Midterm project

Shuoqi Huang

10/18/2019

Introduction

The rapid development of human society brings a lot of benefits. The qualities of people's life increase dramatically due to these changes. However, some people say that human activities destroy the ecological system. To check this idea, I explode the data of forest area, agriculture land, and co2 emissions from the top three largest economies in the world.

```
##Check NA in data
sum(is.na(wb_dat))
```

```
## [1] 0
```

```
## The number of NA is 0 in my data.
```

```
## Data clean part
wb_countries <- wbcountries()
names(wb_countries)
```

```
##  [1] "iso3c"         "iso2c"         "country"       "capital"
##  [5] "long"          "lat"           "regionID"      "region_iso2c"
##  [9] "region"        "adminID"       "admin_iso2c"   "admin"
## [13] "incomeID"      "income_iso2c"  "income"        "lendingID"
## [17] "lending_iso2c" "lending"
```

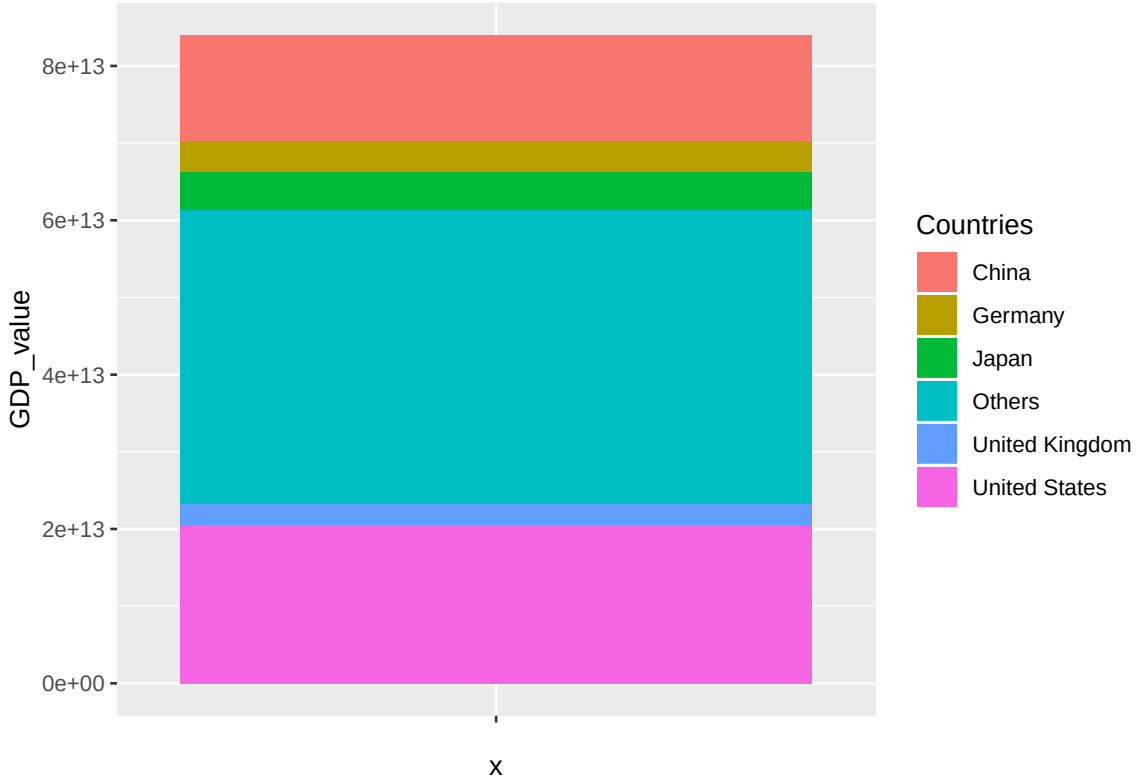
```
wb_dat <- merge(wb_dat, y = wb_countries[c("iso2c", "region")], by = "iso2c", all.x = TRUE)
wb_dat <- subset(wb_dat, region != "Aggregates")

wb_dat$indicatorID[wb_dat$indicatorID == "AG.LND.FRST.K2"] <- "Forest_area"
wb_dat$indicatorID[wb_dat$indicatorID == "AG.LND.FRST.ZS"] <- "Forest_area_%"
wb_dat$indicatorID[wb_dat$indicatorID == "AG.LND.TOTL.K2"] <- "Land area_sqkm"
wb_dat$indicatorID[wb_dat$indicatorID == "AG.LND.AGRI.ZS"] <- "Agricultural_land_%"
wb_dat$indicatorID[wb_dat$indicatorID == "EN.ATM.CO2E.PC"] <- "CO2 emissions"
wb_dat$indicatorID[wb_dat$indicatorID == "SP.POP.GROW"] <- "population_growth"
wb_dat$indicatorID[wb_dat$indicatorID == "NY.GDP.MKTP.CD"] <- "GDP_$"
```

Methods:

Comparing the latest GDP of countries and select the top three largest economies of the world.

```
## Create dataframe
GDP_Country <- wb_dat %>%
  filter(indicatorID == "GDP_$", date == "2018")
GDP_Country <- GDP_Country[order(GDP_Country$value, decreasing = TRUE),]
GDP_US <- GDP_Country[1,4]
GDP_CN <- GDP_Country[2,4]
GDP_JP <- GDP_Country[3,4]
GDP_DE <- GDP_Country[4,4]
```

```
GDP_GB <- GDP_Country[5,4]
GDP_OTHERS <- sum(GDP_Country$value) - sum(GDP_US,GDP_CN,GDP_JP,GDP_DE,GDP_GB)

GDP_value <- c(GDP_US,GDP_CN,GDP_JP,GDP_DE,GDP_GB,GDP_OTHERS)
Countries <- c('United States', 'China', 'Japan', 'Germany', 'United Kingdom', 'Others')
data1 <- data.frame(Countries= Countries, GDP_value = GDP_value)

## Plot a bar chart about the GDP .
p <- ggplot(data = data1, aes(x = '' , y = GDP_value, fill = Countries)) +
  geom_bar(stat = 'identity', position = 'stack', width = 1)
p
```



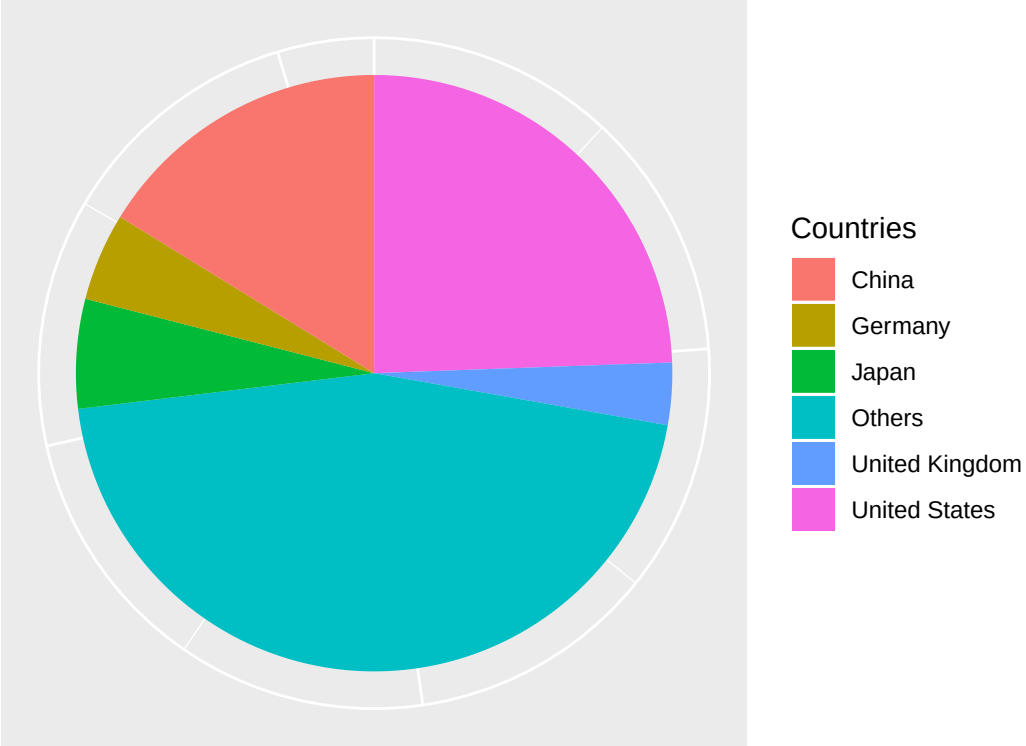

```
## Plot a pie chart about the gdp.
label_value <- paste('(', round(data1$GDP_value/sum(data1$GDP_value) * 100, 1), '%)',
                sep = '')
label_value
```

```
## [1] "(24.4%)" "(16.2%)" "(5.9%)"  "(4.8%)"  "(3.4%)"  "(45.3%)"
```

```
label <- paste(data1$Countries, label_value, sep = '')
label
```

```
## [1] "United States(24.4%)" "China(16.2%)"         "Japan(5.9%)"
## [4] "Germany(4.8%)"        "United Kingdom(3.4%)" "Others(45.3%)"
```

```
p + coord_polar(theta = 'y') + labs(x = '', y = '', title = '') +
  theme(axis.text = element_blank()) +
  theme(axis.ticks = element_blank())
```



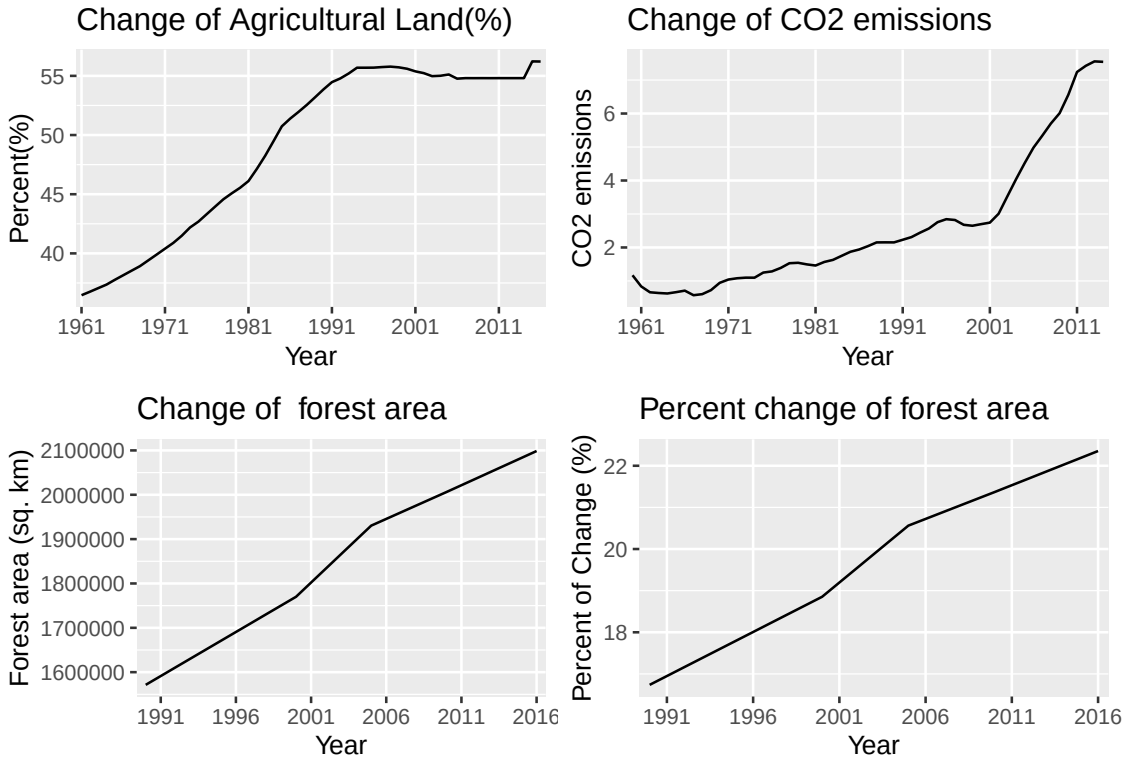
We can see that the United States, China, and Japan are the three countries with higher GDP.

Trend of agricultural Land, CO2 emissions, forest area in three countries:

China

```
## select data about percent of agricultural land in China.
China_Agricultural <- wb_dat %>%
  filter(country == "China",indicatorID == "Agricultural_land_%")
## make a plot of change of agricultural land from 1961 to 2016.
a1 = ggplot(data = China_Agricultural, aes(x = date, y =value, group = 1)) +
  geom_line() + scale_x_discrete("Year",breaks = seq(1961,2016,10))+
  labs(y = "Percent(%)", title = "Change of Agricultural Land(%)")

## select data about co2 emissions in China.
China_CO2_emissions <- wb_dat %>% filter(country == "China",
  indicatorID == "CO2 emissions" )
## make a plot of change of co2 emissions from 1961 to 2016.
a2 = ggplot(data = China_CO2_emissions, aes(x = date, y =value, group = 1)) +
  geom_line() + scale_x_discrete("Year",breaks = seq(1961,2016,10))+
  labs(y = "CO2 emissions",
       title = "Change of CO2 emissions")

## select data about forest area in China.
China_forest <- wb_dat %>% filter(country == "China",
  indicatorID == "Forest_area" )
## make a plot of change of forest area from 1990 to 2016.
a3 = ggplot(data = China_forest, aes(x = date, y =value, group = 1)) +
```

```
  geom_line() + scale_x_discrete("Year",breaks = seq(1991,2016,5))+
  labs(y = "Forest area (sq. km)", title = "Change of  forest area")


## select data about percent of forest area in China.
China_forest_percent <- wb_dat %>% filter(country == "China",
  indicatorID == "Forest_area_%" )
## make a plot about change of percent of forest area from 1990 to 2016.
a4 = ggplot(data = China_forest_percent, aes(x = date, y =value, group = 1)) +
  geom_line() + scale_x_discrete("Year",breaks = seq(1991,2016,5))+
  labs(y = "Percent of Change (%)", title = "Percent change of forest area")
grid.arrange(a1,a2,a3,a4,ncol=2)
```

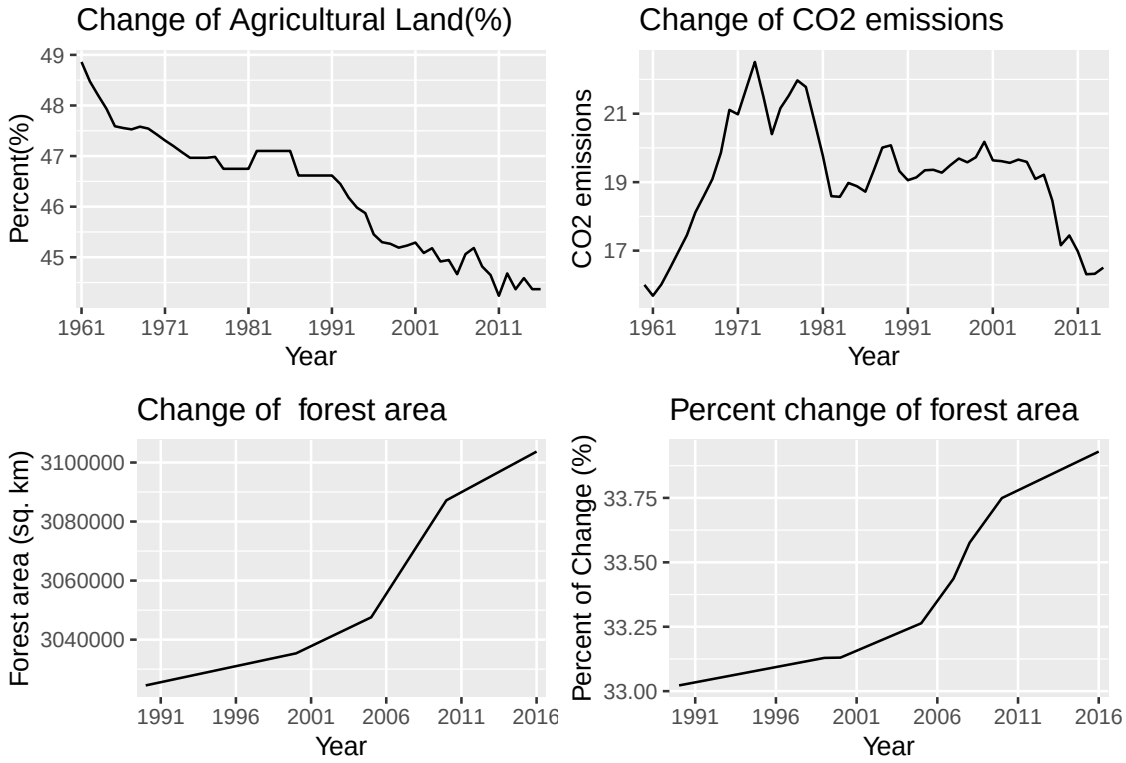


We can see that agricultural land in China increased rapidly from 1961 to 1996 and kept steady from 1996 to 2016. The CO2 emissions metric tons per capita increased a lot. However, the forest in China increased steadily from 1991 to 2016.

The United States

```
## select data about percent of agricultural land in the United States.
US_Agricultural <- wb_dat %>%
  filter(country == "United States",indicatorID == "Agricultural_land_%")
## make a plot of change of agricultural land from 1961 to 2016.
b1 = ggplot(data = US_Agricultural, aes(x = date, y =value, group = 1)) +
  geom_line() + scale_x_discrete("Year",breaks = seq(1961,2016,10))+
  labs(y = "Percent(%)", title = "Change of Agricultural Land(%)")


## select data about co2 emissions in the United States.
```

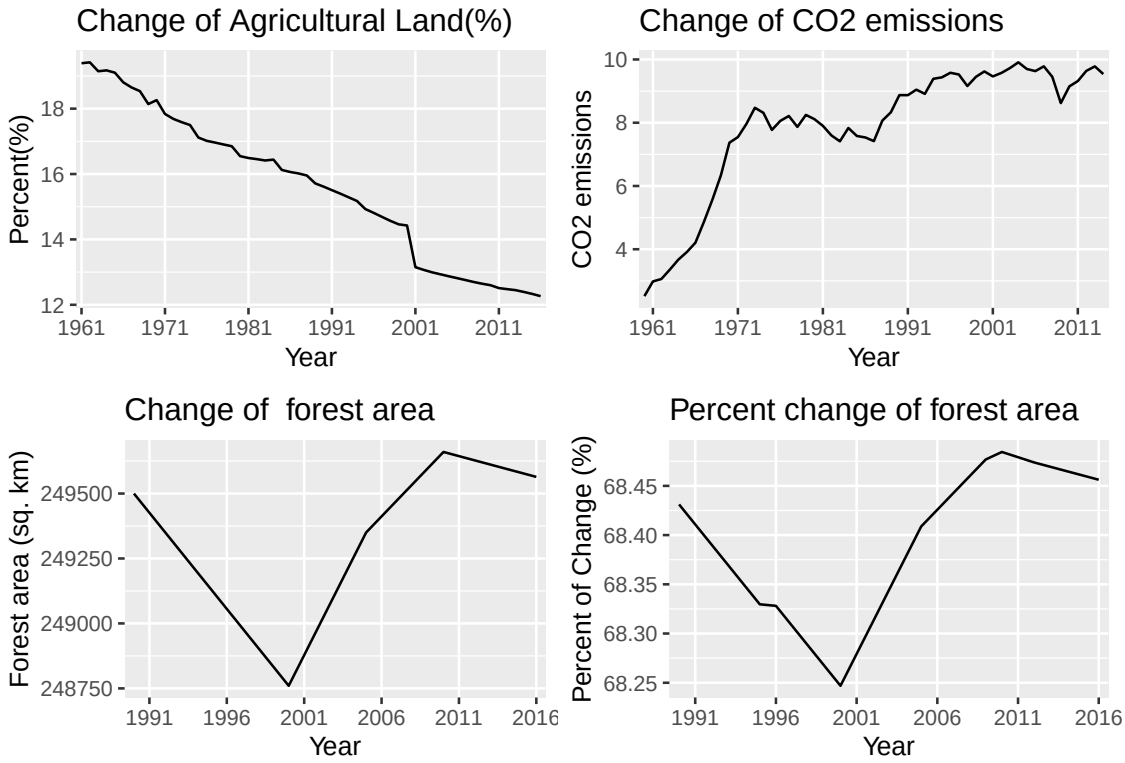
```r
US_CO2_emissions <- wb_dat %>% filter(country == "United States",
  indicatorID == "CO2 emissions" )
## make a plot of change of co2 emissions from 1960 to 2014.
b2 = ggplot(data = US_CO2_emissions, aes(x = date, y =value, group = 1)) +
  geom_line() + scale_x_discrete("Year",breaks = seq(1961,2016,10))+
  labs(y = "CO2 emissions",
       title = "Change of CO2 emissions")

## select data about forest area in the United States.
US_forest <- wb_dat %>% filter(country == "United States",
  indicatorID == "Forest_area" )
## make a plot of change of forest area from 1990 to 2016.
b3 = ggplot(data = US_forest, aes(x = date, y =value, group = 1)) +
  geom_line() + scale_x_discrete("Year",breaks = seq(1991,2016,5))+
  labs(y = "Forest area (sq. km)", title = "Change of  forest area")

## select data about percent of forest area in the United States.
US_forest_percent <- wb_dat %>% filter(country == "United States",
  indicatorID == "Forest_area_%" )
## make a plot about change of percent of forest area from 1990 to 2016.
b4 = ggplot(data = US_forest_percent, aes(x = date, y =value, group = 1)) +
  geom_line() + scale_x_discrete("Year",breaks = seq(1991,2016,5))+
  labs(y = "Percent of Change (%)", title = "Percent change of forest area")
grid.arrange(b1,b2,b3,b4,ncol=2)
```



The agricultural land in the United States decreased from 1961 to 2016 but the differences were not so huge. The CO2 emissions metric tons per capita increased from 1961 to 1973 and then began to decrease slowly. Moreover, the forest area in the United States also increased.

Japan

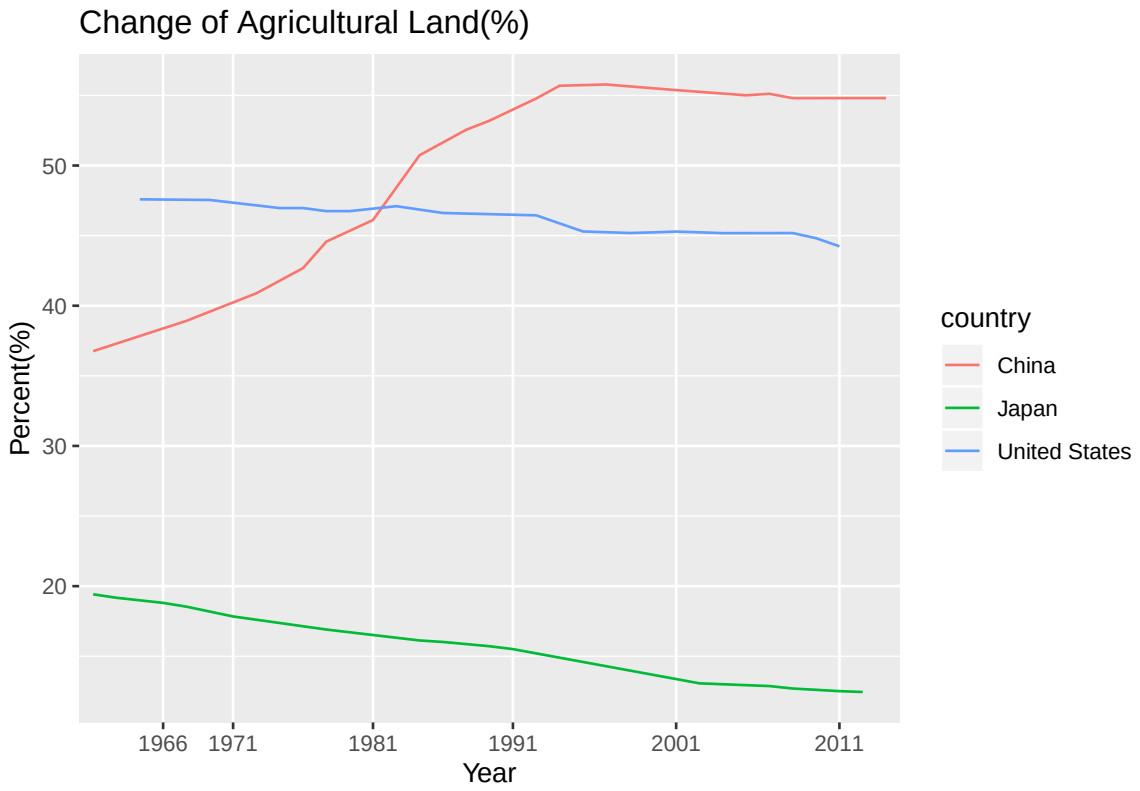
```
## select data about percent of agricultural land in Japan.
JP_Agricultural <- wb_dat %>%
  filter(country == "Japan",indicatorID == "Agricultural_land_%")
## make a plot of change of agricultural land from 1961 to 2016.
c1 = ggplot(data = JP_Agricultural, aes(x = date, y =value, group = 1)) +
  geom_line() + scale_x_discrete("Year",breaks = seq(1961,2016,10))+
  labs(y = "Percent(%)", title = "Change of Agricultural Land(%)")

## select data about co2 emissions in Japan.
JP_CO2_emissions <- wb_dat %>% filter(country == "Japan",
  indicatorID == "CO2 emissions" )
## make a plot of change of co2 emissions from 1960 to 2014.
c2 = ggplot(data = JP_CO2_emissions, aes(x = date, y =value, group = 1)) +
  geom_line() + scale_x_discrete("Year",breaks = seq(1961,2016,10))+
  labs(y = "CO2 emissions",
       title = "Change of CO2 emissions")

## select data about forest area in the Japan.
JP_forest <- wb_dat %>% filter(country == "Japan",
  indicatorID == "Forest_area" )
## make a plot of change of forest area from 1990 to 2016.
c3 = ggplot(data = JP_forest, aes(x = date, y =value, group = 1)) +
  geom_line() + scale_x_discrete("Year",breaks = seq(1991,2016,5))+
  labs(y = "Forest area (sq. km)", title = "Change of  forest area")

## select data about percent of forest area in Japan.
JP_forest_percent <- wb_dat %>% filter(country == "Japan",
  indicatorID == "Forest_area_%" )
## make a plot about change of percent of forest area from 1990 to 2016.
c4 = ggplot(data = JP_forest_percent, aes(x = date, y =value, group = 1)) +
  geom_line() + scale_x_discrete("Year",breaks = seq(1991,2016,5))+
  labs(y = "Percent of Change (%)", title = "Percent change of forest area")
grid.arrange(c1,c2,c3,c4,ncol=2)
```



The percent of agricultural land in Japan decreased dramatically from 1961 to 2016. The CO2 emissions metric tons per capita in Japan also increased from 1961 to 2016. The change in forest land in Japan was not huge.

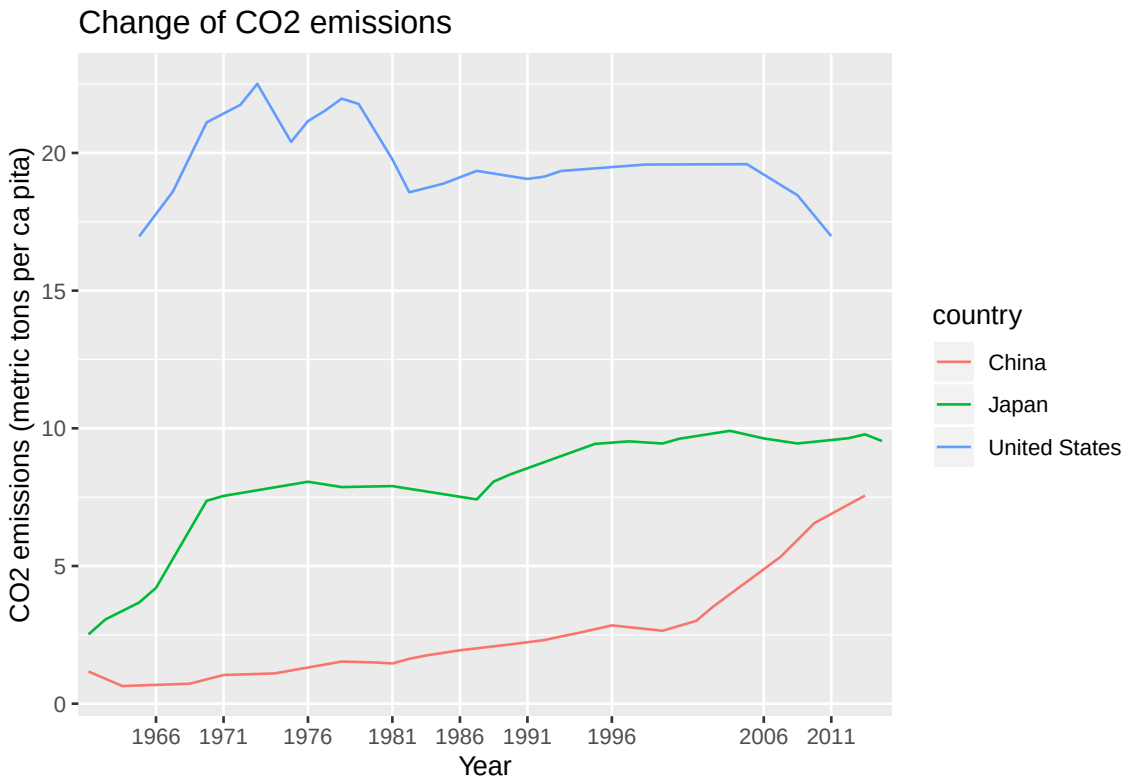
Comparing three countries with three categories.

```
## select data about percent of agricultural land from these three countries.
al_dat = wb_dat %>% filter(indicatorID == "Agricultural_land_%" &
                country == c("China","Japan","United States"))
## Plot a graph that compares percent changes of agricultural land among these three countries.
ggplot(al_dat) + geom_line(aes(y=value,x = date, group = country,color = country)) +
  scale_x_discrete("Year",breaks = seq(1961,2016,5))+
labs(y = "Percent(%)", title = "Change of Agricultural Land(%)")
```



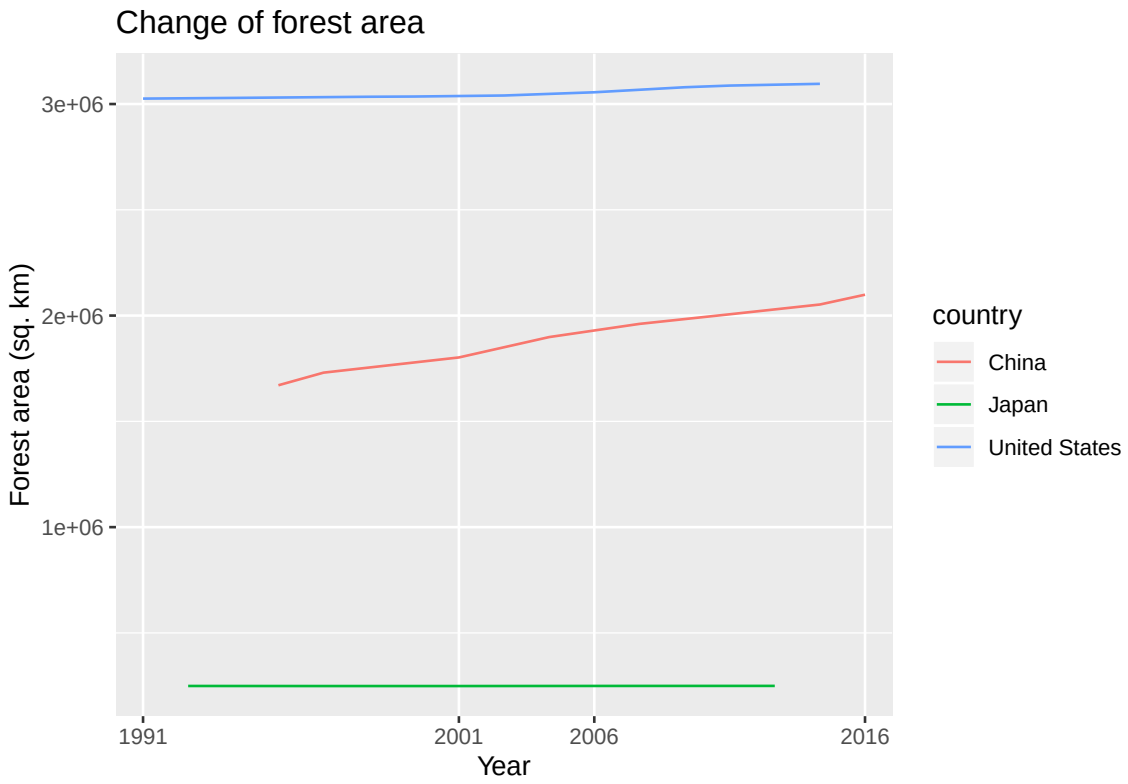
China has the largest percent of agricultural land among these countries.

```
## select data about co2 emissions from these three countries.
co2_datt = wb_dat %>% filter(indicatorID == "CO2 emissions" &
               country == c("China","Japan","United States"))
## Plot a graph that compares changes of co2 emissions among these three countries.
ggplot(co2_datt) + geom_line(aes(y=value,x = date, group = country,color = country)) +
  scale_x_discrete("Year",breaks = seq(1961,2016,5))+
labs(y = "CO2 emissions (metric tons per ca pita)",
     title = "Change of CO2 emissions")
```



We can see that the United States has the highest CO2 emissions metric tons per capita.

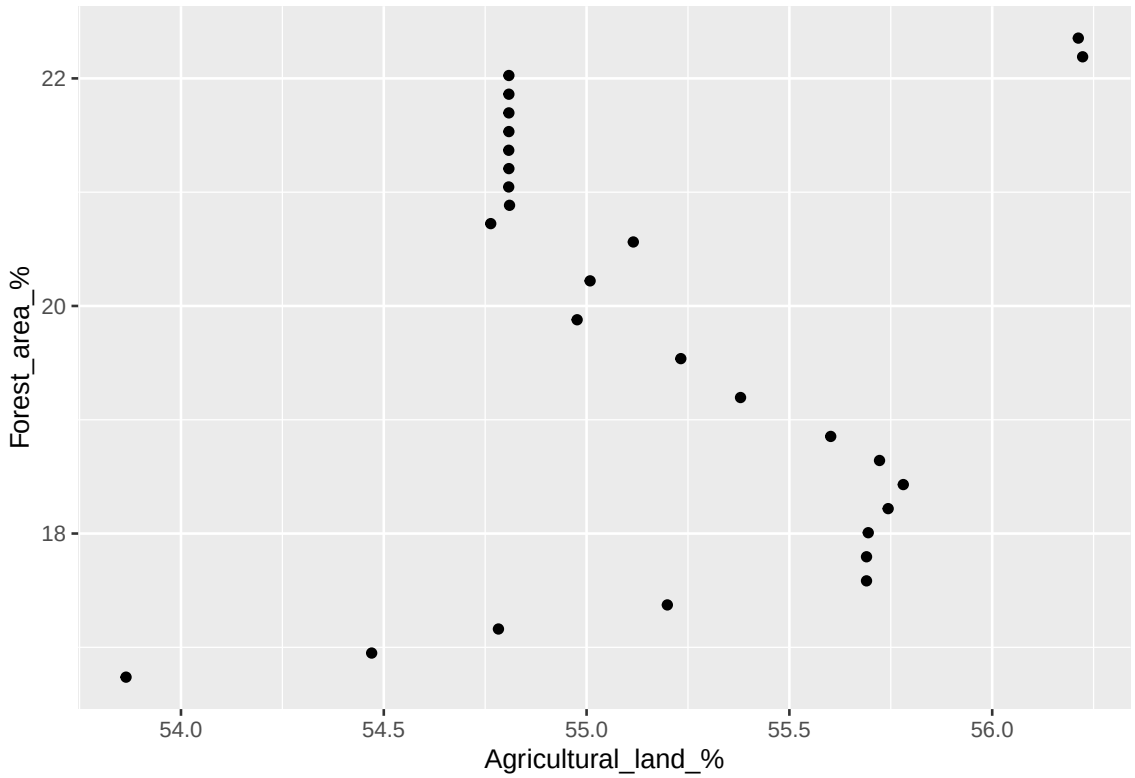
```
## select data about forest area from these three countries.
FA_datt = wb_dat %>% filter(indicatorID == "Forest_area" &
               country == c("China","Japan","United States"))
## Plot a graph that compares changes of forest area among these three countries.
ggplot(FA_datt) + geom_line(aes(y=value,x = date, group = country,color = country)) +
  scale_x_discrete("Year",breaks = seq(1991,2016,5))+
labs(y = "Forest area (sq. km)",
     title = "Change of forest area")
```



The United States has the largest forest area. The slope of forest area in China is the steepest.

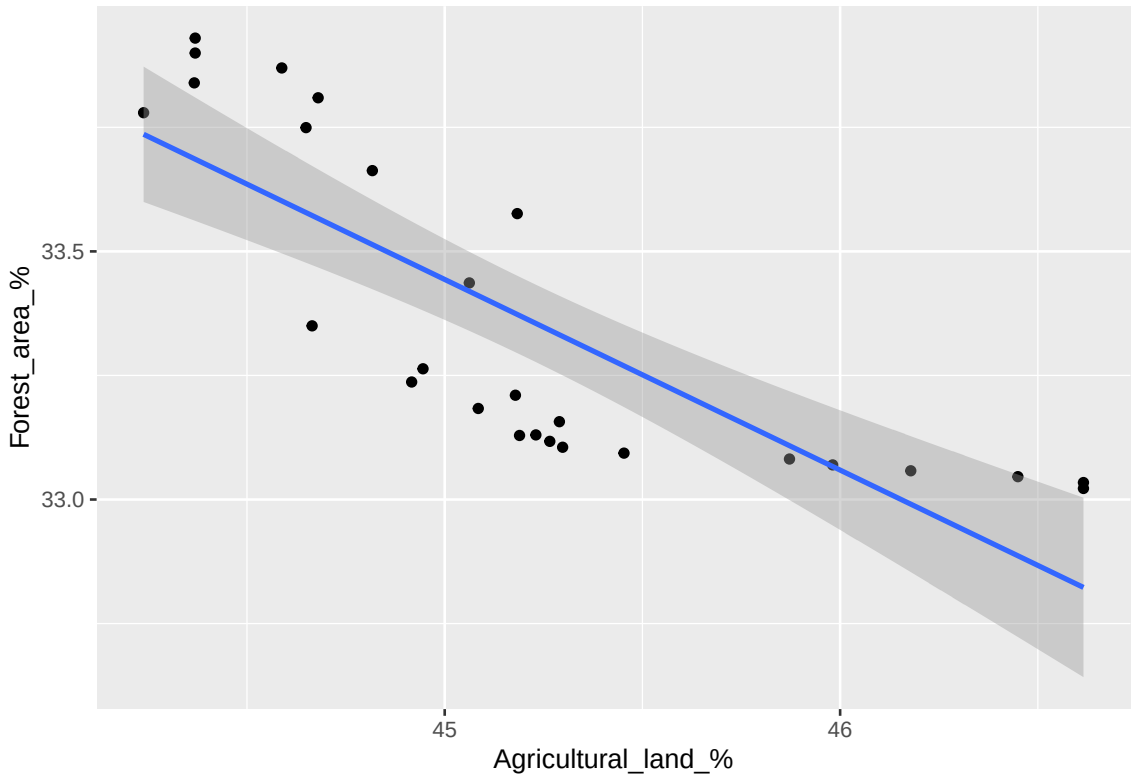
Relationship between percent of agricultural land and percent of forest land.

```
## Create table that contains percent of agricultural land and forest area in China after 1989.

relation_data1 <- wb_dat %>% filter((indicatorID == "Forest_area_%"|indicatorID ==
                "Agricultural_land_%") &
                country == "China")

select_rel <- relation_data1 %>% select(date,value,indicatorID,country)
data_need <- select_rel %>% pivot_wider(names_from = indicatorID,
            values_from = value)
data_1990 <- data_need %>% filter(data_need$date > 1989)

## Plot the graph about relationship between percent of agricultural land and forest area.
ggplot(data = data_1990, aes(x = `Agricultural_land_%`, y = `Forest_area_%`)) +
  geom_point()
```



We can see that there is no obvious relationship between percent of agricultural land and forest area.

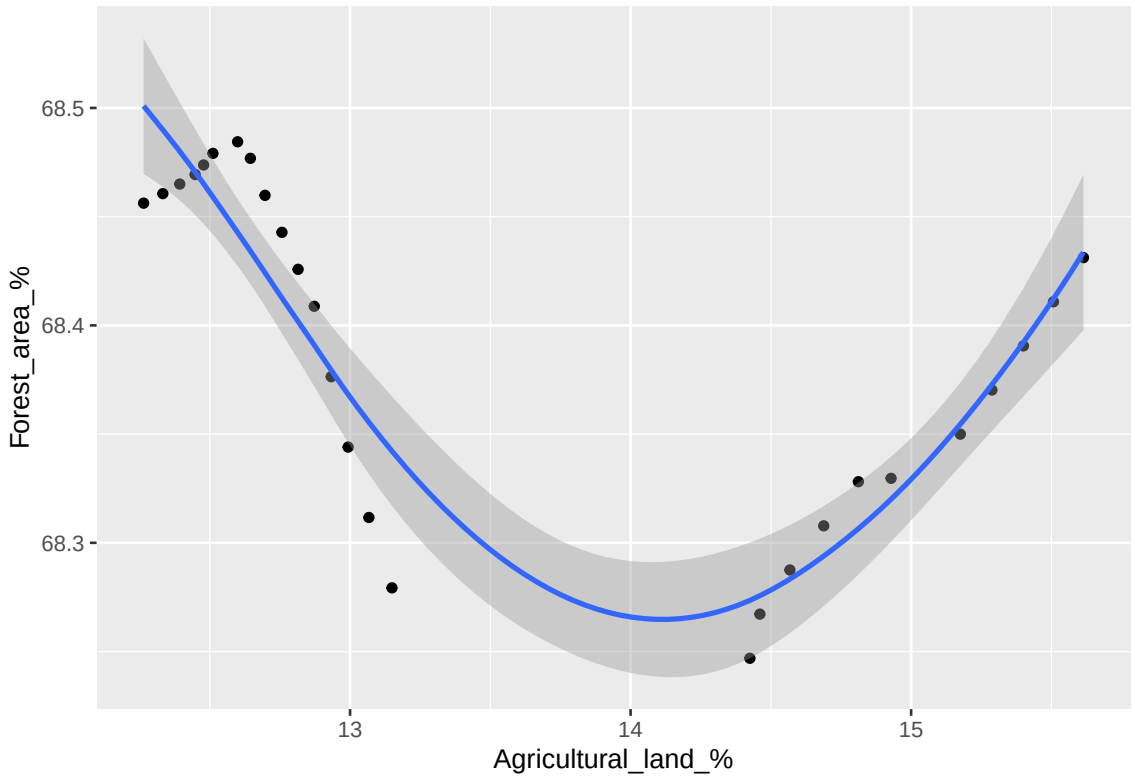
```
## Create table that contains percent of agricultural land and forest area in the United States after 1
relation_data2 <- wb_dat %>% filter((indicatorID == "Forest_area_%"|
                indicatorID == "Agricultural_land_%") &
                country == "United States")

select_rel1 <- relation_data2 %>% select(date,value,indicatorID,country)
data_need1 <- select_rel1 %>% pivot_wider(names_from = indicatorID,values_from = value)
data_1990_1 <- data_need1 %>% filter(data_need1$date > 1989)

## Plot the graph about relationship between percent of agricultural land and forest area.
ggplot(data = data_1990_1, aes(x = `Agricultural_land_%`, y = `Forest_area_%`)) +
  geom_point() + geom_smooth(method = "lm")
```



From this plot, we can see that the percent of forest land decreases when percent of agricultural land increases. Therefore, this is a negative linear relationship.

```
## Create table that contains percent of agricultural land and forest area in Japan after 1989.
relation_data3 <- wb_dat %>% filter((indicatorID == "Forest_area_%"|
                indicatorID == "Agricultural_land_%") &
                country == "Japan")

select_rel2 <- relation_data3 %>% select(date,value,indicatorID,country)
data_need2 <- select_rel2 %>% pivot_wider(names_from = indicatorID,values_from = value)
data_1990_2 <- data_need2 %>% filter(data_need2$date > 1989)

## Plot the graph about relationship between percent of agricultural land and forest area.
ggplot(data = data_1990_2, aes(x = `Agricultural_land_%`, y = `Forest_area_%`)) +
  geom_point() + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



We can see that there is certain pattern between percent of agricultural land and forest area. Perhaps, these two variables have a non-linear relationship.

Conclusion:

We know that situation differs in every country. In the United States, increase of percent of agriculture land causes decrease the percent of forest area. This evidence shows that human activities affect the ecological system. However, the relationship is not significant or hard to tell from the previous analyses. We should look more factors that relate to change of forest area.