Group_27

Stage 1 Source an interesting dataset

1. Import data and packages

```
library(dplyr)
library(ggplot2)
#Annual articles published in scientific and technical journals per million people
article_per_million_people <- read.csv('https://ourworldindata.org/grapher/scientific-publ
#Data source: World Bank (2023); United Nations (2022)

#Annual patent applications per million people
patent_per_million_people <- read.csv('https://ourworldindata.org/grapher/patent-applicati
#Data source: World Bank (2023); United Nations (2022)

#Number of Research & Development researchers per million people
researcher_per_million_people <- read.csv('https://ourworldindata.org/grapher/researchers-
#Data source: UNESCO Institute for Statistics (2025)

#Research & Development spending as a share of GDP
RD_spending_proportion <- read.csv('https://ourworldindata.org/grapher/research-spending-g
#Data source: UNESCO Institute for Statistics (2025)

#Annual patent applications related to AI per million people
patent_ai_per_million_people <- read.csv('https://ourworldindata.org/grapher/artificial-in
#Data source: Center for Security and Emerging Technology (2024); Population based on vari

#World Bank income groups
income_group <- read.csv('https://ourworldindata.org/grapher/world-bank-income-groups.csv?
#Data source: World Bank (2024)
```

2. Merge data based on the same columns

```
merged_data <- Reduce(function(x, y) merge(x, y, by = c("Entity", "Code", "Year"), all = T
                      list(article_per_million_people,
                           patent_per_million_people,
                           researcher_per_million_people,
                           RD_spending_proportion,
                           patent_ai_per_million_people,
                           income_group))
```

3. Remove blank data and update the index and column names

```
#remove blank data and update index
cleaned_data <- na.omit(merged_data)
row.names(cleaned_data) <-NULL
#update column names
cleaned_data <- cleaned_data %>%
  rename(
    article_per_million_people = articles_per_million,
    patent_per_million_people = patents_per_million,
    researcher_per_million_people = Researchers.in.R.D..per.million.people.,
    RD_spending_proportion = Research.and.development.expenditure....of.GDP.,
    patent_ai_per_million_people = num_patent_applications_per_mil__field_all
  )
```

4. Check data

```
head(cleaned_data)
```

```
     Entity Code Year article_per_million_people patent_per_million_people
1 Argentina  ARG 2013                   186.2688                 15.169291
2 Argentina  ARG 2016                   194.3170                 20.243547
3 Argentina  ARG 2017                   195.7600                  8.920745
4 Argentina  ARG 2018                   204.6856                  9.569142
5 Argentina  ARG 2019                   200.7969                  9.878085
6 Argentina  ARG 2020                   216.0437                 20.650131
  researcher_per_million_people RD_spending_proportion
1                      1198.984                0.61849
```

```
2                      1260.701                0.55815
3                      1212.457                0.55631
4                      1227.404                0.48830
5                      1231.517                0.47813
6                      1256.267                0.54154
  patent_ai_per_million_people                classification
1                   0.02348385 Upper-middle-income countries
2                   0.02277888 Upper-middle-income countries
3                   0.02257902         High-income countries
4                   0.15675780 Upper-middle-income countries
5                   0.17788266 Upper-middle-income countries
6                   0.22127830 Upper-middle-income countries
```

```r
str(cleaned_data)
```

```
'data.frame':   290 obs. of  9 variables:
 $ Entity                       : chr  "Argentina" "Argentina" "Argentina" "Argentina" ...
 $ Code                         : chr  "ARG" "ARG" "ARG" "ARG" ...
 $ Year                         : int  2013 2016 2017 2018 2019 2020 2013 2014 2015 2016 ...
 $ article_per_million_people   : num  186 194 196 205 201 ...
 $ patent_per_million_people    : num  15.17 20.24 8.92 9.57 9.88 ...
 $ researcher_per_million_people: num  1199 1261 1212 1227 1232 ...
 $ RD_spending_proportion       : num  0.618 0.558 0.556 0.488 0.478 ...
 $ patent_ai_per_million_people : num  0.0235 0.0228 0.0226 0.1568 0.1779 ...
 $ classification               : chr  "Upper-middle-income countries" "Upper-middle-income
 - attr(*, "na.action")= 'omit' Named int [1:8260] 1 2 3 4 5 6 7 8 9 10 ...
  ..- attr(*, "names")= chr [1:8260] "1" "2" "3" "4" ...
```

```r
summary(cleaned_data)
```

```
    Entity              Code                Year     
 Length:290         Length:290         Min.   :2013  
 Class :character   Class :character   1st Qu.:2015  
 Mode  :character   Mode  :character   Median :2017  
                                       Mean   :2017  
                                       3rd Qu.:2019  
                                       Max.   :2020  
 article_per_million_people patent_per_million_people
 Min.   :  13.01            Min.   :   0.682         
 1st Qu.: 471.05            1st Qu.:  38.661         
```

```
Median :1109.73            Median : 110.165
Mean   :1079.79            Mean   : 300.668
3rd Qu.:1511.73            3rd Qu.: 233.707
Max.   :2655.37            Max.   :3481.109
researcher_per_million_people RD_spending_proportion
Min.   :  70.48               Min.   :0.1203
1st Qu.:1732.12               1st Qu.:0.8998
Median :3623.70               Median :1.3771
Mean   :3598.85               Mean   :1.7094
3rd Qu.:5194.75               3rd Qu.:2.3877
Max.   :8614.64               Max.   :4.7957
patent_ai_per_million_people classification
Min.   :  0.00356            Length:290
1st Qu.:  0.23706            Class :character
Median :  0.67502            Mode  :character
Mean   :  4.84308
3rd Qu.:  2.56590
Max.   :159.97383
```

5. Questions of interest

1. Does R&D investment in East Asian countries have a significant impact on patent applications related to AI?
2. How does R&D investment affect patent applications and articles published in scientific and technical journals in different countries?

Stage 2. Analyse an interesting data set (R part)

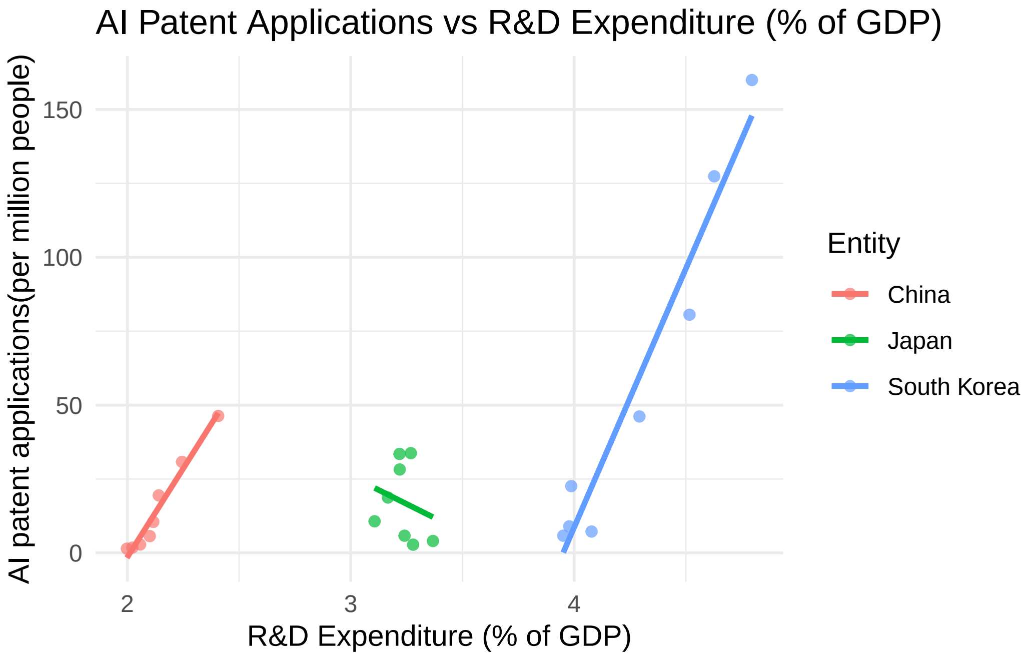
Question 1:

Does R&D investment (number of R&D researchers per million people, R&D spending as a share of GDP) in East Asian countries (China, Japan, South Korea) have a significant impact on patent applications related to AI (per million people)?

```r
east_asia_countries <- c("China", "Japan", "South Korea")
east_asia_data <- cleaned_data %>%
  filter(Entity %in% east_asia_countries)
#x1
#east_asia_data$RD_spending_proportion
#x2
```

```
#east_asia_data$researcher_per_million_people
#y
#east_asia_data$patent_ai_per_million_people
```

```
#y ~ x1 scatterplot
ggplot(east_asia_data, aes(x = RD_spending_proportion,
                           y = patent_ai_per_million_people,
                           color= Entity)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE, aes(group = Entity)) +
  labs(title = "AI Patent Applications vs R&D Expenditure (% of GDP)",
       x = "R&D Expenditure (% of GDP)",
       y = "AI patent applications per million people ") +
  theme_minimal()
```



The graph shows a positive correlation between R&D expenditure and AI patent applications in China and South Korea.

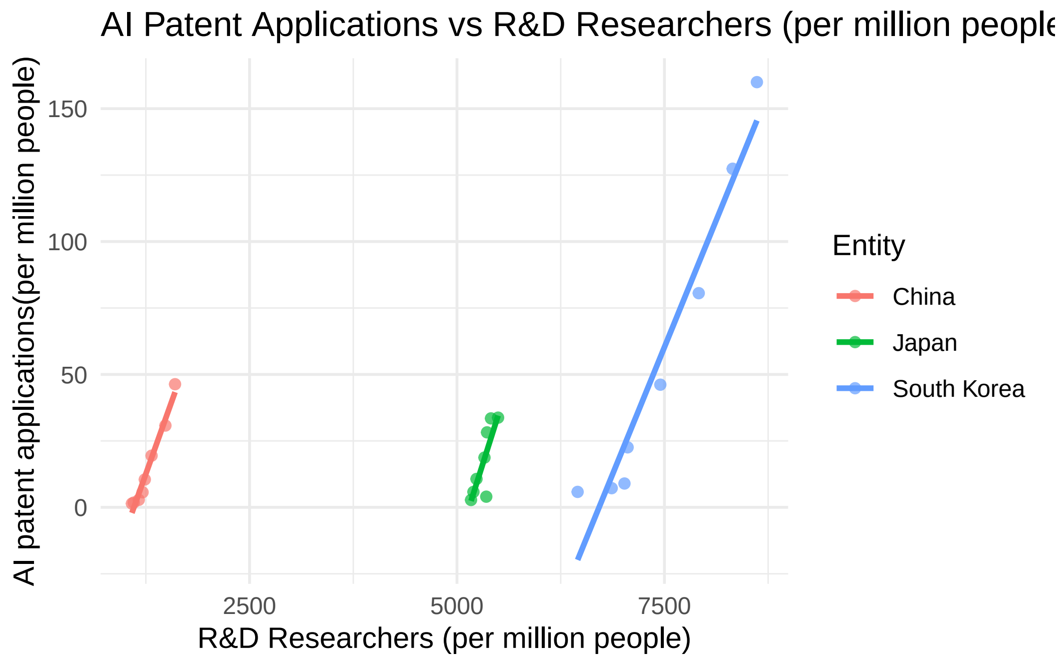
Countries with higher R&D spending tend to have more AI-related patents

China (Red): Most data points are concentrated in the lower R&D investment range (around 2%-2.5%). AI patent applications are relatively low, though some points show higher values.

Japan (Green): Data points are distributed within the moderate R&D investment range (3.0%-3.5%). AI patent applications appear stable without extreme fluctuations. The situation in Japan is quite unique, as the fitted line shows a negative correlation, which may indicate: There are few data points, which affects the regression results. Japan's AI patent applications may rely more on other factors such as market demand and policy support, rather than just R&D expenditures.

South Korea (Blue): South Korea has the highest R&D investment (above 4%). AI patent applications show greater dispersion, with some points significantly higher than those of China and Japan.

```r
#y ~ x2 scatterplot
ggplot(east_asia_data, aes(x = researcher_per_million_people,
                           y = patent_ai_per_million_people,
                           color= Entity)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE, aes(group = Entity)) +
  labs(title = "AI Patent Applications vs R&D Researchers (per million people)",
       x = "R&D Researchers (per million people)",
       y = "AI patent applications per million people ") +
  theme_minimal()
```



There is a positive correlation between R&D researchers per million people and AI patent applications.

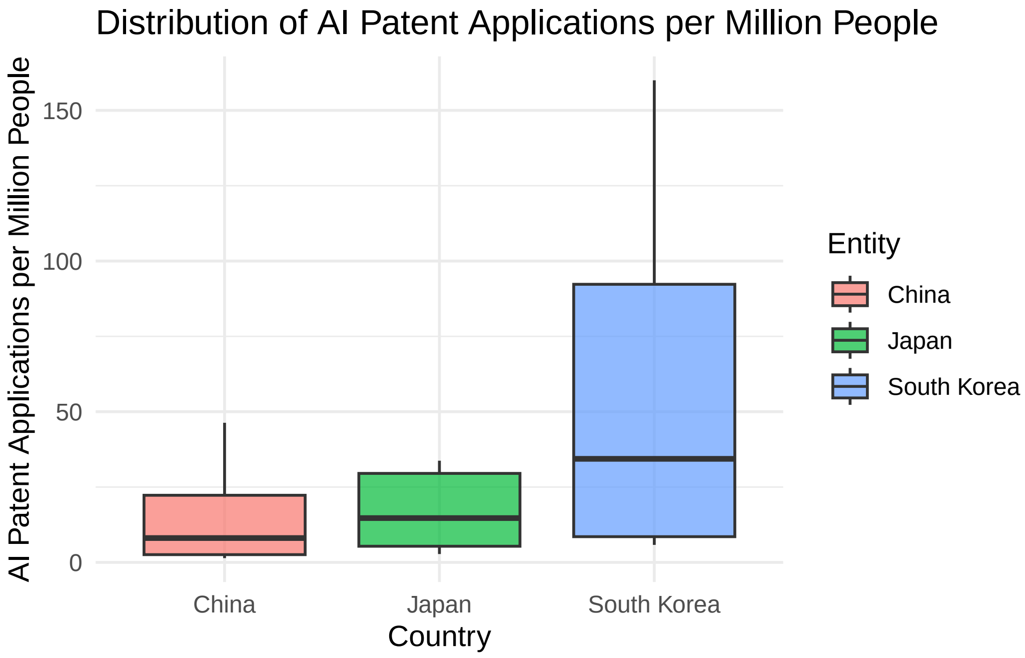
China (Red): Concentrated in the lower range of R&D researchers (below 3,000 per million). AI patent applications remain relatively low but show some increasing trend.

Japan (Green): Positioned in the middle range (around 3,500-5,000 researchers per million). AI patent applications are relatively stable but do not exhibit extreme values.

South Korea (Blue): Has the highest R&D researcher density (above 7,000 per million). Displays a higher variance in AI patent applications, with some significantly high values.

South Korea leads in both R&D researchers and AI patent applications, showcasing a highly productive AI innovation ecosystem. China, despite having the lowest number of researchers per million people, shows a strong positive relationship with AI patents, implying high efficiency or large-scale R&D efforts.

```
#y boxplot
ggplot(east_asia_data, aes(x = Entity,
                           y = patent_ai_per_million_people, fill = Entity)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Distribution of AI Patent Applications per Million People",
       x = "Country",
       y = "AI Patent Applications per Million People") +
  theme_minimal()
```



The boxplot shows South Korea exhibits the widest spread in AI patent applications per million people. China and Japan have more compact distributions, suggesting less variation in their AI patent outputs.

South Korea has the largest variability, which suggests that its AI patent production fluctuates significantly over time.

China and Japan have relatively stable distributions, meaning their AI patent outputs are more consistent.

```r
#summary
summary(east_asia_data)
```

```
   Entity               Code                Year     
Length:24          Length:24          Min.   :2013  
Class :character   Class :character   1st Qu.:2015  
Mode  :character   Mode  :character   Median :2016  
                                      Mean   :2016  
                                      3rd Qu.:2018  
                                      Max.   :2020  
article_per_million_people patent_per_million_people
Min.   : 259.0             Min.   : 512.3           
1st Qu.: 415.7             1st Qu.: 928.7           
Median : 804.9             Median :2041.7           
Mean   : 801.8             Mean   :2016.2           
3rd Qu.:1195.5             3rd Qu.:3155.5           
Max.   :1398.2             Max.   :3481.1           
researcher_per_million_people RD_spending_proportion
Min.   :1082                  Min.   :1.998         
1st Qu.:1444                  1st Qu.:2.219         
Median :5342                  Median :3.230         
Mean   :4686                  Mean   :3.216         
3rd Qu.:6903                  3rd Qu.:3.980         
Max.   :8615                  Max.   :4.796         
patent_ai_per_million_people classification    
Min.   :  1.418              Length:24         
1st Qu.:  5.771              Class :character  
Median : 14.701              Mode  :character  
Mean   : 29.784                                
3rd Qu.: 33.535                                
Max.   :159.974                                
```

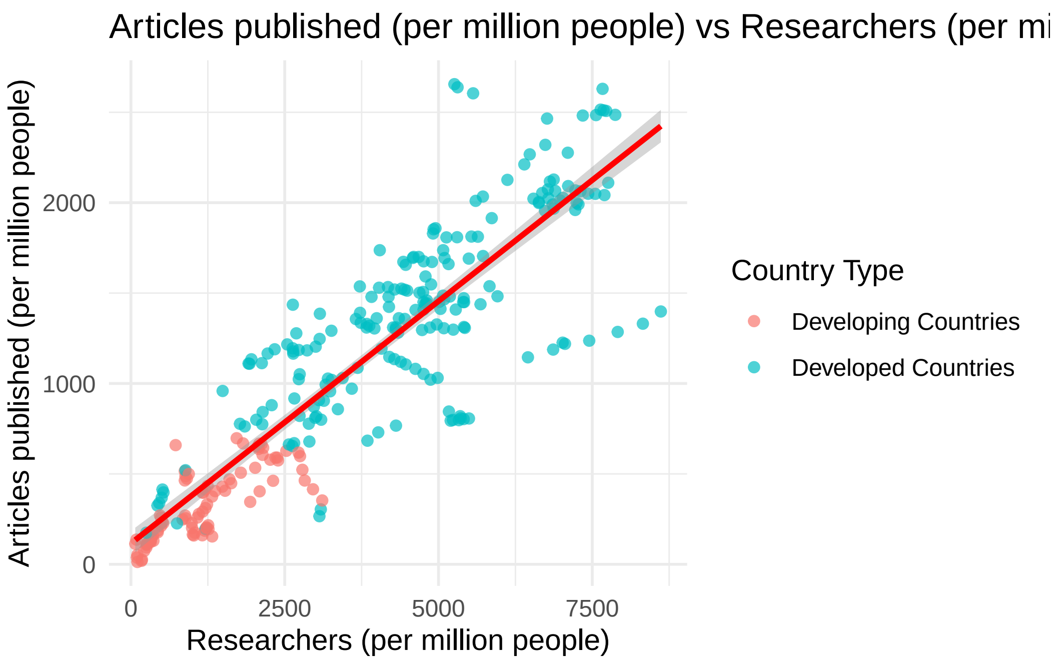
Question 2:

How does R&D investment (number of R&D researchers per million people, R&D spending as a share of GDP) affect patent applications (per million people) and articles published in scientific and technical journals (per million people) in different countries (developed countries, developing countries)?

```
#add a new cloumn that reprents the country type: developed countries or developing countr
cleaned_data <- cleaned_data %>%
  mutate(Developed_Status = factor(
    case_when(
      classification == "High-income countries" ~ "Developed Countries",  # developed coun
      classification %in% c("Upper-middle-income countries", "Lower-middle-income countrie
    ),
    levels = c("Developing Countries", "Developed Countries")
  ))

#x1
#cleaned_data$researcher_per_million_people
#x2
#cleaned_data$RD_spending_proportion
#x3
#cleaned_data$Developed_Status
#y1
#cleaned_data$article_per_million_people
#y2
#cleaned_data$patent_per_million_people
```

```
#y1 ~ x1 scatterplot
ggplot(cleaned_data, aes(x = researcher_per_million_people,
                         y = article_per_million_people,
                         color= Developed_Status)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "lm", color = "red", se = TRUE) +
  labs(title = "Articles published (per million people) vs Researchers (per million people
       x = "Researchers (per million people)",
       y = "Articles published (per million people)",
       color = "Country Type") +
  theme_minimal()
```

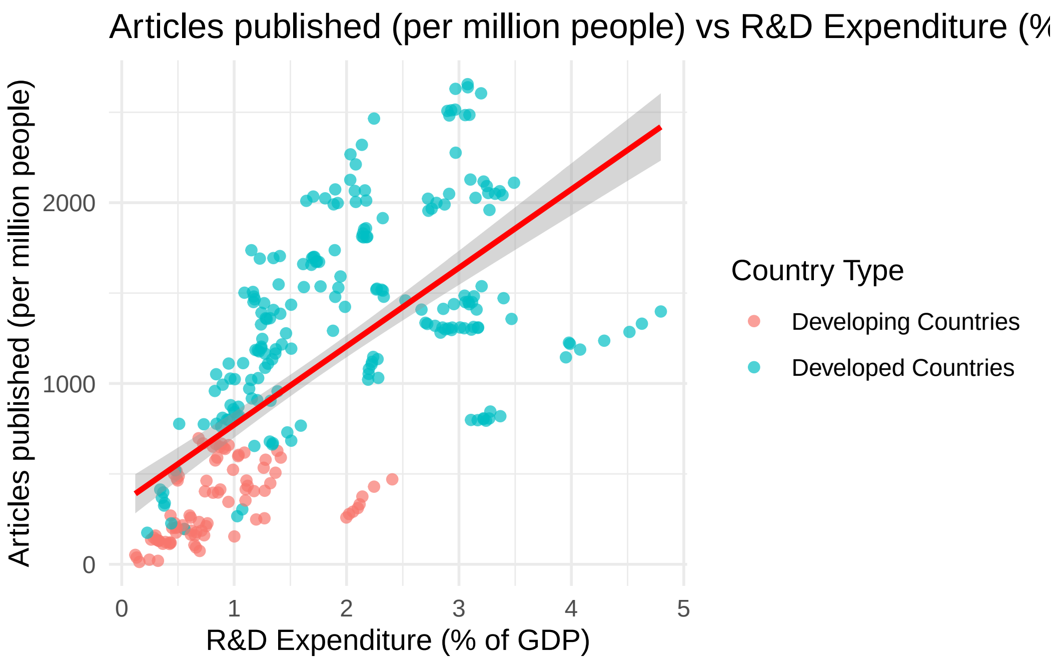


The scatterplot effectively shows a **positive correlation** between the number of researchers per million people and the number of published articles per million people.

The **linear trend** suggests that countries with more researchers tend to publish more articles.

Developed countries tend to have more researchers and more published articles than developing countries.

```
#y1 ~ x2 scatterplot
ggplot(cleaned_data, aes(x = RD_spending_proportion,
                         y = article_per_million_people,
                         color= Developed_Status)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "lm", color = "red", se = TRUE) +
  labs(title = "Articles published (per million people) vs R&D Expenditure (% of GDP)",
       x = "R&D Expenditure (% of GDP)",
       y = "Articles published (per million people)",
       color = "Country Type") +
  theme_minimal()
```

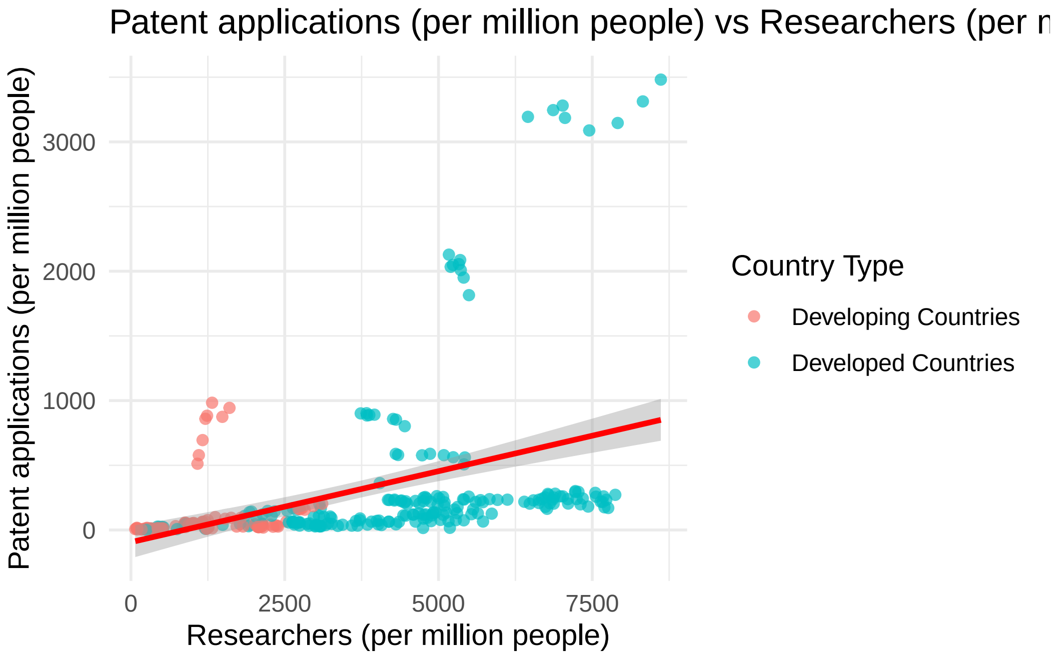


This scatterplot shows that there is a **positive correlation** between annual articles published per million people and R&D spending as a share of GDP.

The **linear trend** of this data indicates that more investment in the R&D leads to more published articles.

Developed countries tend to have more R&D investment and more articles published in scientific and technical journals than developing countries.

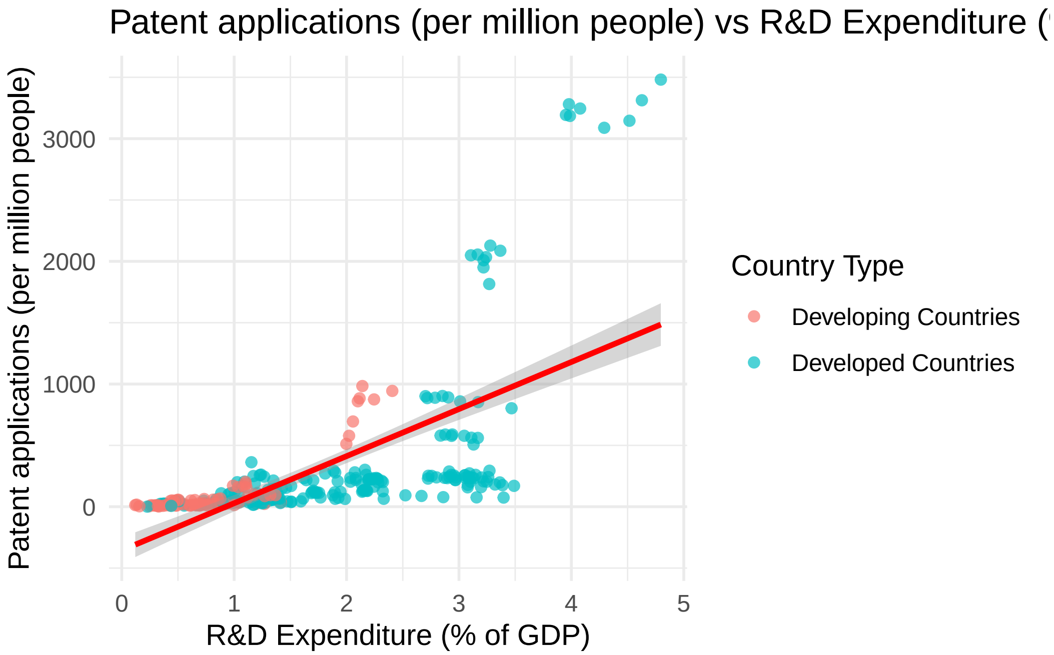
```
#y2 ~ x1 scatterplot
ggplot(cleaned_data, aes(x = researcher_per_million_people,
                            y = patent_per_million_people,
                            color= Developed_Status)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "lm", color = "red", se = TRUE) +
  labs(title = "Patent applications (per million people) vs Researchers (per million peopl
       x = "Researchers (per million people)",
       y = "Patent applications (per million people)",
       color = "Country Type") +
  theme_minimal()
```



Developing countries tend to have **fewer researchers** and **lower patent applications**, while developed countries show **greater dispersion and higher values.** But the correlation between the two variables is not very clear.

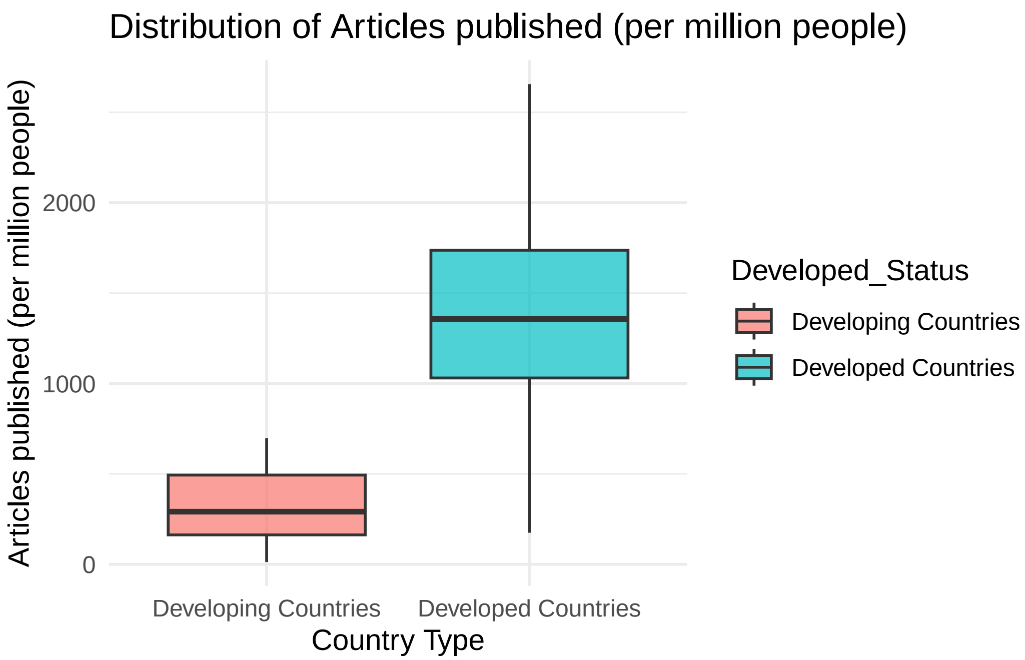
Some developing countries appear to have **higher-than-expected patent applications** given their number of researchers—this could be worth further investigation.

```
#y2 ~ x2 scaterplot
ggplot(cleaned_data, aes(x = RD_spending_proportion,
                          y = patent_per_million_people,
                          color= Developed_Status)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "lm", color = "red", se = TRUE) +
  labs(title = "Patent applications (per million people) vs R&D Expenditure (% of GDP)",
       x = "R&D Expenditure (% of GDP)",
       y = "Patent applications (per million people)",
       color = "Country Type") +
  theme_minimal()
```



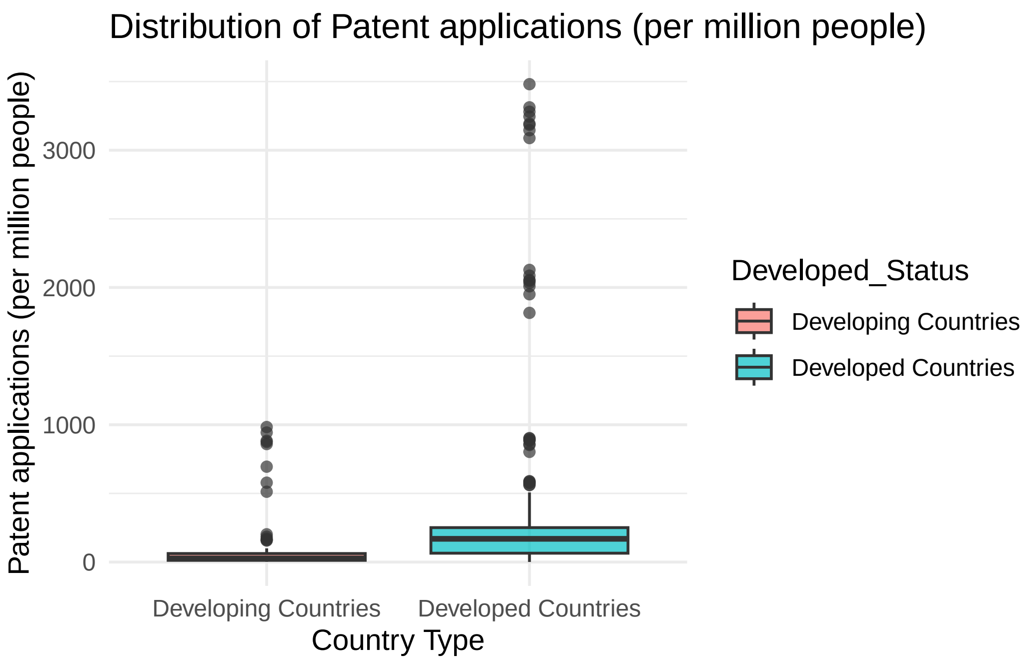
Developing countries generally have **lower R&D expenditure** an **lower patent applications**, but some show unexpectedly high patent activity. While developed countries display **greater variation in both R&D spending and patent applications**, with some countries exhibiting a strong innovative output.

```r
#y1 boxplot
ggplot(cleaned_data, aes(x = Developed_Status,
                         y = article_per_million_people, fill = Developed_Status)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Distribution of Articles published (per million people)",
       x = "Country Type",
       y = "Articles published (per million people)") +
  theme_minimal()
```



This plot shows that developed countries have **higher median** and **wider distribution**, while developing countries have **lower** number of published articles.

```
#y2 boxplot
ggplot(cleaned_data, aes(x = Developed_Status,
                         y = patent_per_million_people, fill = Developed_Status)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Distribution of Patent applications (per million people)",
       x = "Country Type",
       y = "Patent applications (per million people)") +
  theme_minimal()
```



This plot indicates that both developing countries and developed countries have **low level** of patent applications, but developed countries tend to have **larger** number of that.

```r
#summary
summary(cleaned_data)
```

```
   Entity              Code                Year     
Length:290         Length:290         Min.   :2013  
Class :character   Class :character   1st Qu.:2015  
Mode  :character   Mode  :character   Median :2017  
                                      Mean   :2017  
                                      3rd Qu.:2019  
                                      Max.   :2020  
article_per_million_people patent_per_million_people
Min.   :  13.01            Min.   :   0.682         
1st Qu.: 471.05            1st Qu.:  38.661         
Median :1109.73            Median : 110.165         
Mean   :1079.79            Mean   : 300.668         
3rd Qu.:1511.73            3rd Qu.: 233.707         
Max.   :2655.37            Max.   :3481.109         
researcher_per_million_people RD_spending_proportion
Min.   :  70.48               Min.   :0.1203        
```

```
1st Qu.:1732.12               1st Qu.:0.8998
Median :3623.70               Median :1.3771
Mean   :3598.85               Mean   :1.7094
3rd Qu.:5194.75               3rd Qu.:2.3877
Max.   :8614.64               Max.   :4.7957
patent_ai_per_million_people classification                    Developed_Status
Min.   :  0.00356            Length:290         Developing Countries: 87
1st Qu.:  0.23706            Class :character   Developed Countries :203
Median :  0.67502            Mode  :character
Mean   :  4.84308
3rd Qu.:  2.56590
Max.   :159.97383
```

Export datasets

```
write.csv(east_asia_data, "Group_27_Data_1.csv", row.names = FALSE)
write.csv(cleaned_data, "Group_27_Data_2.csv", row.names = FALSE)
```