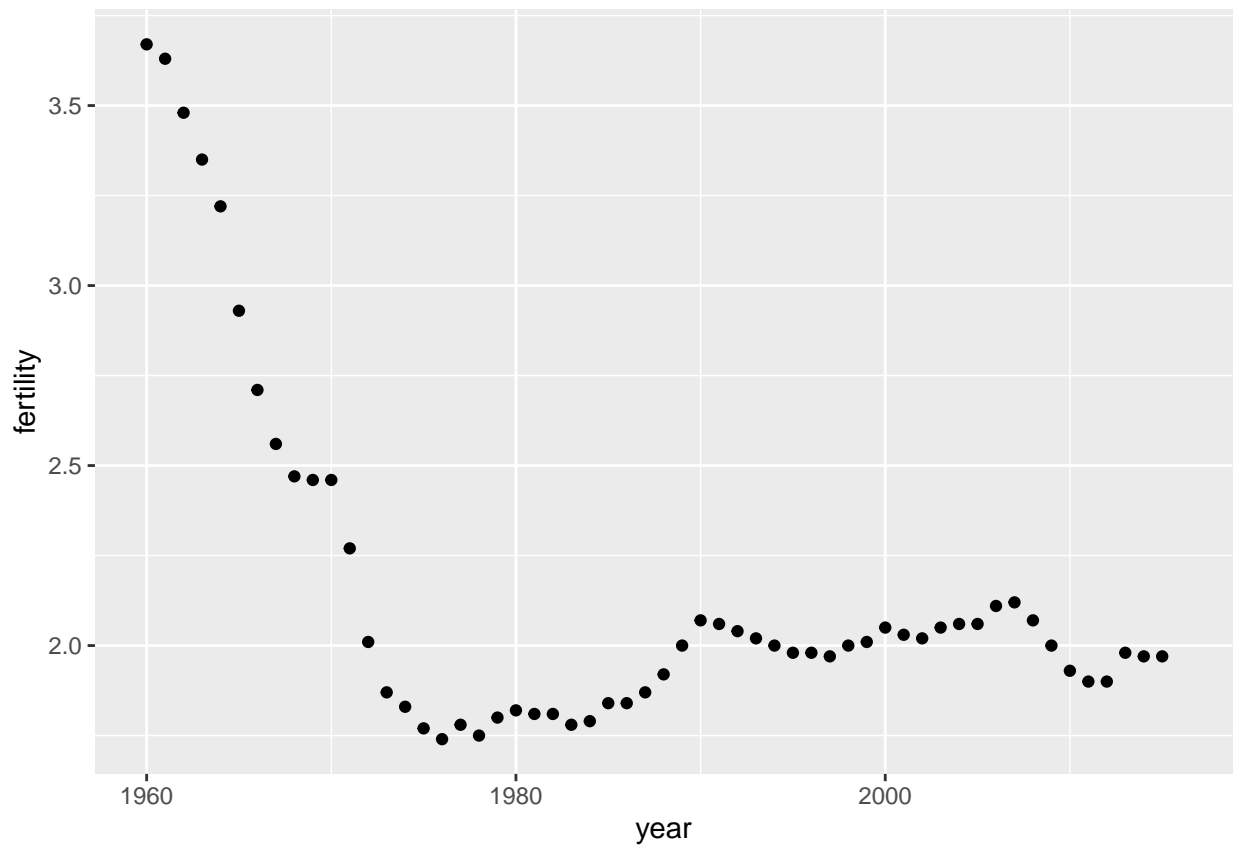
```
2012      2
```

```
## [1] TRUE TRUE
```

Time Series plot()

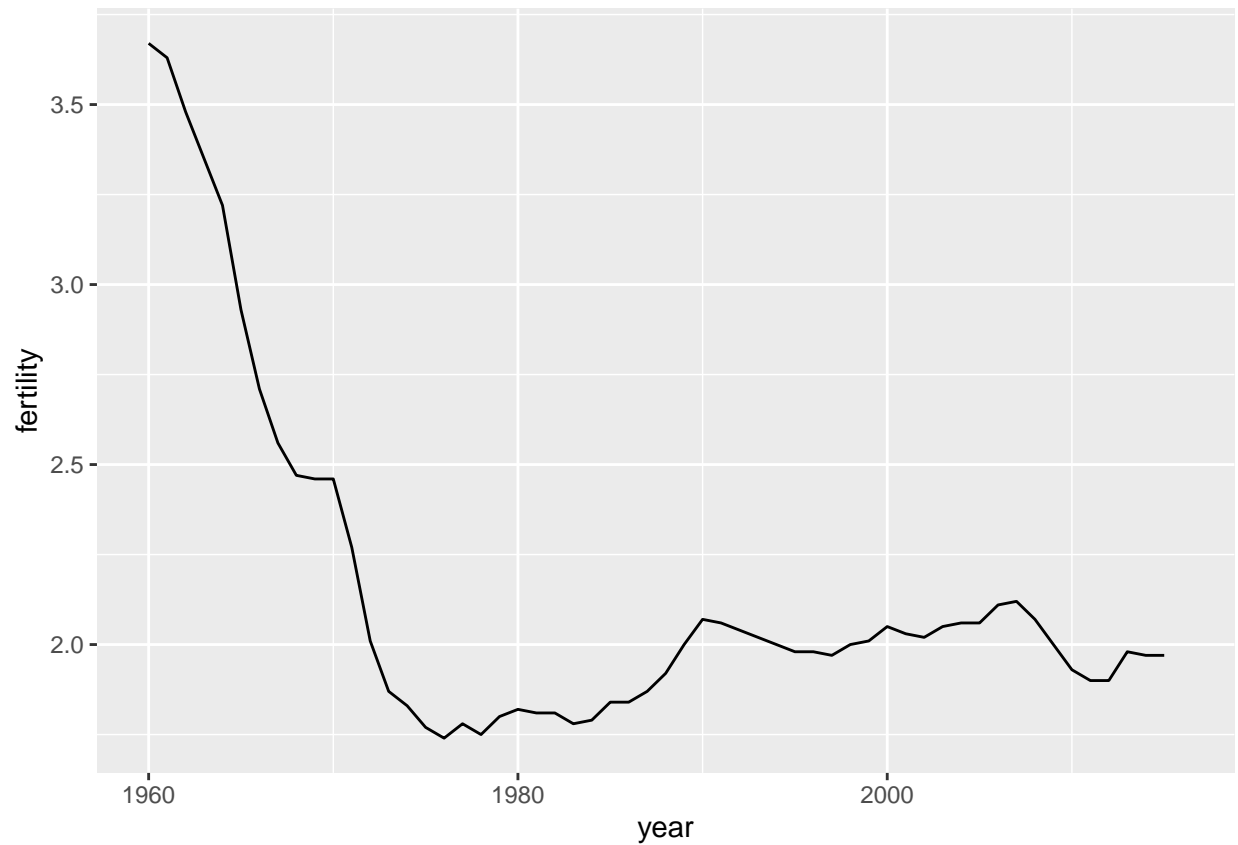
```
gapminder %>%  
  filter(country=="United States") %>%  
  ggplot(aes(year,fertility)) +  
  geom_point()
```

Warning: Removed 1 rows containing missing values (geom_point).



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

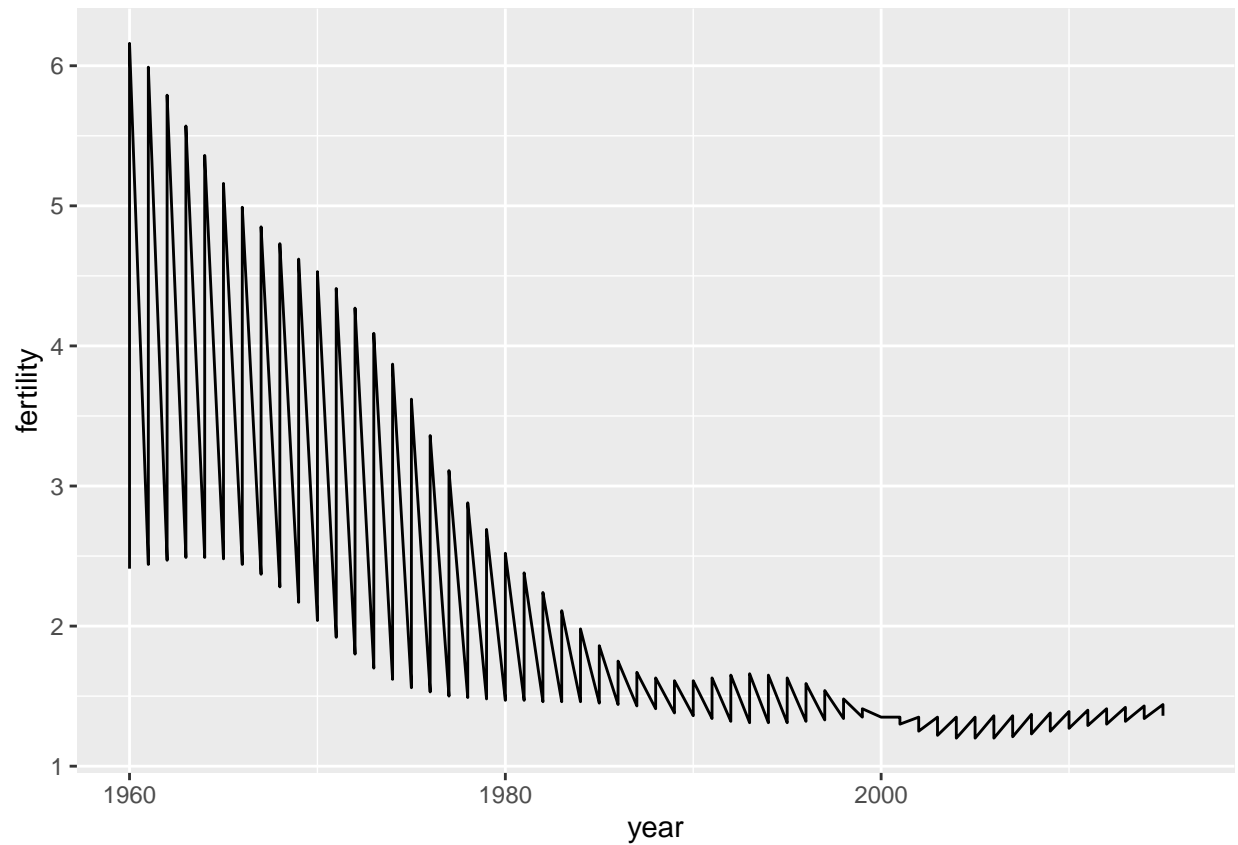
Warning: Removed 1 row(s) containing missing values (geom_path).



```
countries = c("South Korea", "Germany")

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

```
## Warning: Removed 2 row(s) containing missing values (geom_path).
```

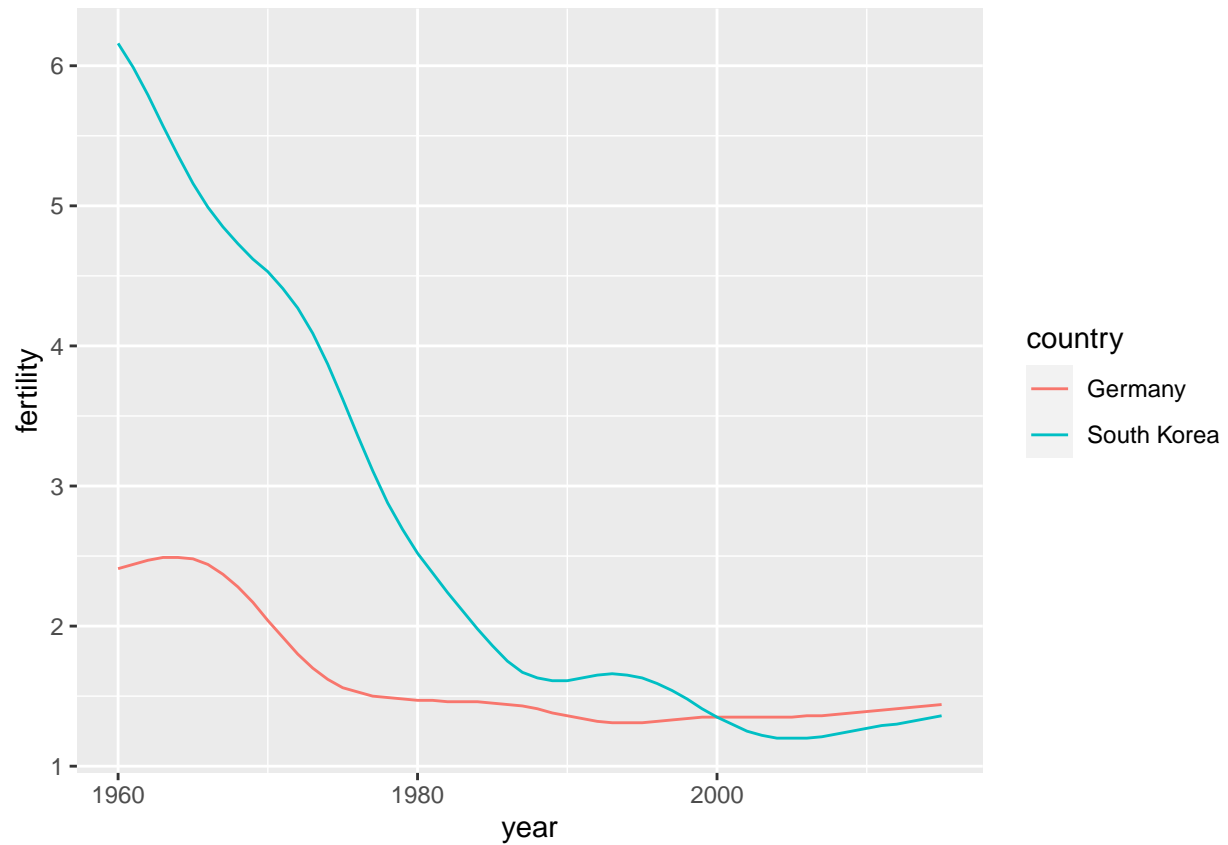


`ggplot(group=)`

group!

```
countries = c("South Korea", "Germany")

gapminder %>%
  filter(country %in% countries & !is.na(fertility)) %>%
  ggplot(aes(year, fertility, group=country)) +
  geom_line(aes(color=country))
```

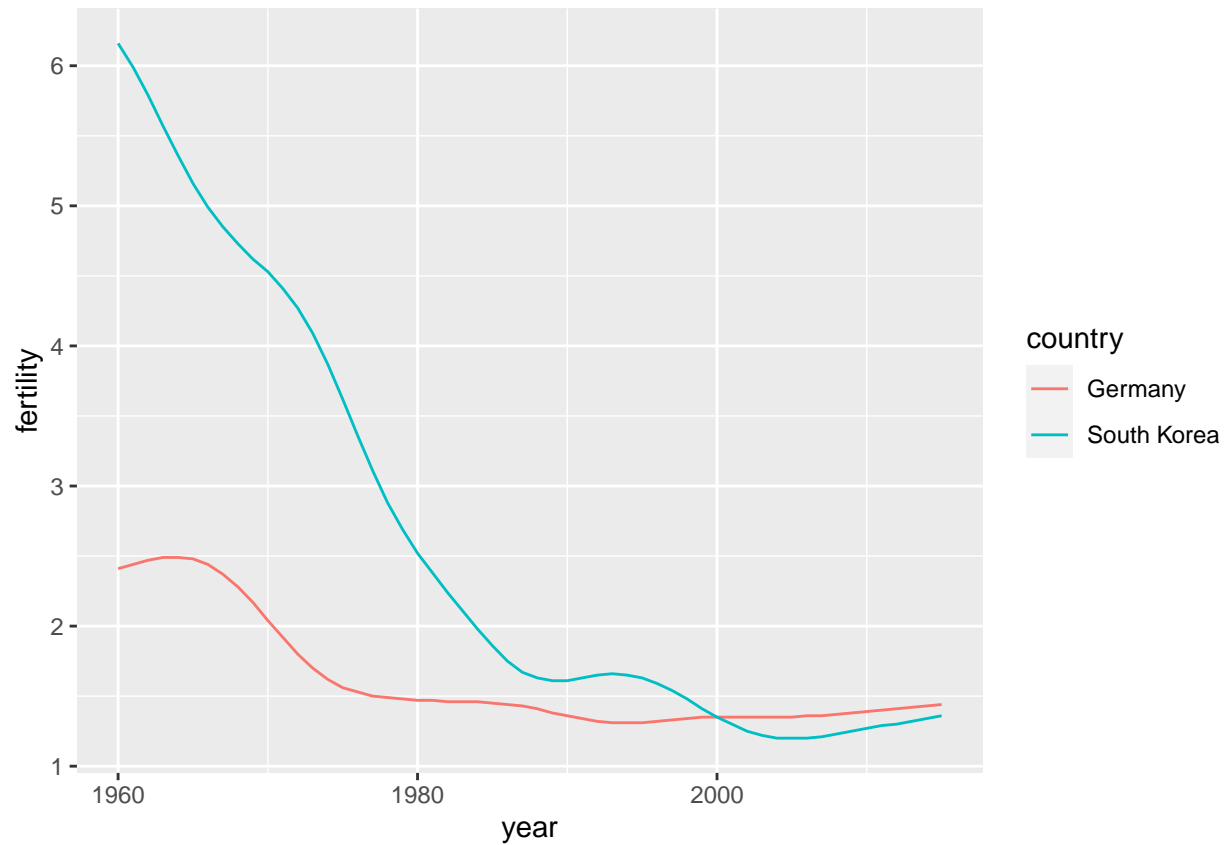


color

group

```
countries = c("South Korea","Germany")

gapminder %>%
  filter(country %in% countries & !is.na(fertility)) %>%
  ggplot(aes(year,fertility))+
  geom_line(aes(color=country))# ggplot
```

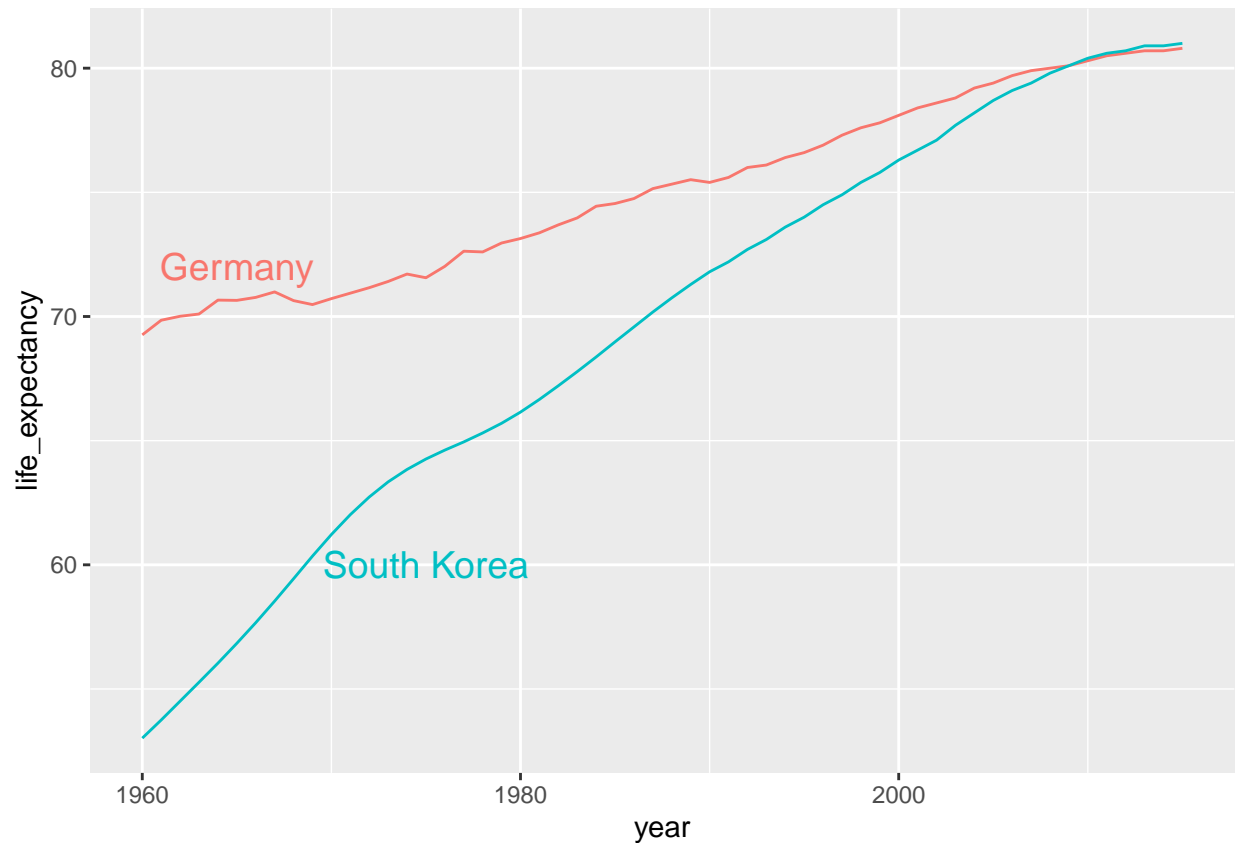


legend

label

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

gapminder %>%
  filter(country %in% countries & !is.na(fertility)) %>%
  ggplot(aes(year,life_expectancy,color=country))+
  geom_line()+
  geom_text(data=labels,aes(x,y,label=country),size=5)+
  #geom_text x,y labels x,y aes(x,y)
  theme(legend.position = "None")
```



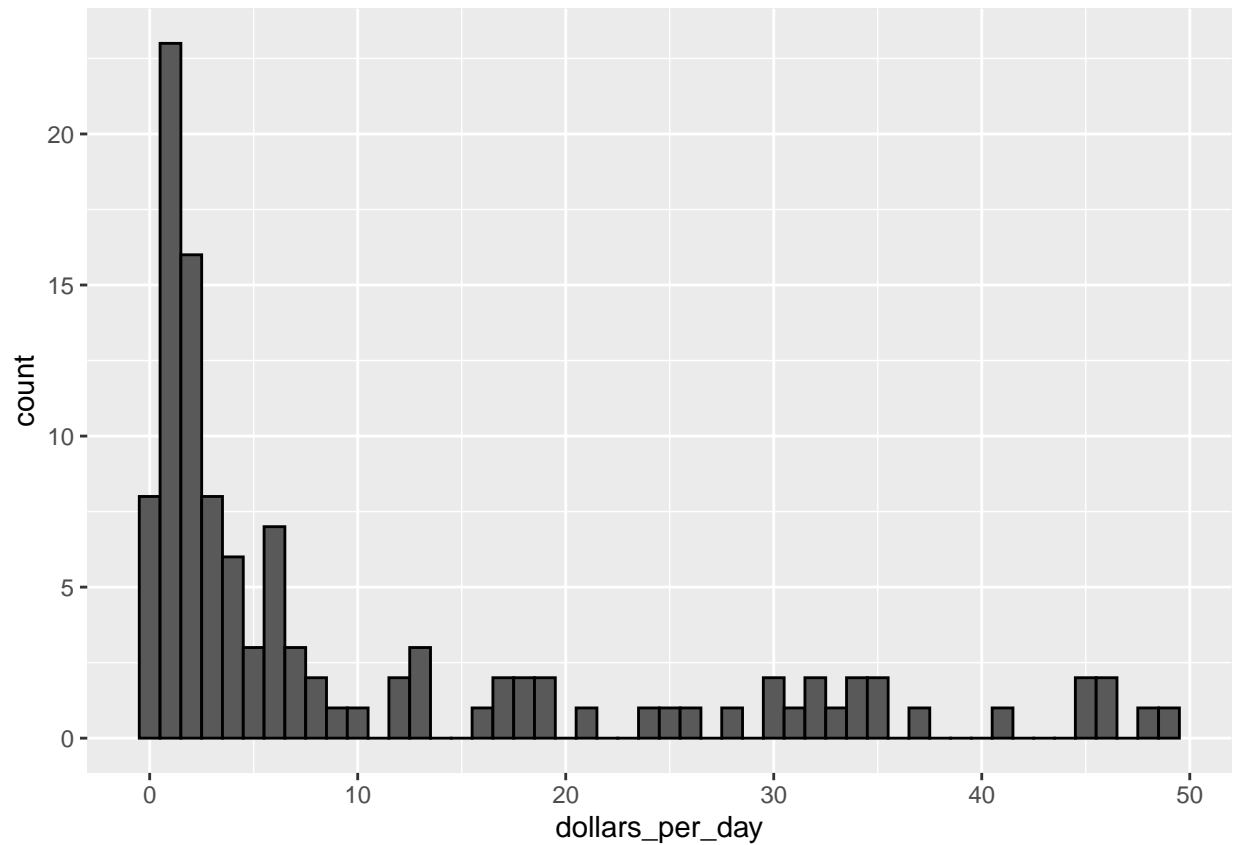
Han Rosling's quiz(2) ## ? 2\$

```
gapminder=gapminder %>%
  mutate(dollars_per_day = gdp/population/365)
names(gapminder)
```

```
## [1] "country"      "year"         "infant_mortality" "life_expectancy"
## [5] "fertility"    "population"   "gdp"            "continent"
## [9] "region"      "dollars_per_day"
```

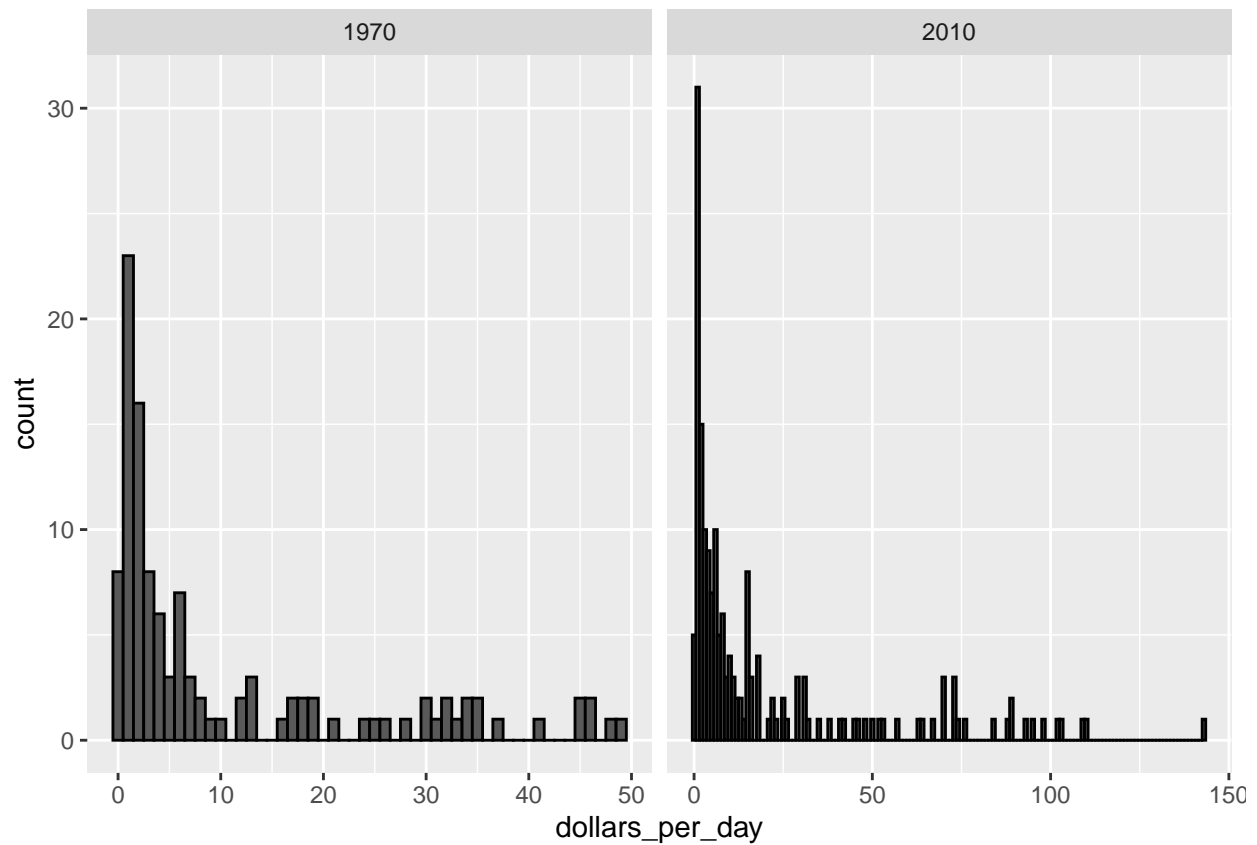
1970

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

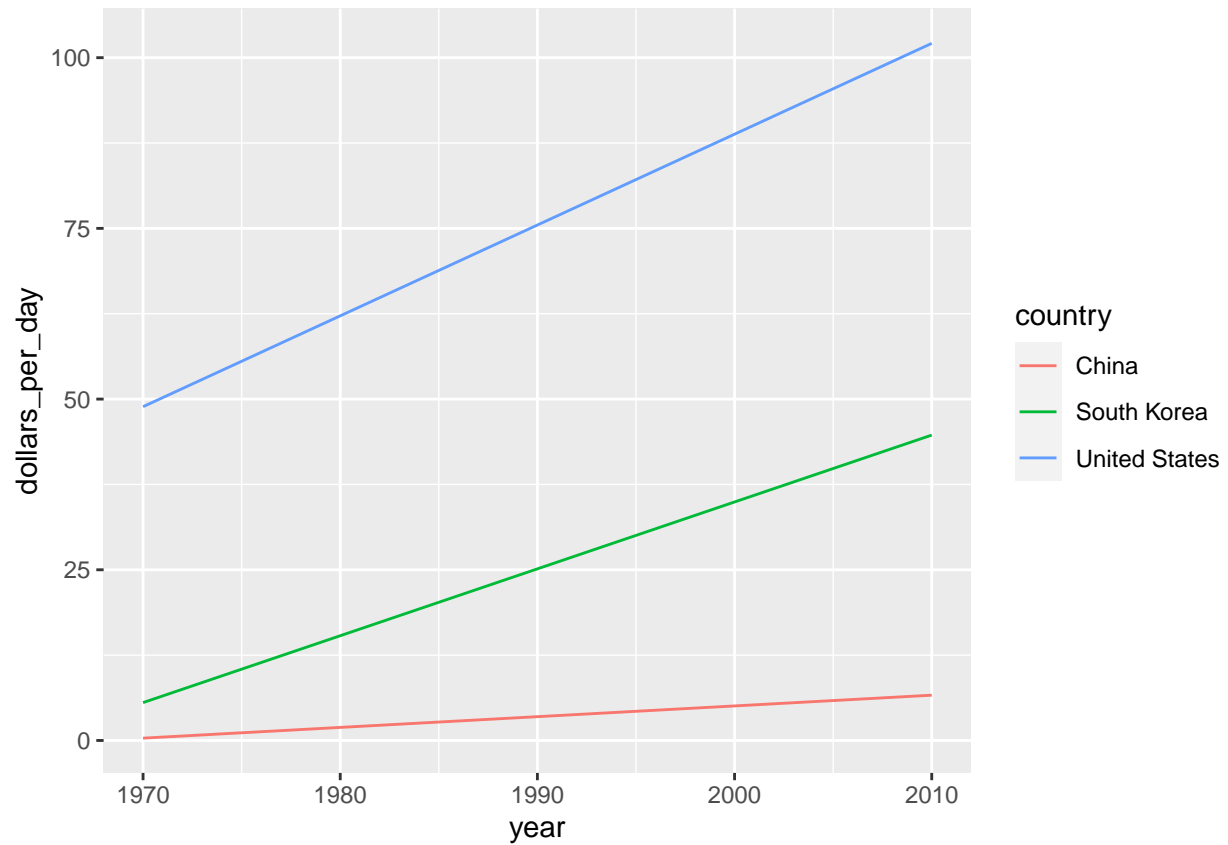


1970 2010

```
past_year = 1970
gapminder %>%
  filter(year %in% c(1970,2010) & !is.na(gdp)) %>%
  ggplot(aes(dollars_per_day))+
  geom_histogram(binwidth=1,color="black")+
  facet_grid(~year,scale="free")
```

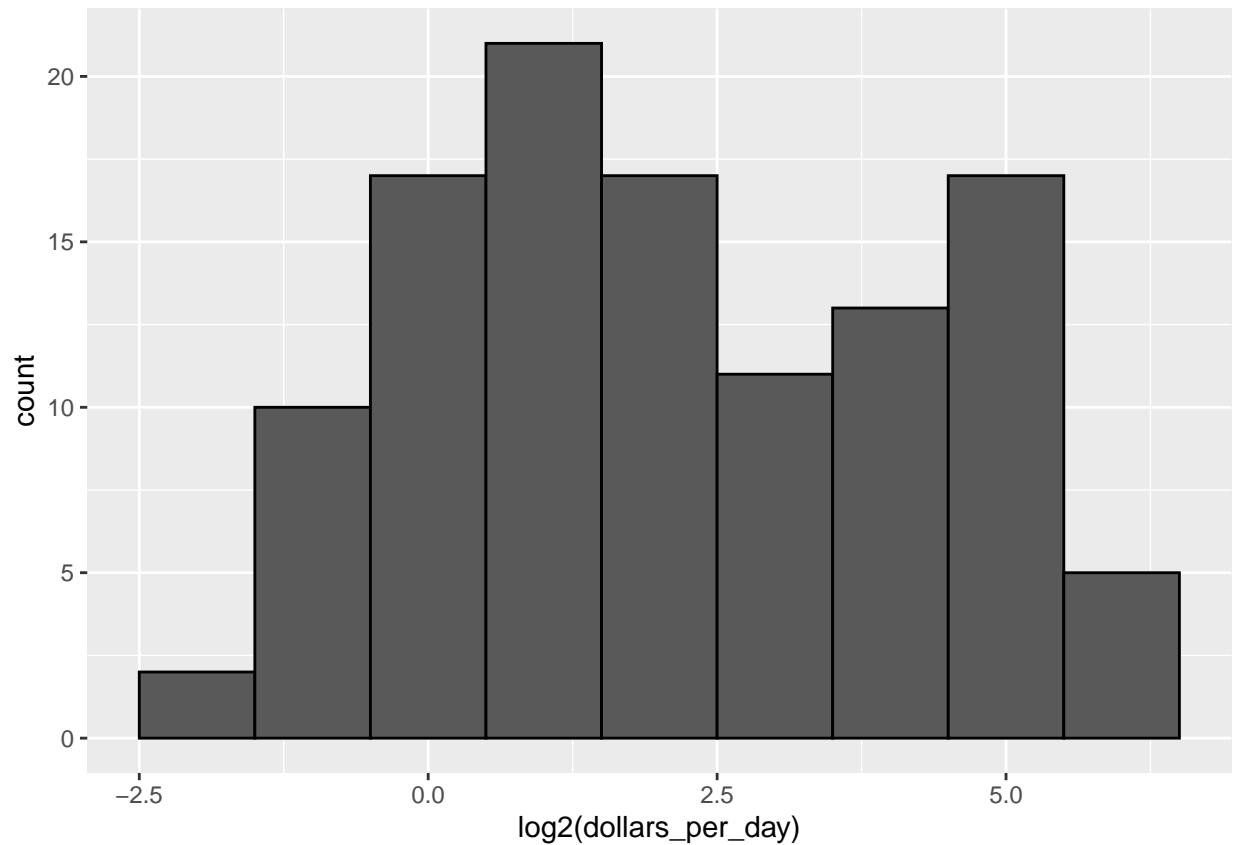
```
gapminder %>%
  filter(year %in% c(1970,2010) & !is.na(gdp) & country %in% c("United States","South Korea","China")) %>%
  ggplot(aes(year,dollars_per_day,color=country))+
  geom_line()
```



log transformation

\$1 \$2 \$4 \$8 \log_2 (hist x 0 ,1 ,2) => **log**

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



log log base => ln (...)

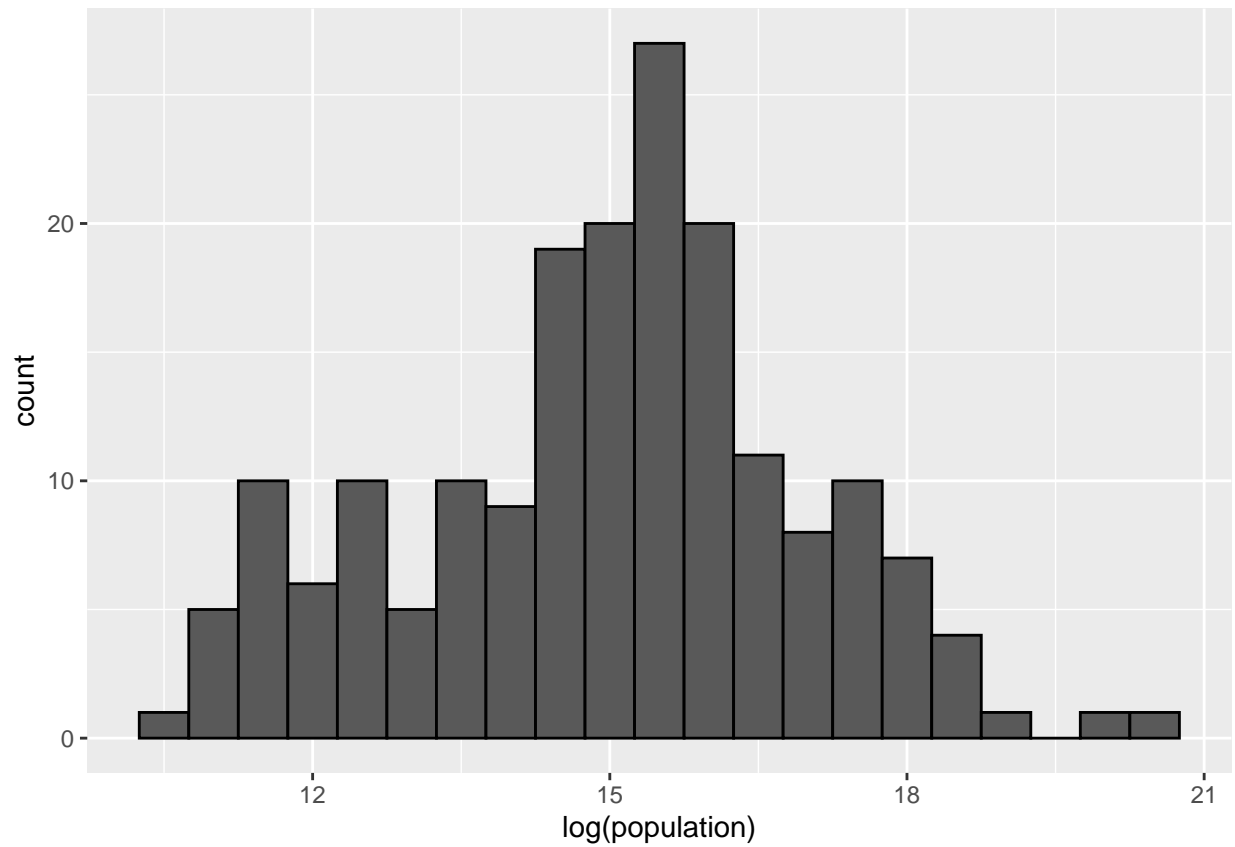
```
a=gapminder %>%
  filter(year==past_year) %>%
  summarise(min=min(population),max=max(population))
```

a

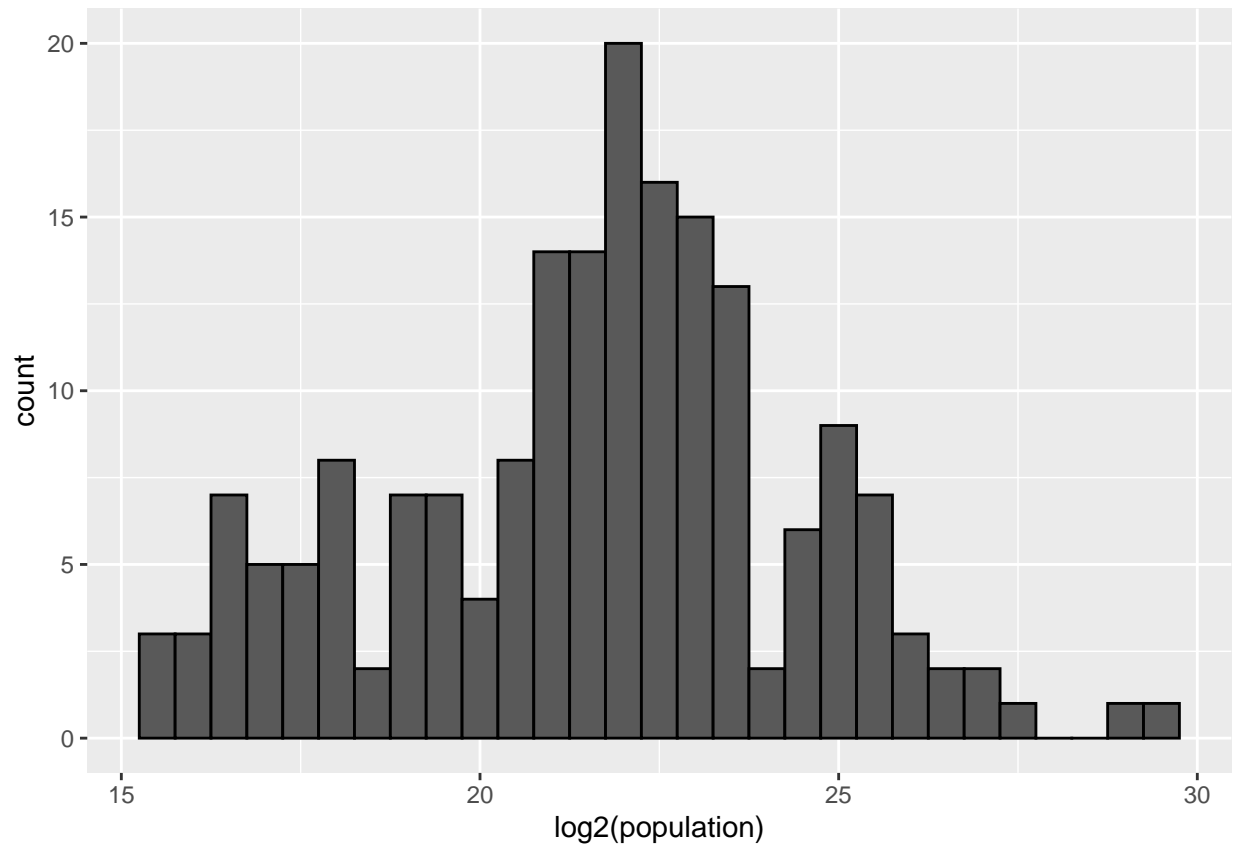
```
## # A tibble: 1 x 2
##   min      max
##   <dbl>   <dbl>
## 1 46075 808510713
```

pop log10

```
gapminder %>%
  filter(year==past_year) %>%
  ggplot(aes(log(population)))+
  geom_histogram(binwidth=0.5,color="black")
```



```
gapminder %>%  
  filter(year==past_year) %>%  
  ggplot(aes(log2(population)))+  
  geom_histogram(binwidth=0.5,color="black")
```

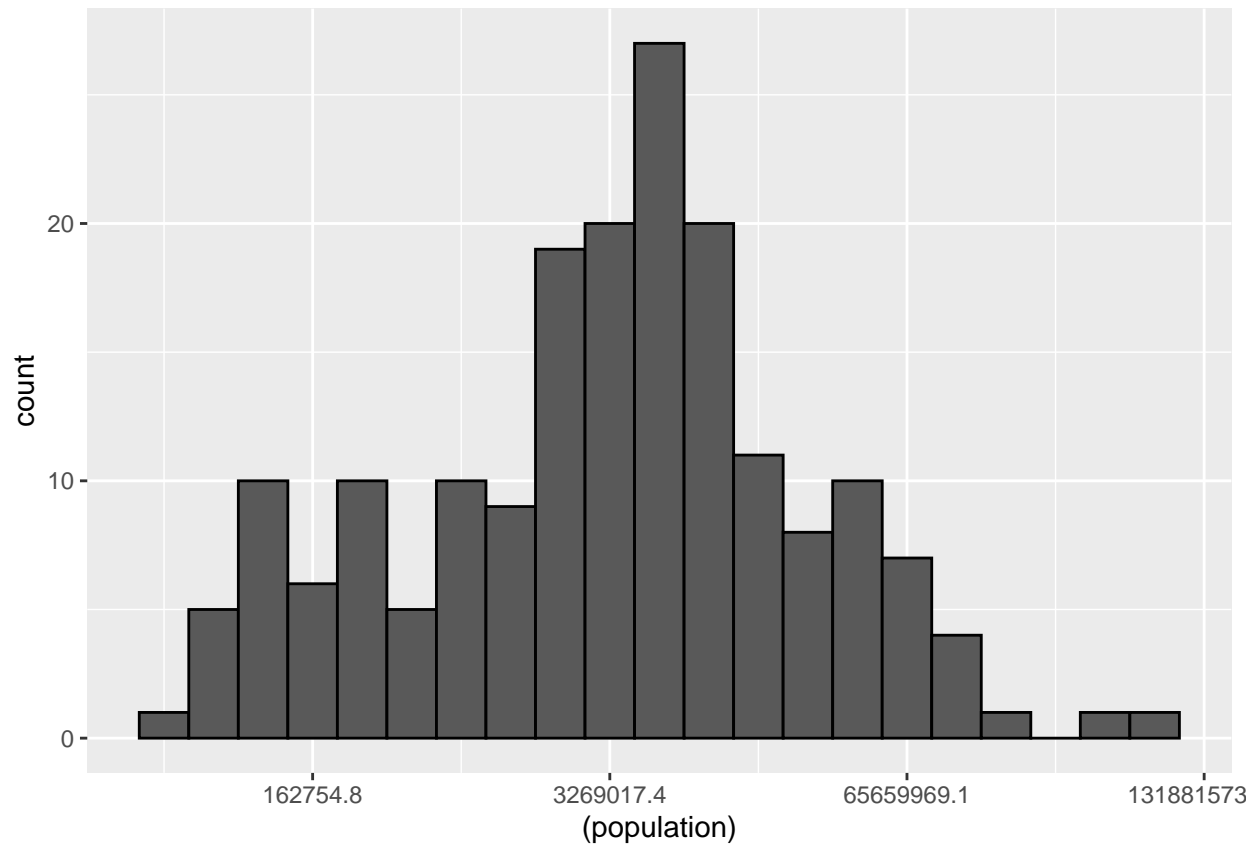


=>log10

scale_x_continuous

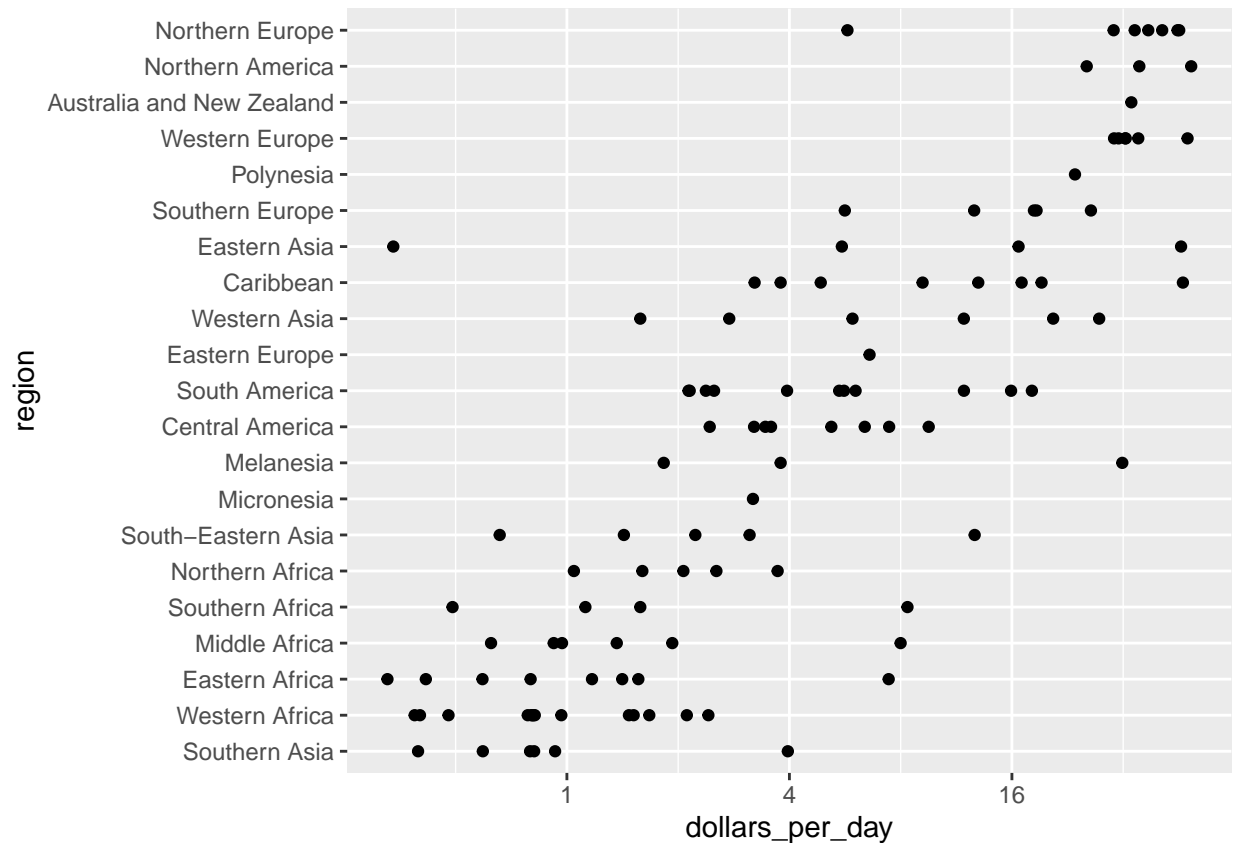
log

```
gapminder %>%  
  filter(year==past_year) %>%  
  ggplot(aes((population)))+  
  geom_histogram(binwidth=0.5,color="black")+  
  scale_x_continuous(trans="log")
```



hist

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



Ridge plot

plot ###

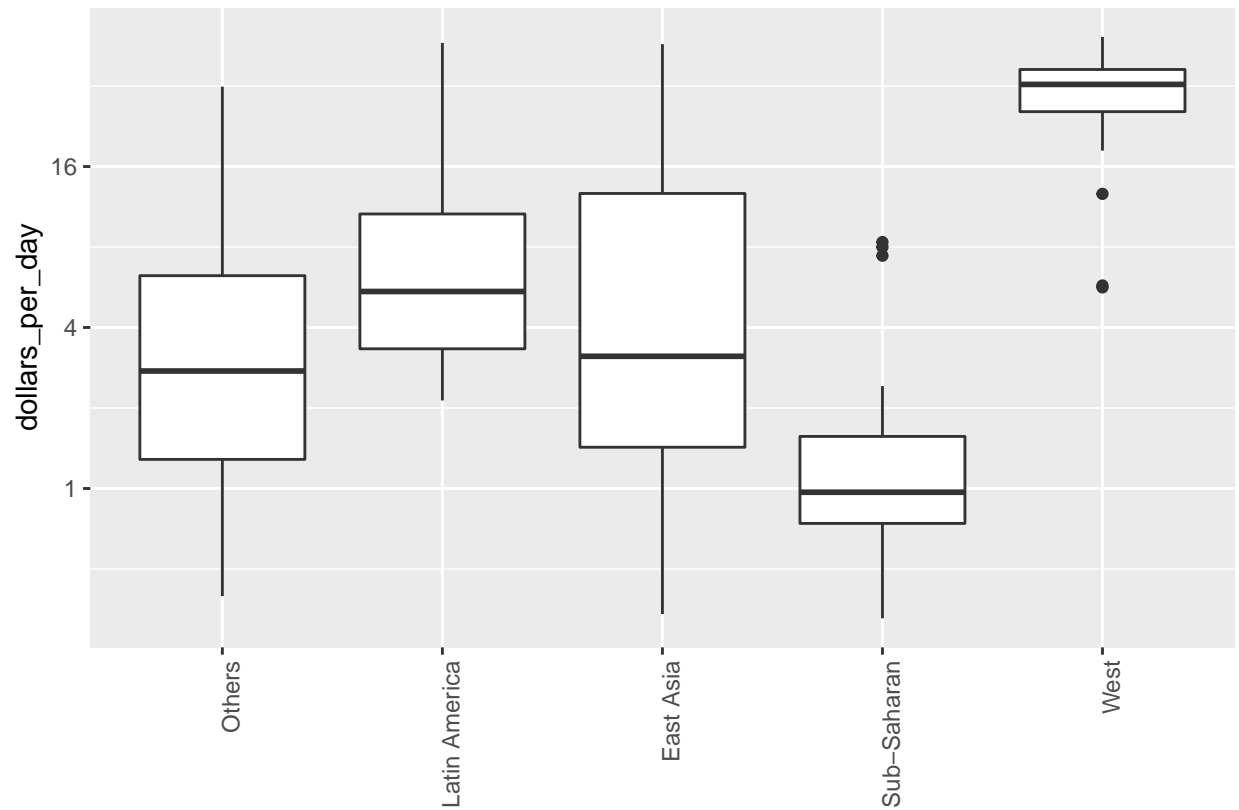
```
gapminder <- gapminder %>%
  mutate(group = case_when(
    region %in% c("Western Europe", "Northern Europe",
      "Southern Europe", "Northern America",
      "Australia and New Zealand") ~ "West",
    region %in% c("Eastern Asia",
      "South-Eastern Asia") ~ "East Asia",
    region %in% c("Caribbean", "Central America",
      "South America") ~ "Latin America",
    continent == "Africa" &
    region != "Northern Africa" ~ "Sub-Saharan",
    TRUE ~ "Others"))

gapminder = gapminder %>%
  mutate(group = factor(group, levels=c("Others", "Latin America", "East Asia", "Sub-Saharan", "West")))
```

```
p=gapminder %>%
  filter(year==past_year & !is.na(gdp)) %>%
```

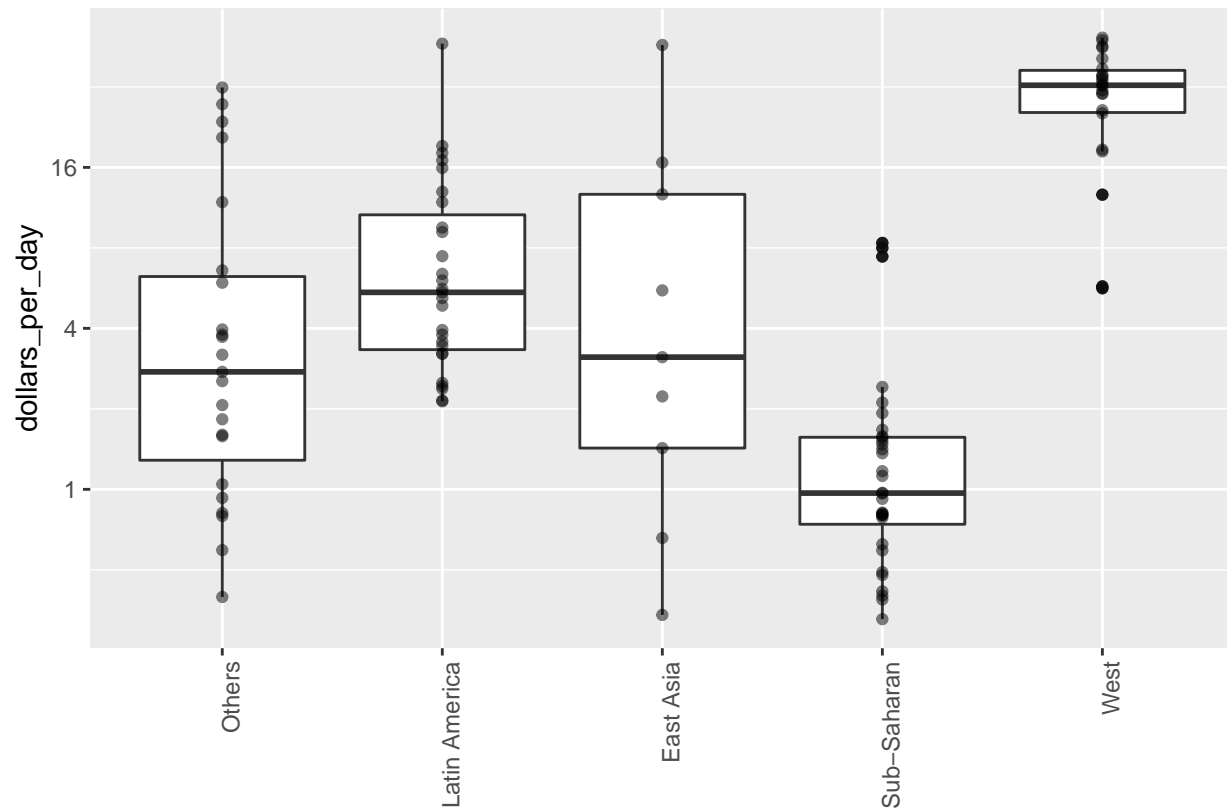
```
ggplot(aes(group,dollars_per_day))+
  geom_boxplot()+
  scale_y_continuous(trans="log2")+
  xlab("")+ #x
  theme(axis.text.x=element_text(angle=90,hjust=1)) #x
```

p



boxplot

```
p+geom_point(alpha=0.5)
```

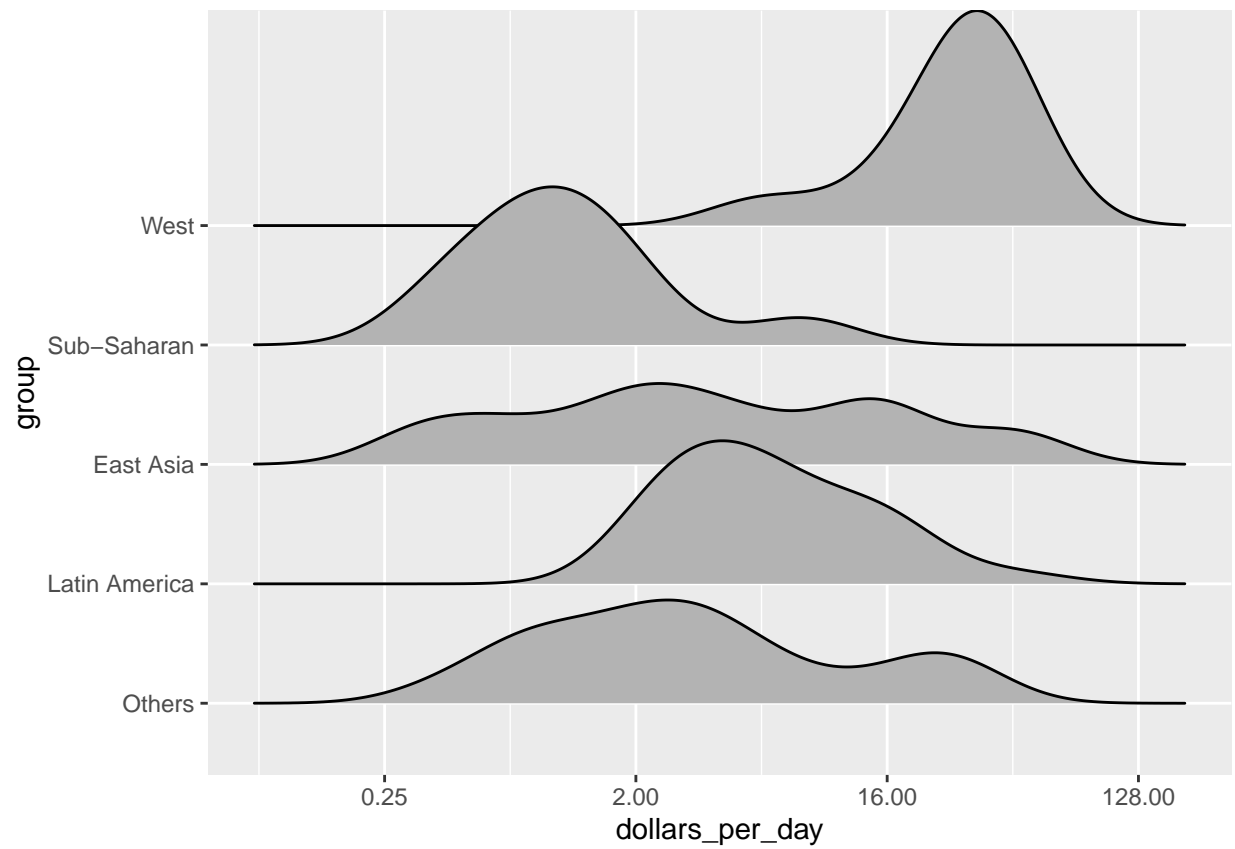
```
library(ggribes)
```

```
## Warning: 'ggribes' R 4.1.1
```

Ridge plot

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

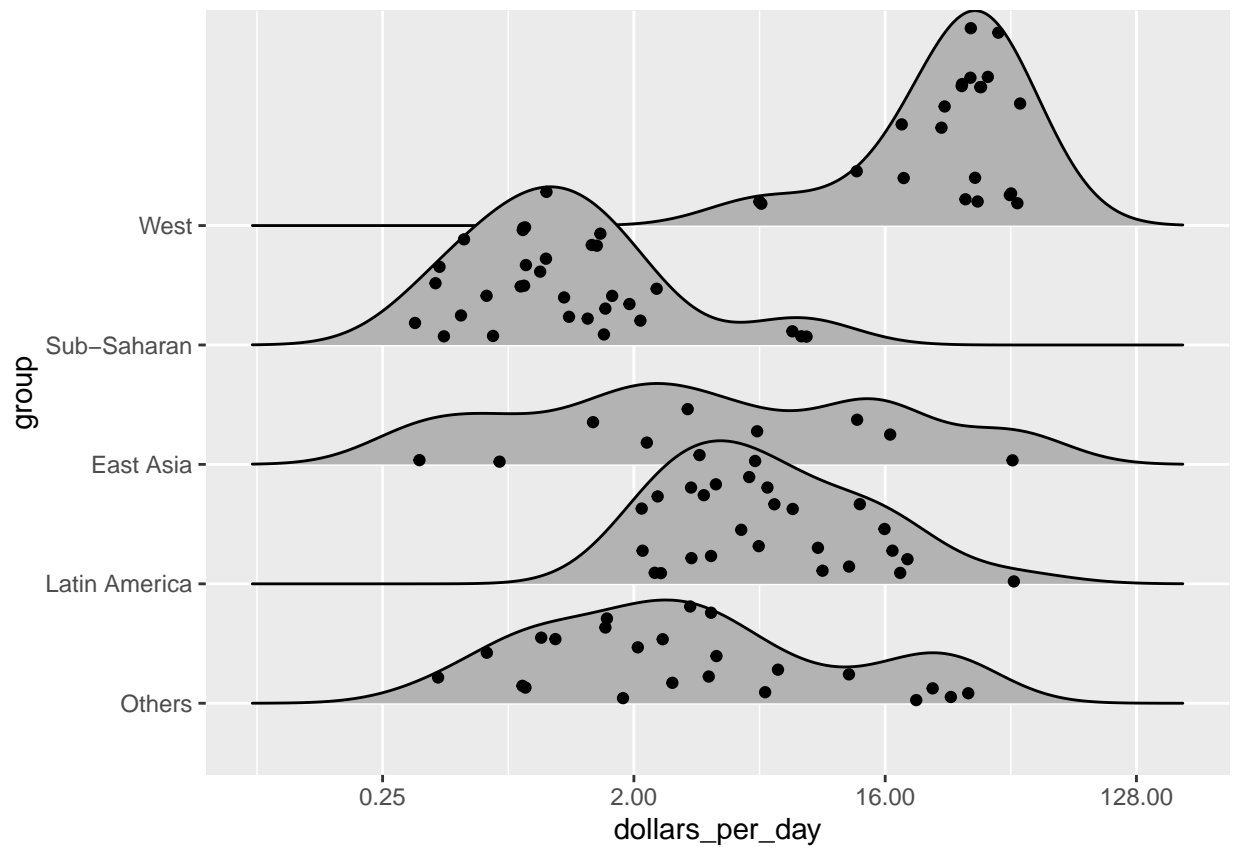
```
## Picking joint bandwidth of 0.648
```



```
p +  
  geom_density_ridges(jittered_points = TRUE)
```

Jittered_points

```
## Picking joint bandwidth of 0.648
```

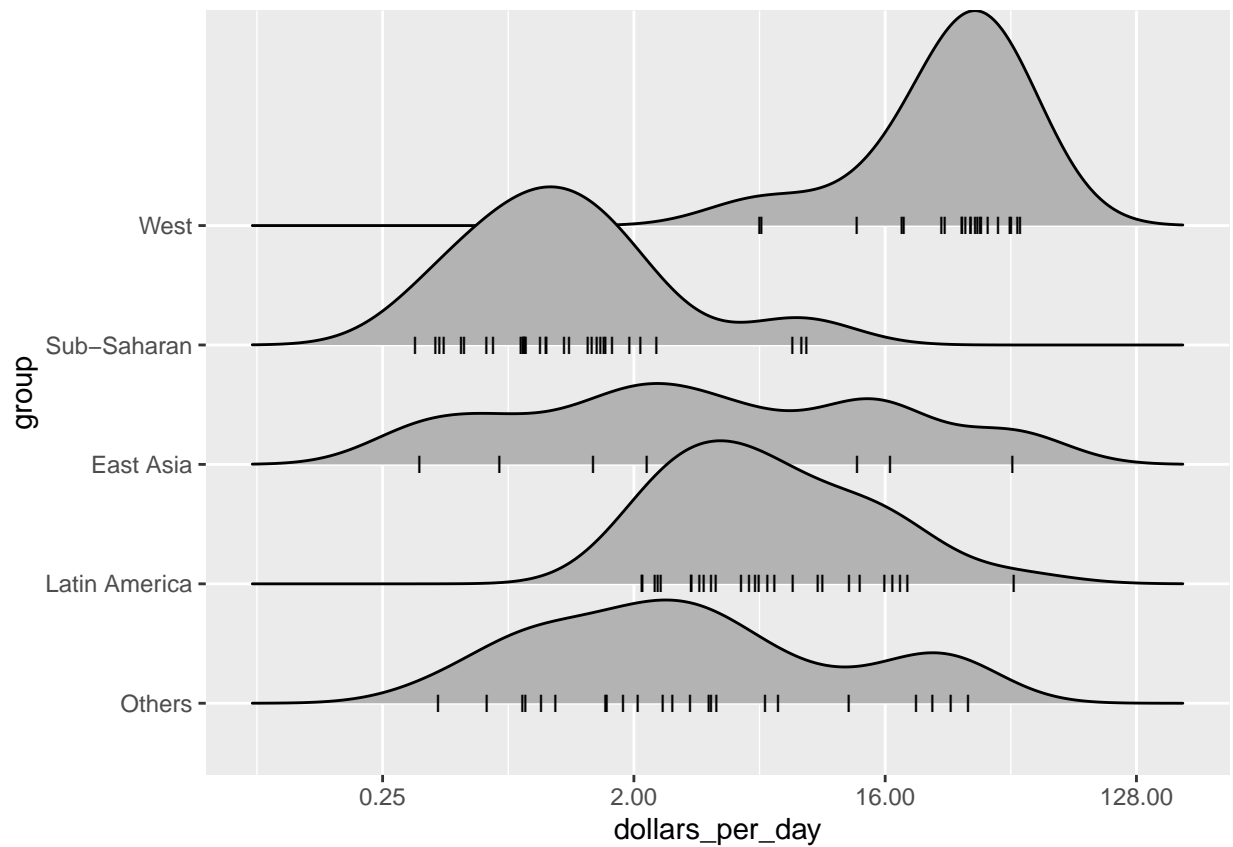


=> y

rug representation data point

```
p+geom_density_ridges(jittered_points = T,
  position=
    position_points_jitter(height=0),
  point_shape='l',point_size=3,
  point_alpha= 1, alpha=1)
```

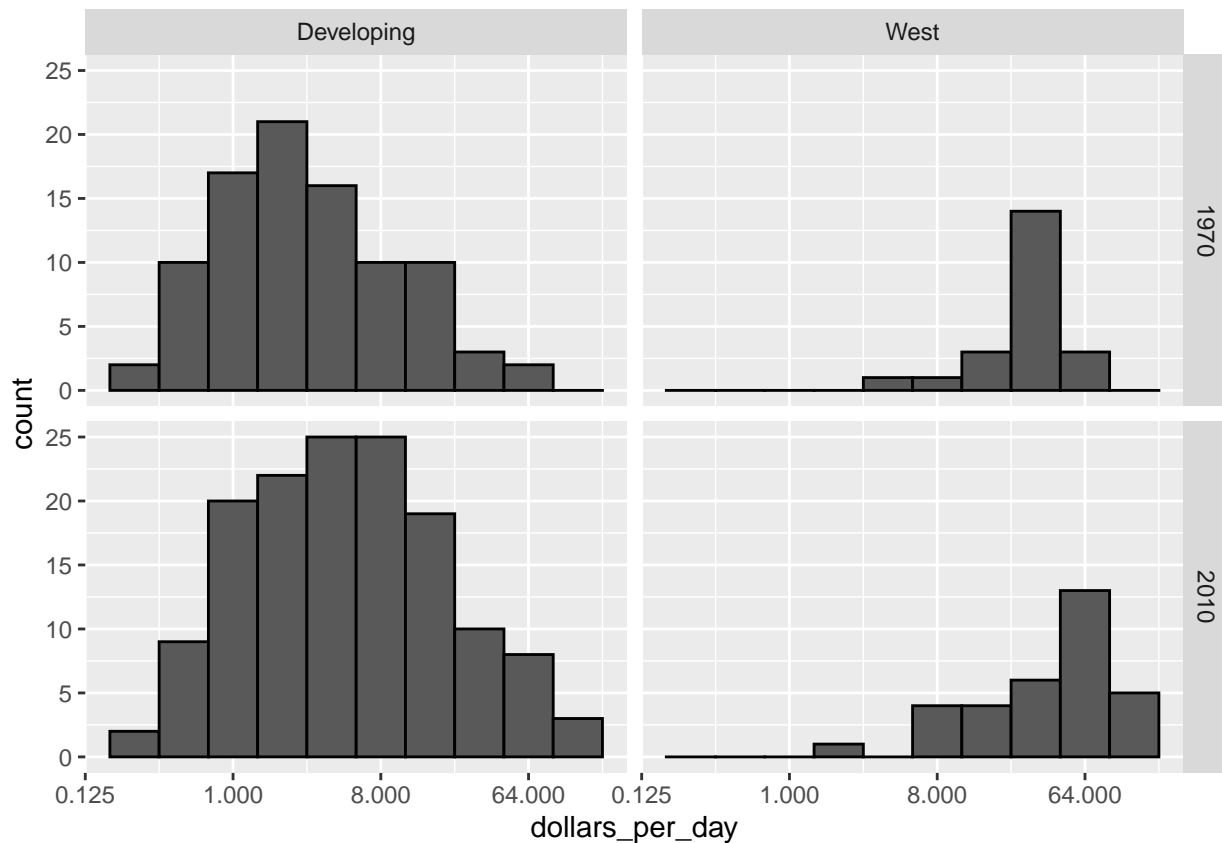
Picking joint bandwidth of 0.648



```

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

```



intersect

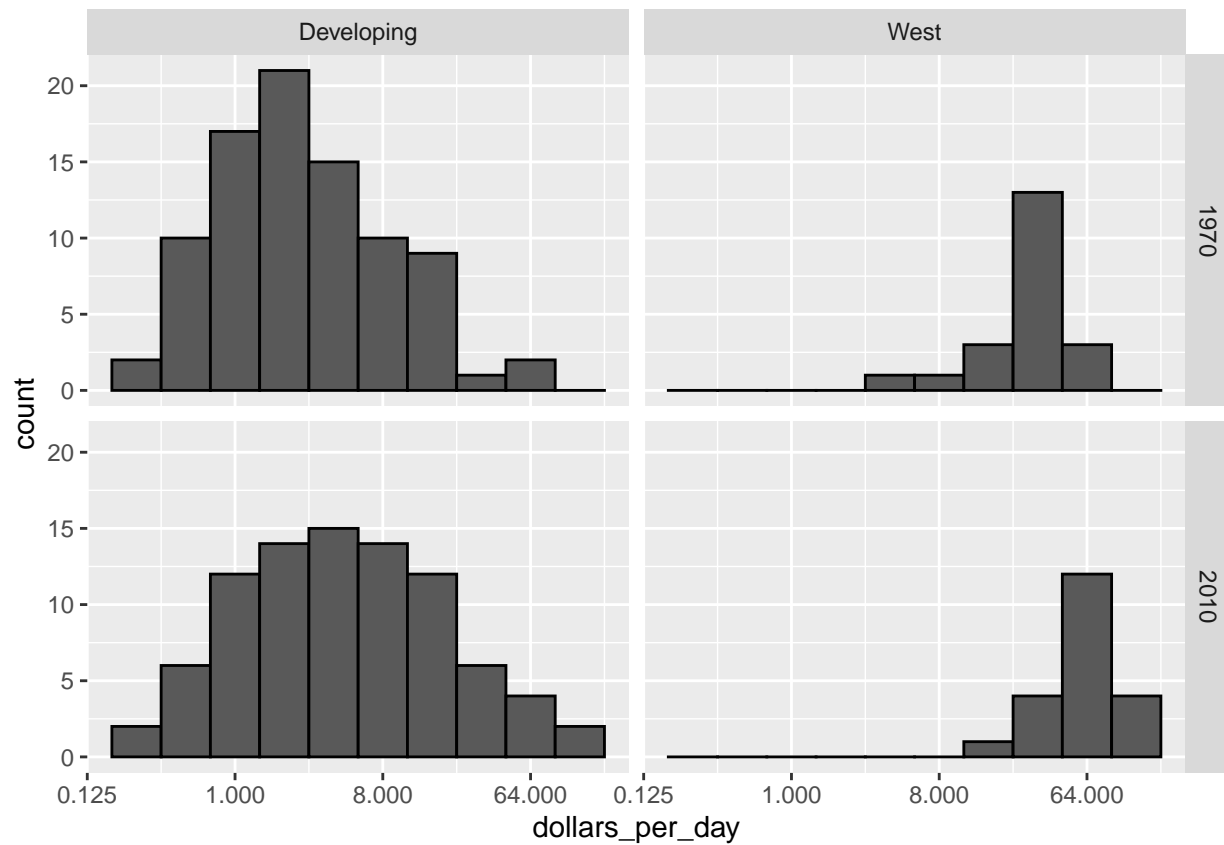
1 2 ==> 1970 2010

```
country_list_1 = gapminder %>%
  filter(year==past_year & !is.na(dollars_per_day)) %>%
  pull(country)

country_list_2 = gapminder %>%
  filter(year==present_year & !is.na(dollars_per_day)) %>%
  pull(country)

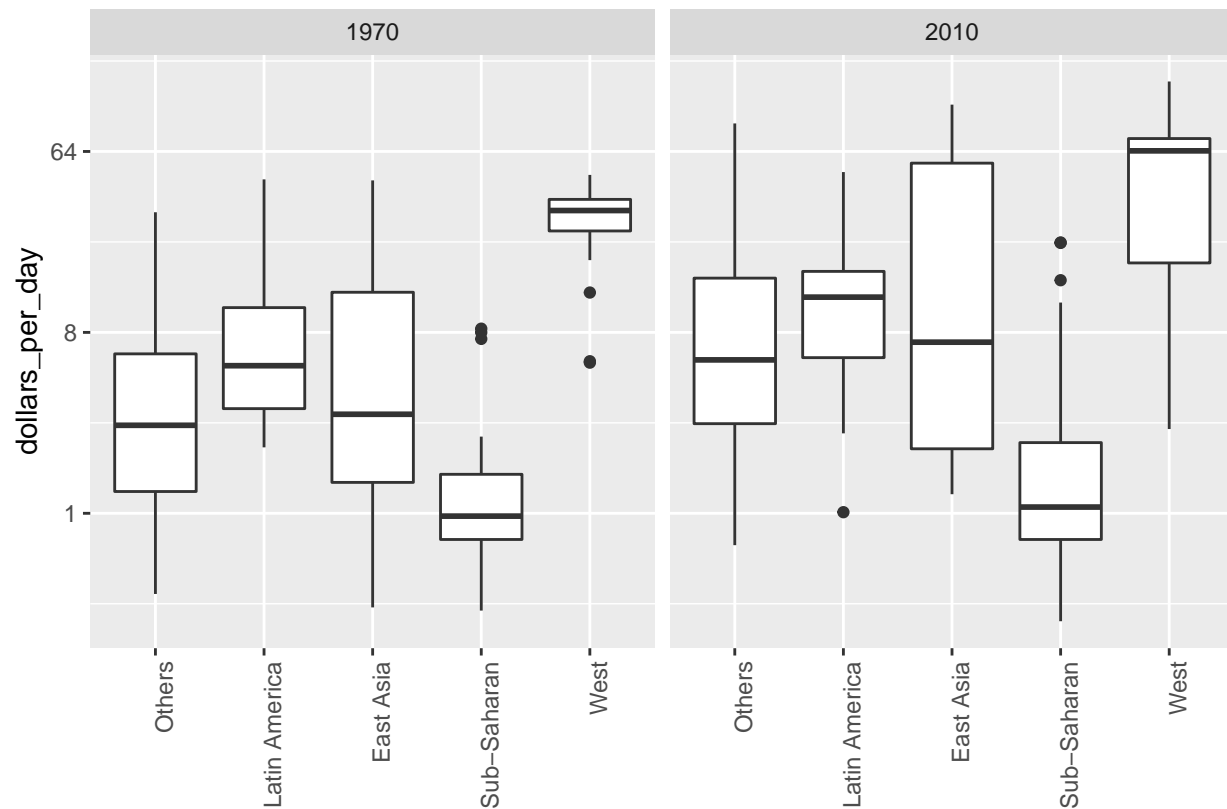
country_list=intersect(country_list_1,country_list_2)
```

```
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)
```



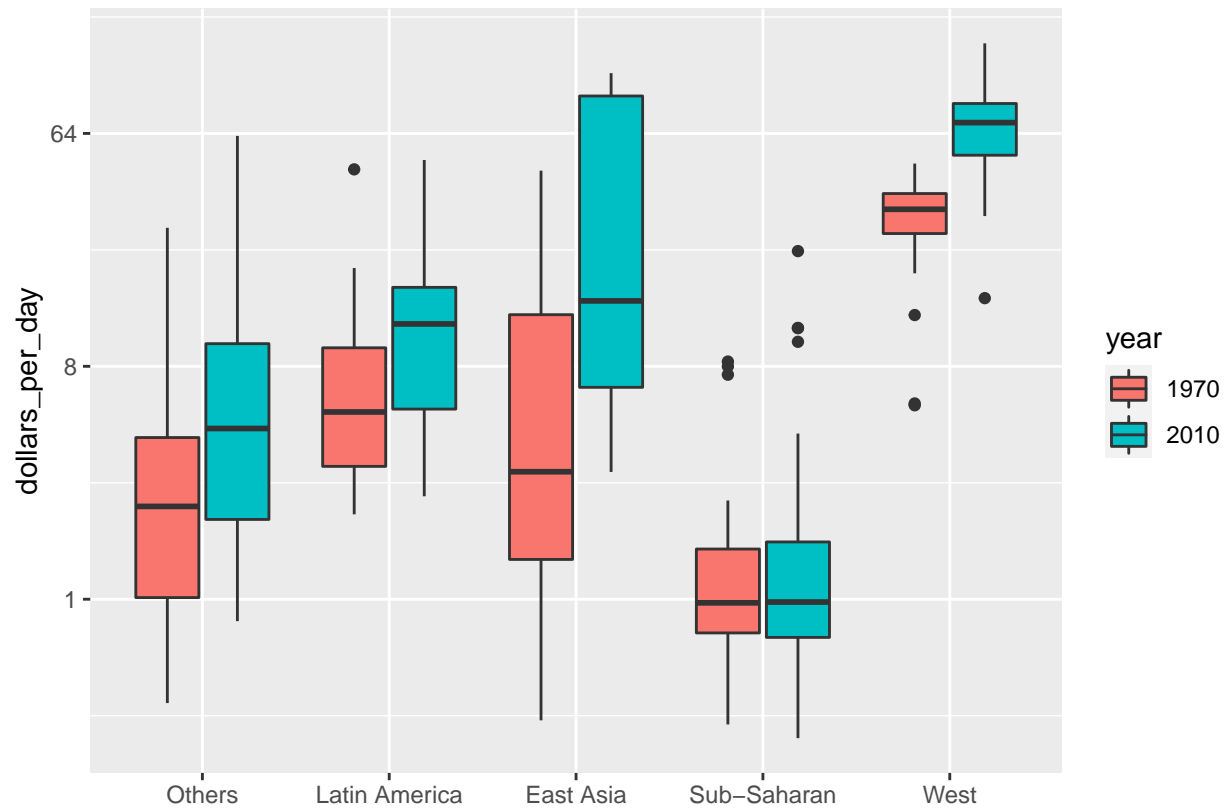
boxplot

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



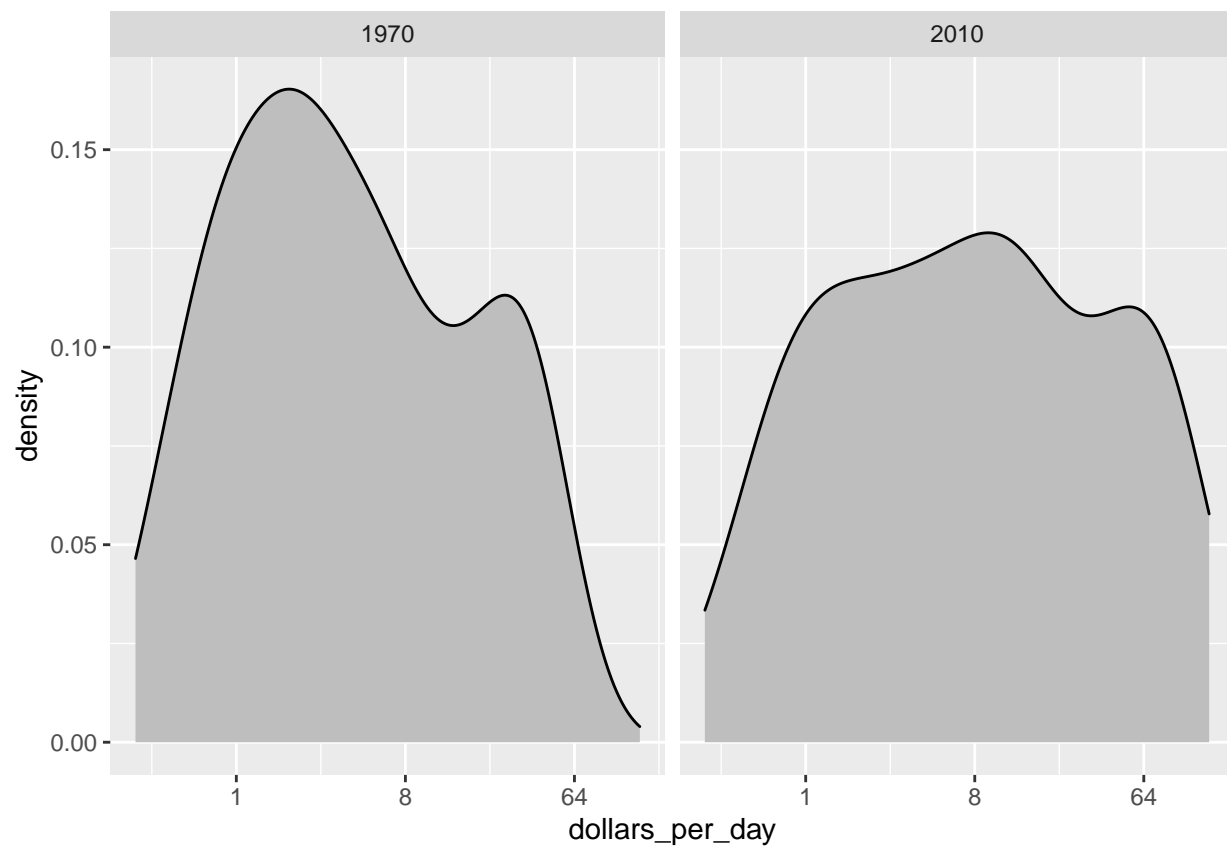
year factor

```
gapminder %>%
  filter(year %in% years & country %in% country_list & !is.na(gdp)) %>%
  mutate(year=factor(year)) %>%
  ggplot(aes(group,dollars_per_day,fill=year))+
  # year numeric fill
  geom_boxplot()+
  scale_y_continuous(trans="log2")+
  xlab("")
```

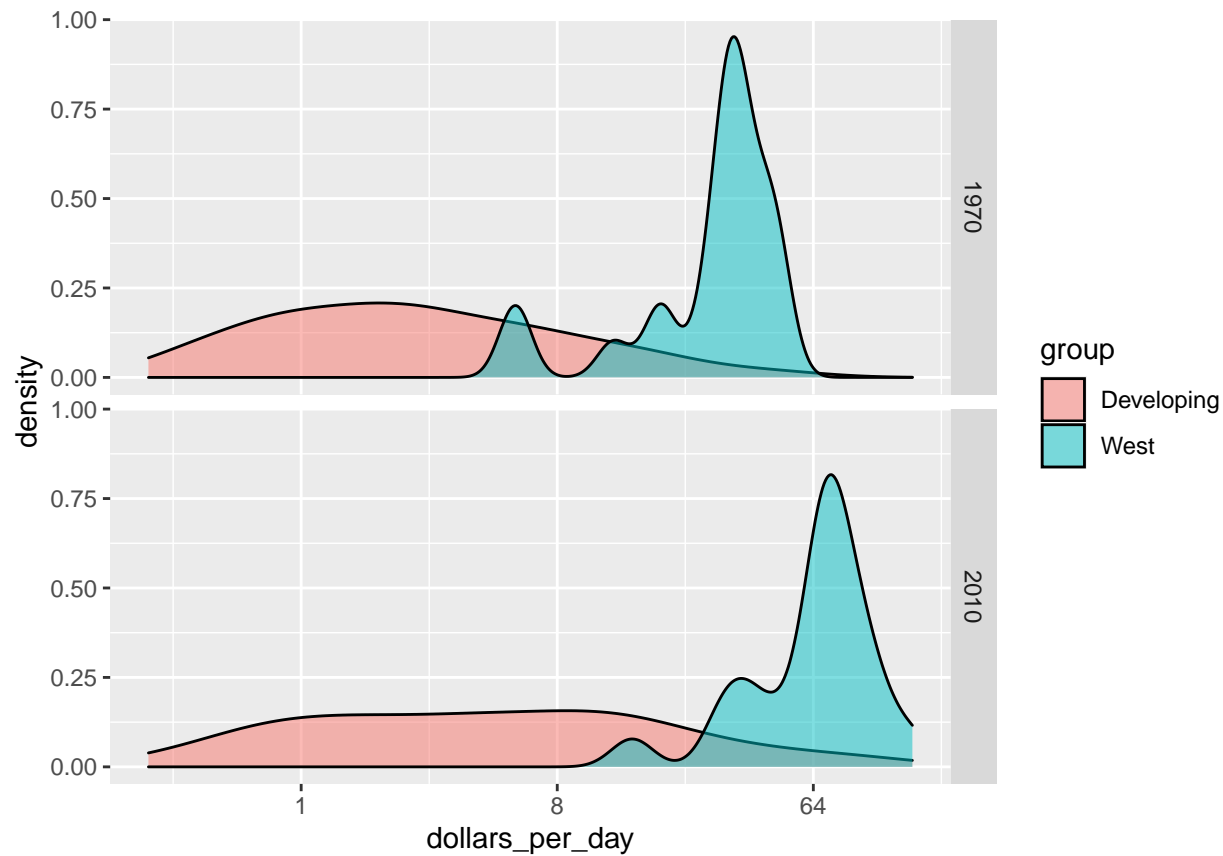


Density

```
gapminder %>%
  filter(year %in% years & !is.na(gdp) & country %in% country_list) %>%
  ggplot(aes(dollars_per_day))+
  geom_density(fill='grey')+
  scale_x_continuous(trans="log2")+
  facet_grid(~year)
```

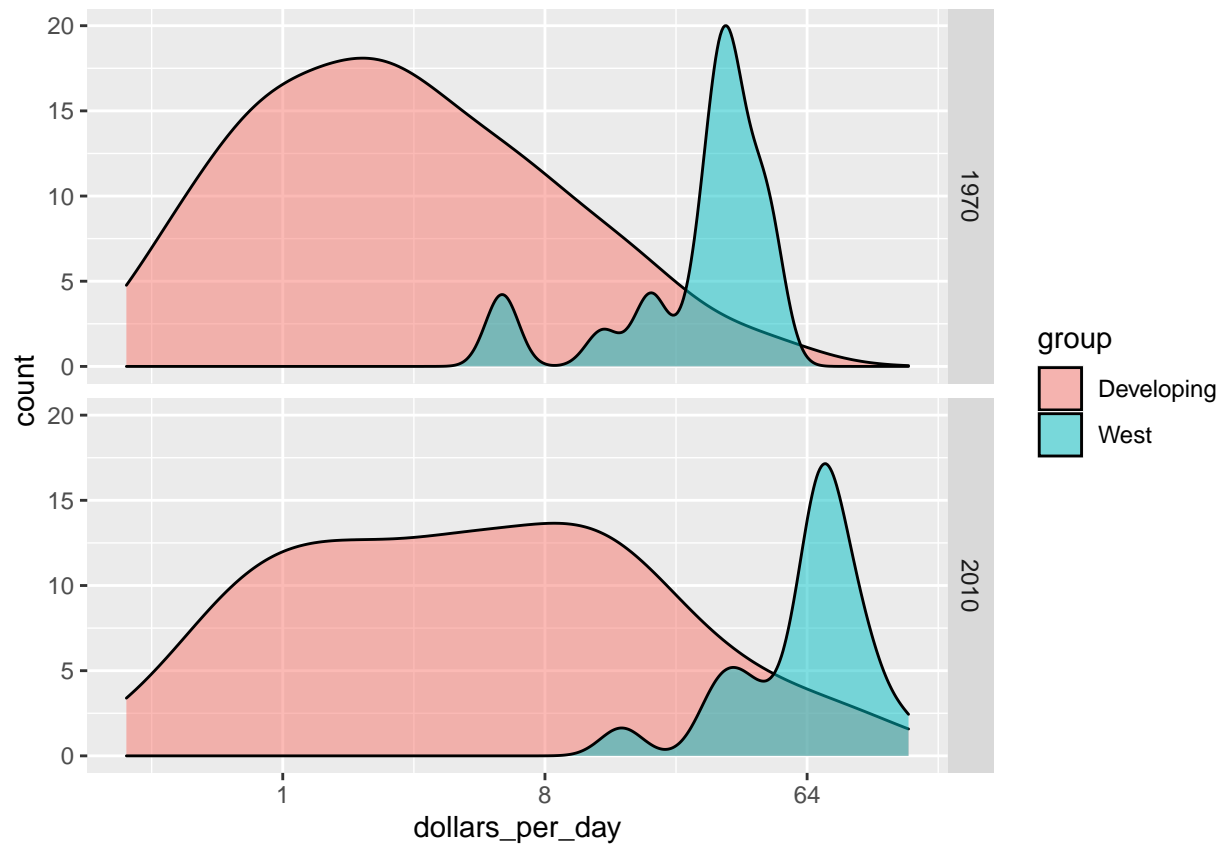
```
gapminder %>%  
  filter(year %in% years & country %in% country_list) %>%  
  mutate(group = ifelse(group=="West", "West", "Developing")) %>%  
  ggplot(aes(dollars_per_day, fill=group)) +  
  geom_density(alpha=.5) +  
  scale_x_continuous(trans="log2") +  
  facet_grid(year~.)
```



-> , -> ?

Density y count ==> Density (Developing)

```
gapminder %>%
  filter(year %in% years & country %in% country_list) %>%
  mutate(group = ifelse(group=="West", "West", "Developing")) %>%
  ggplot(aes(dollars_per_day, fill=group, y=..count..)) +
  #
  geom_density(alpha=.5)+
  scale_x_continuous(trans="log2")+
  facet_grid(year~.)
```



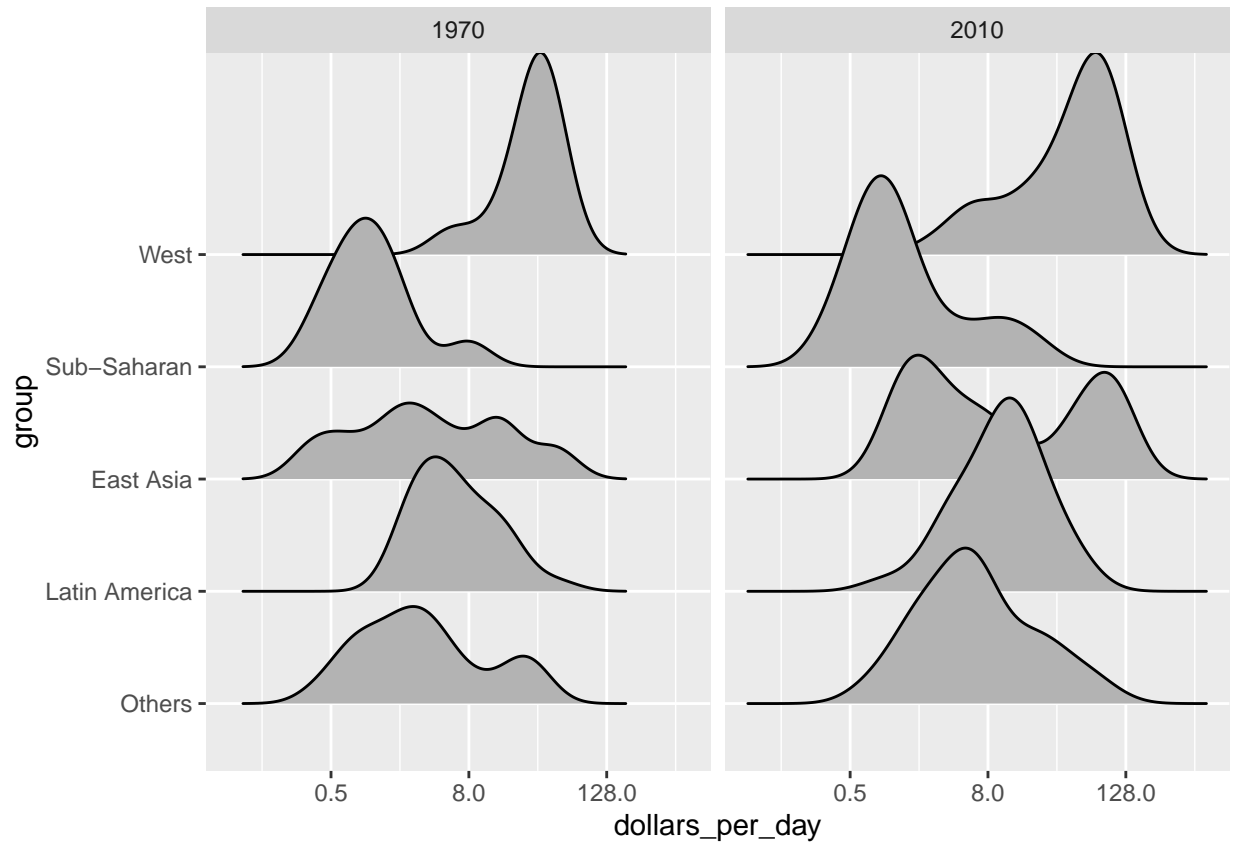
ridges ,

```
gapminder %>%
  filter(year %in% years & !is.na(dollars_per_day)) %>%
  ggplot(aes(dollars_per_day, group)) +
  scale_x_continuous(trans="log2") +
  geom_density_ridges(adjust=1.5) + #adjust
  facet_grid(~year)
```

```
## Warning: Ignoring unknown parameters: adjust
```

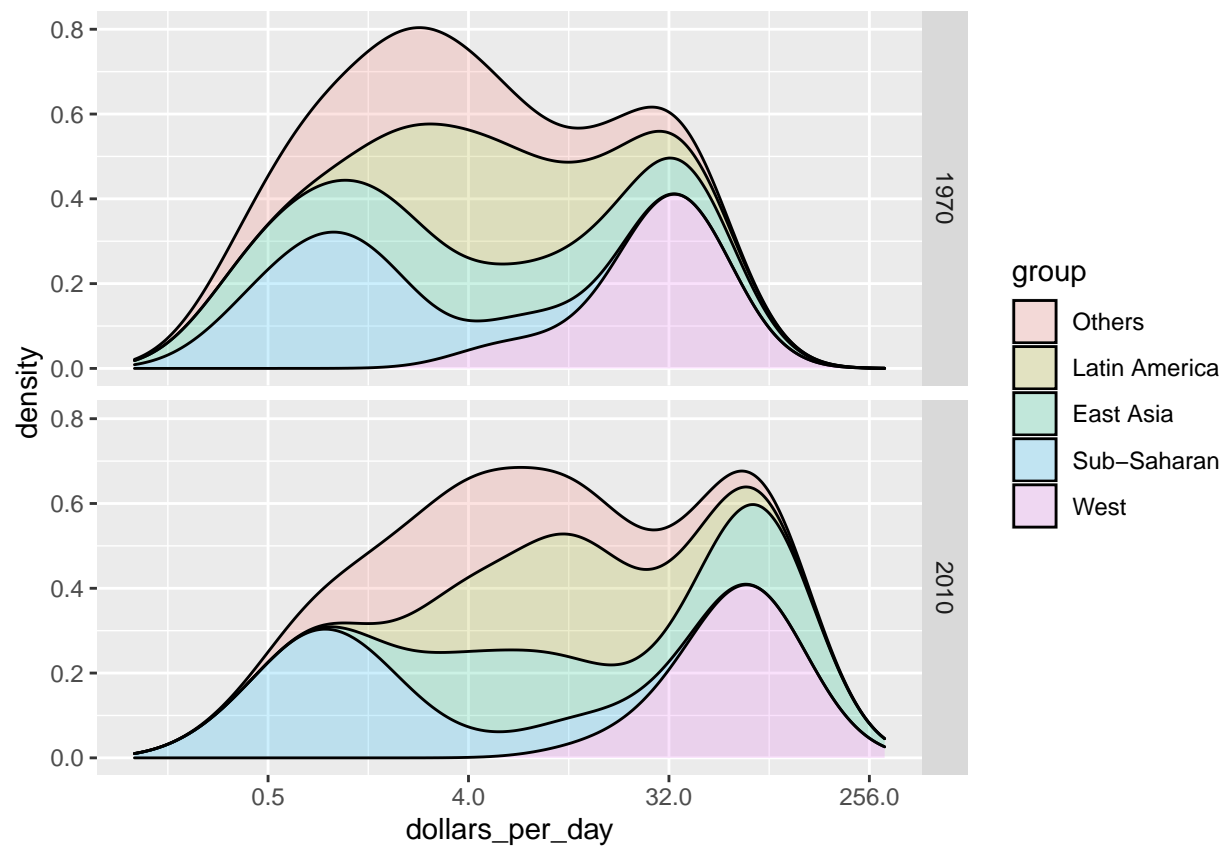
```
## Picking joint bandwidth of 0.648
```

```
## Picking joint bandwidth of 0.726
```



Stacking

```
gapminder %>%
  filter(year %in% years & country %in% country_list) %>%
  group_by(year) %>%#
  mutate(weight=population/sum(population)*2) %>%#
  ungroup() %>%#
  ggplot(aes(dollars_per_day,fill=group))+
  scale_x_continuous(trans="log2",limit=c(0.125,300))+
  geom_density(alpha=0.2,bw=0.75,position="stack")+
  facet_grid(year~.)
```



r-graph-gallery.com