data 550

Lili Tang

2023-01-12

Quesion 1

Most important message conveyed in the video is that how statistical data provides insights about the world which contradict common sense(the world is not as poor or as divided as many people think) and how important to get boring statistics into graphic formats where people can instantly understand them

Use of animation in the Gapminder tool(animated bubble chart) is most effective, which dynamically shows how change happened in income between western world and third world, and between the different countries in each region over time.The animation also helps to convey the message that the traditional divide between "developed" and "developing" countries is becoming increasingly irrelevant.

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(dplyr)
library(Hmisc)
```

```
## Loading required package: lattice
## Loading required package: survival
## Loading required package: Formula
##
## Attaching package: 'Hmisc'
##
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
##
```

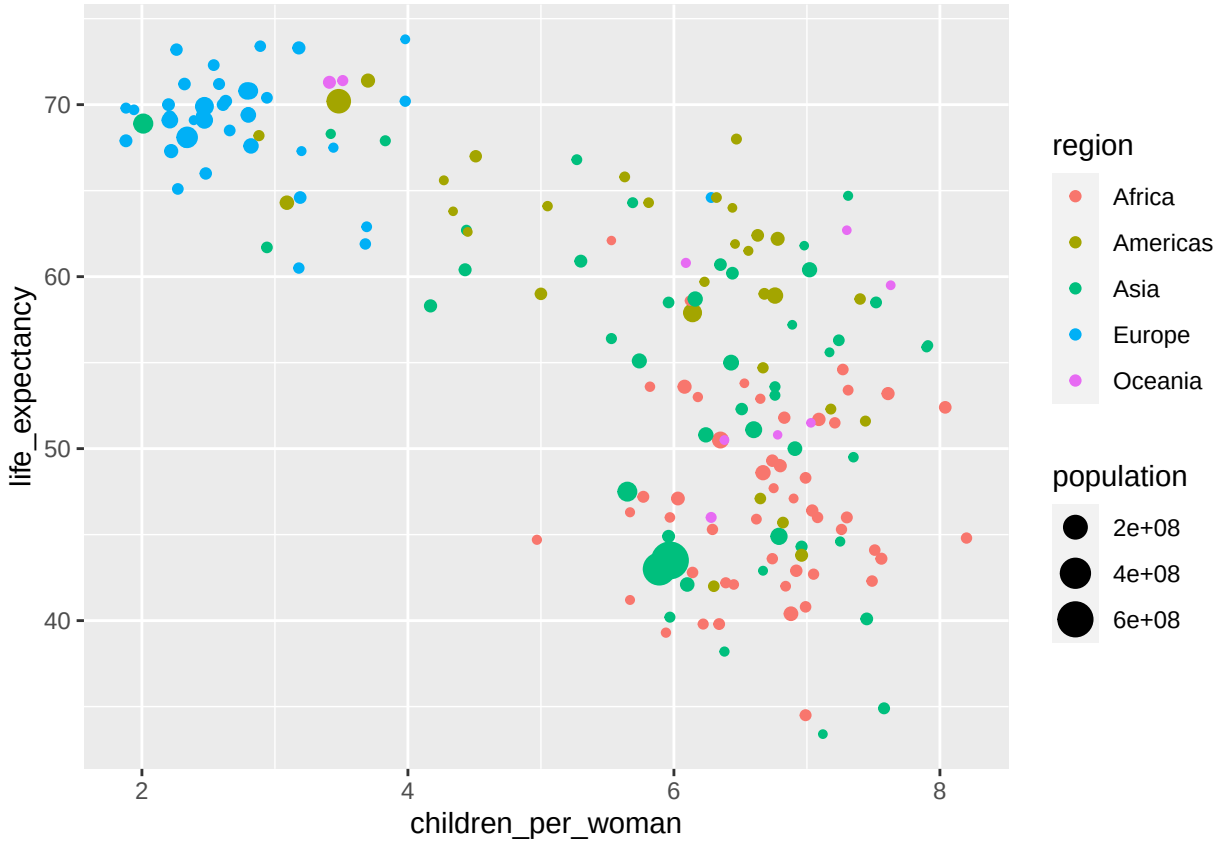
```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

Question 2.2

```
df<-read.csv("https://raw.githubusercontent.com/UofTCoders/workshops-dc-py/master/data/processed/world-
```

```
df_1962 <- subset(df, df$year == 1962)
```

```
df_1962 %>%
  ggplot(aes(
    x=children_per_woman,
    y=life_expectancy,
    color=region,
    size=population))+
  geom_point()
```



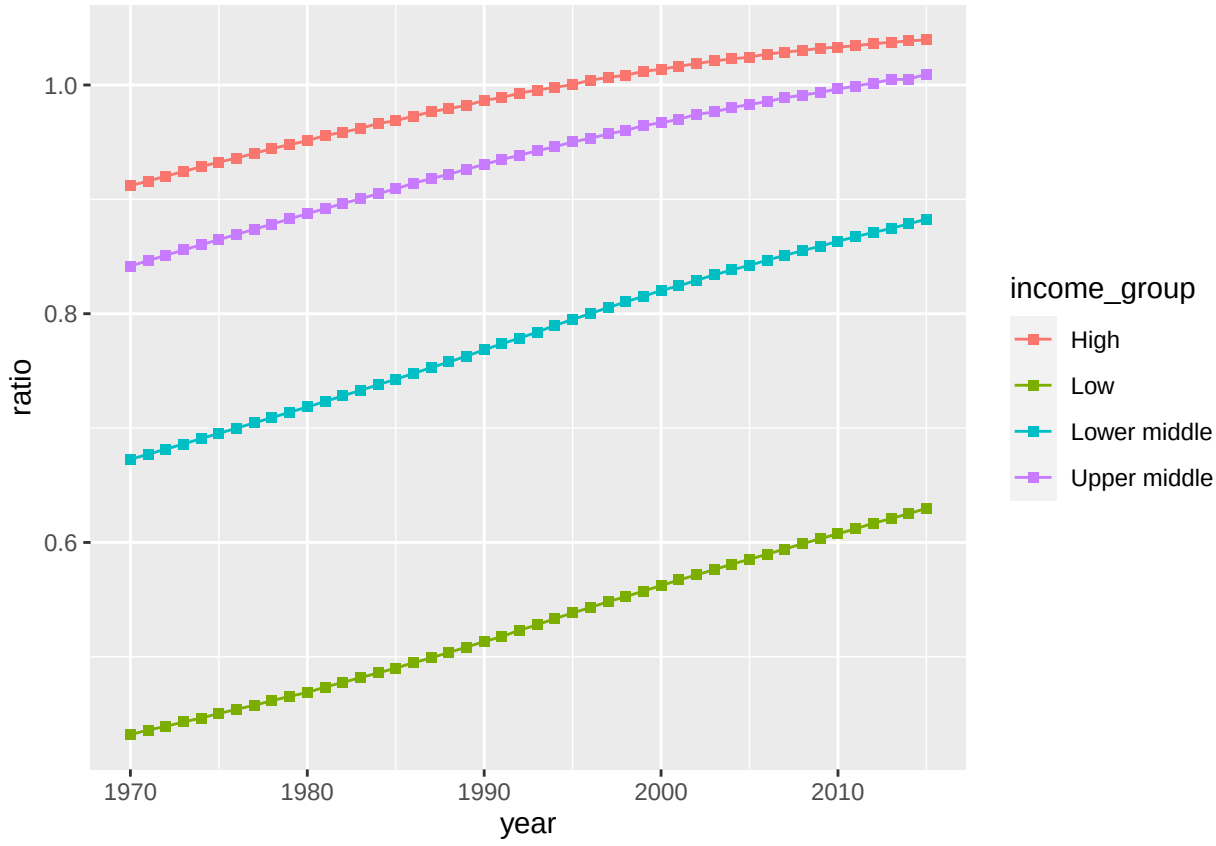
Question 3.2

```
df$ratio  <- df$years_in_school_women/df$years_in_school_men
```

```
df_1970_2015<- subset(df, df$year >=1970 & df$year <=2015 )
```

```
df_1970_2015 %>%
```

```
  ggplot(aes(
    x=year,
    y=ratio,
    group = income_group,
    color=income_group))+
  stat_summary(geom="line", fun=mean)+
  stat_summary(geom="point", fun=mean, shape="square")
```

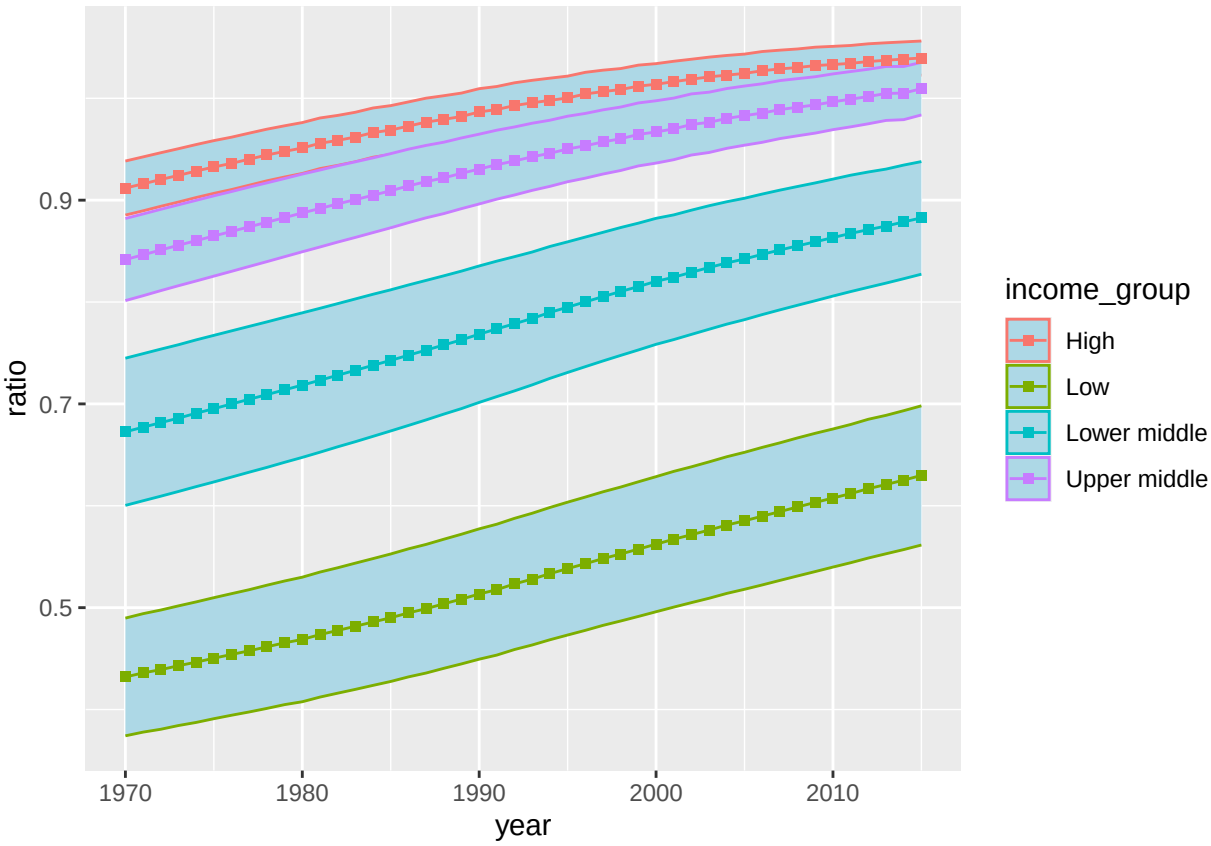


Question 3.4

```
df$ratio  <- df$years_in_school_women/df$years_in_school_men

df_1970_2015<- subset(df, df$year >=1970 & df$year <=2015 )

df_1970_2015 %>%
  ggplot(aes(
    x=year,
    y=ratio,
    group = income_group,
    color=income_group))+
  stat_summary(geom="ribbon", fun.data=mean_cl_normal,
               fun.args=list(conf.int=0.95),fill="lightblue")+
  stat_summary(geom="line", fun=mean)+
  stat_summary(geom="point", fun=mean, shape="square")
```



Question 4.2

```
df_filtered <- subset(df, year %in% c(1918, 1938, 1958, 1978, 1998, 2018) & child_mortality!="NA")


p <- ggplot(df_filtered,aes(
    x=children_per_woman,
    y=child_mortality,
    color=income_group))+
  geom_point(shape=1)+
  facet_wrap(~year,ncol=3,nrow=2)
p
```


```
ggsave("my_plot.png", plot = p, width = 6, height = 4)
```

Conclusion:
1. People in high-income group tend to have smaller family sizes, at the same time with lower child mortality;
2. People in the low-income group tend to have larger family sizes, at the same time with larger child mortality;
3. As time goes by, the overall child mortality largely decreases, although low-income group still has larger family size and higher mortality than high-income group, but the gap becomes much smaller.

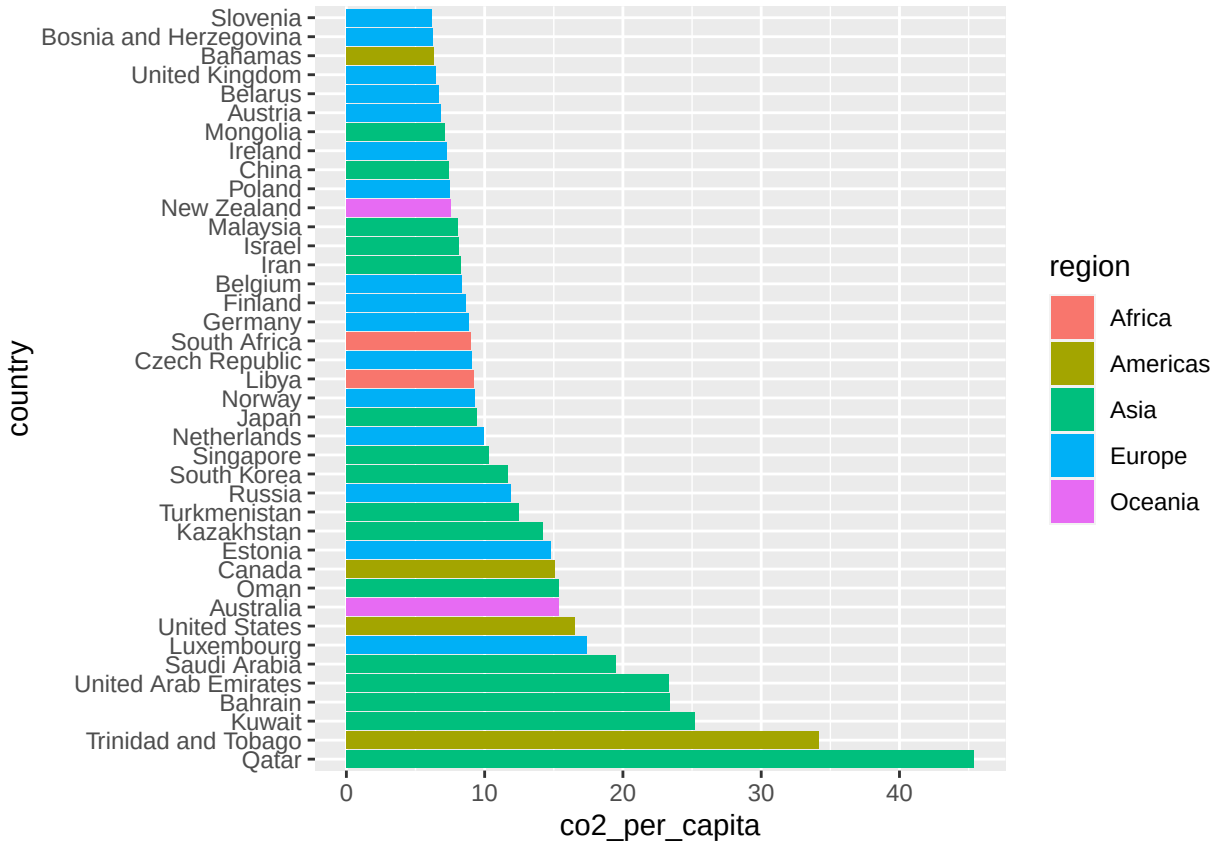
Question 5.2

```
df1 <- subset(df,df$co2_per_capita!="NA")
max(df1$year)
```

```
## [1] 2014
```

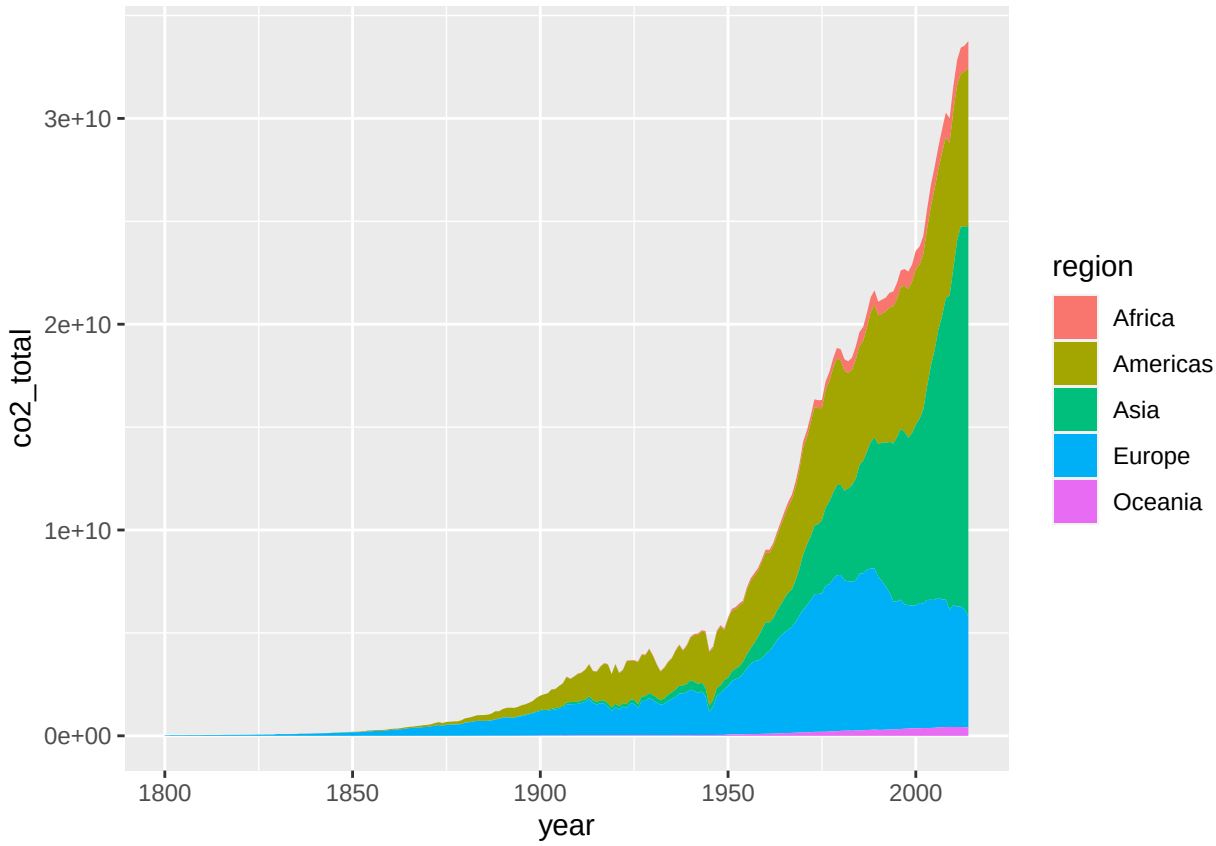
```
df_2014 <- subset(df,df$year==2014)

df_2014 %>% slice_max(co2_per_capita, n = 40) %>%
  ggplot(aes(
    x=co2_per_capita,
    y=reorder(country,desc(co2_per_capita)),
    fill=region))+
  geom_bar(stat="identity")+
  labs( y = "country")
```



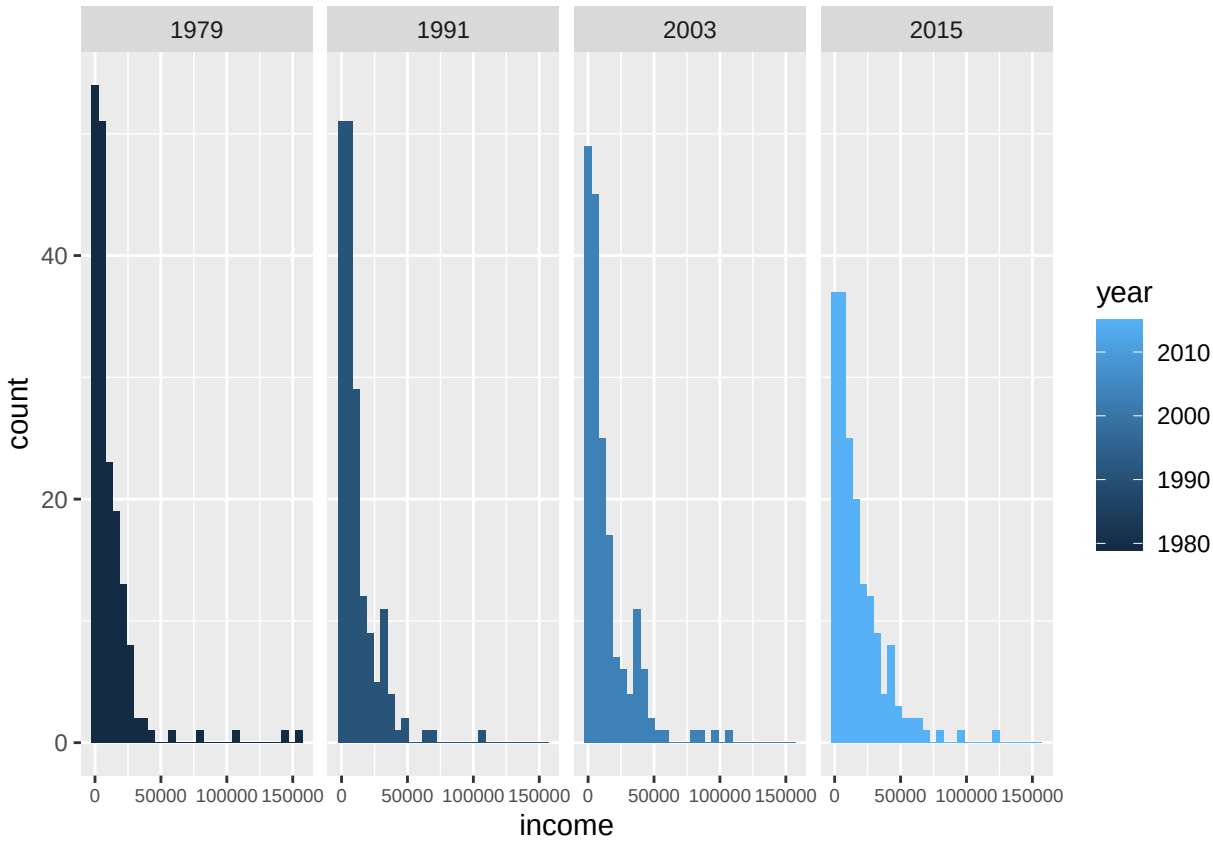
Question 5.4

```
df1$co2_total <- df1$co2_per_capita* df1$population
df1 %>%
  ggplot(aes(
    x=year,
    y=co2_total,
    group=region,
    fill=region))+
  geom_area(stat="summary",fun=sum,position='stack')
```



Question 6.2

```
df_filtered <- subset(df, year %in% c(1979, 1991, 2003,2015))
df_filtered %>%
  ggplot(aes(
    x=income,
    fill=year))+
  geom_histogram(bins = 30)+
  facet_wrap(~year,ncol=4)+
  theme(axis.text.x = element_text(size = 7))
```

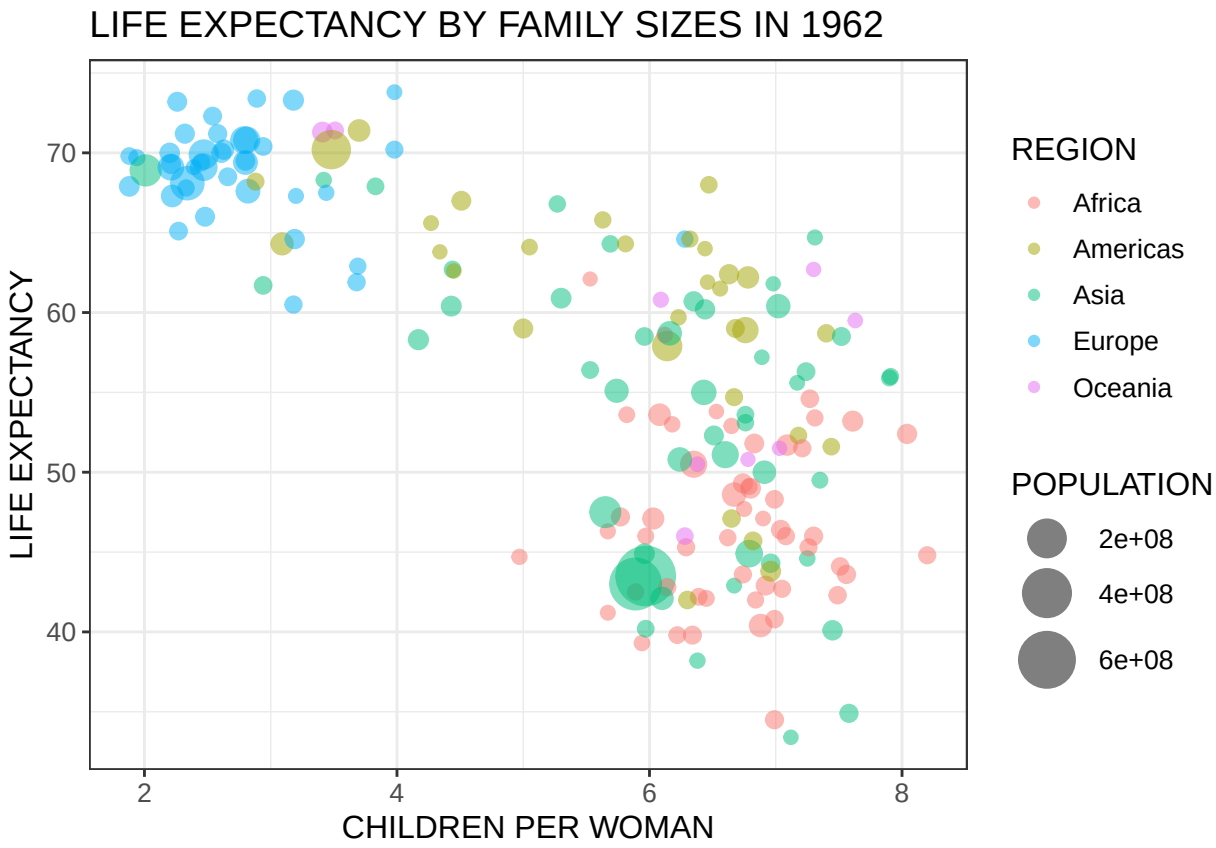


Conclusion:I think the trend is the same as the Rosling's projection, that is income distribution in 2015 would be much more evenly distributed than it was in the past and global income distribution was becoming more similar across countries.

The opinion is true as the above chart shows that the number of lower income decreases over time(from 1980-2015)

Question 7.2

```
df_1962 %>%
  ggplot(aes(
    x=children_per_woman,
    y=life_expectancy,
    color=region,
    size=population))+
  geom_point(alpha = 0.5)+
  ggtitle("LIFE EXPECTANCY BY FAMILY SIZES IN 1962")+
  labs(x = "CHILDREN PER WOMAN", y = "LIFE EXPECTANCY",color="REGION", size="POPULATION")+
  theme_bw()+
  theme(text = element_text(size = 12))+
  scale_size(range = c(2, 10))
```


LIFE EXPECTANCY BY FAMILY SIZES IN 1962
LIFE EXPECTANCY
70
60
50
40
2
4
6
8
CHILDREN PER WOMAN
REGION
Africa
Americas
Asia
Europe
Oceania
POPULATION
2e+08
4e+08
6e+08