ESP 106Lab 3

Vickie Bach

2026-01-25

ESP 106 Lab 3

In this lab we will start by reading merging in data on economic development and indoor and outdoor air pollution. Then we will practice making some graphs with it.

1. First read in the csv files: gdppercapitaandgini and airpollution

Both datasets are from Our World in Data The GDP dataset has GDP per capita and the GINI index (a measure of income inequality)
The air pollution dataset has death rates from indoor and outdoor air pollution - units are in deaths per 100,000 people
Indoor air pollution is the Household Air Pollution from Solid Fuels
Outdoor air pollution is split into particulate matter and ozone

Hint: Make sure to save all material for the lab into one sensible directory, probably one within your Github repository. The .csv files used in this lab are small enough to add to Github if you like. Then set that as your working directory. By default, the working directory for the Rmarkdown file will be the directory where your markdown file is saved. See more info here

Hint: The column names are long and cumbersome (because they contain information about units et) - you might want to rename some of the columns to make them easier to work with

```r
airpollution <- read.csv("data/airpollution.csv")
gdppercapiandgini <- read.csv("data/gdppercapiandgini.csv")

air <-airpollution

colnames(airpollution) = c("Entity","Code", "Year", "pm", "indoor", "ozone", "total_air")
colnames(gdppercapiandgini) = c("Entity", "Code", "Year", "total_population", "Continent", "gini", "GDP

country1 <- "United States"
country2 <- "China"

air_us <- subset(airpollution, Entity == country1)
air_china <- subset(airpollution, Entity == country2)
```

```
## [1] "\Users\74112\Documents\R\Lab 3"
```

2. Chose two countries that you are interested in and make a plot showing the death rates from indoor air pollution and outdoor air pollution (sum of particulate matter and ozone) over time
   Distinguish the countries using different colored lines and the types of pollution using different line types
   Make sure to add a legend and appropriate titles for the axes and plot

Hint: you can see all the different country names using unique(x$Entity) where x is the data frame containing the air pollution data Then create two new data frames that countain only the rows corresponding to each of the two countries you want to look at Create a new column of total outdoor air pollution deaths by summing death rates from particulate matter and ozone Use these to make your plot and add the lines you need

Hint: you might have to set the y scale manually to make sure your plot is wide enough to show both countries. You can do this using the “ylim” argument in plot

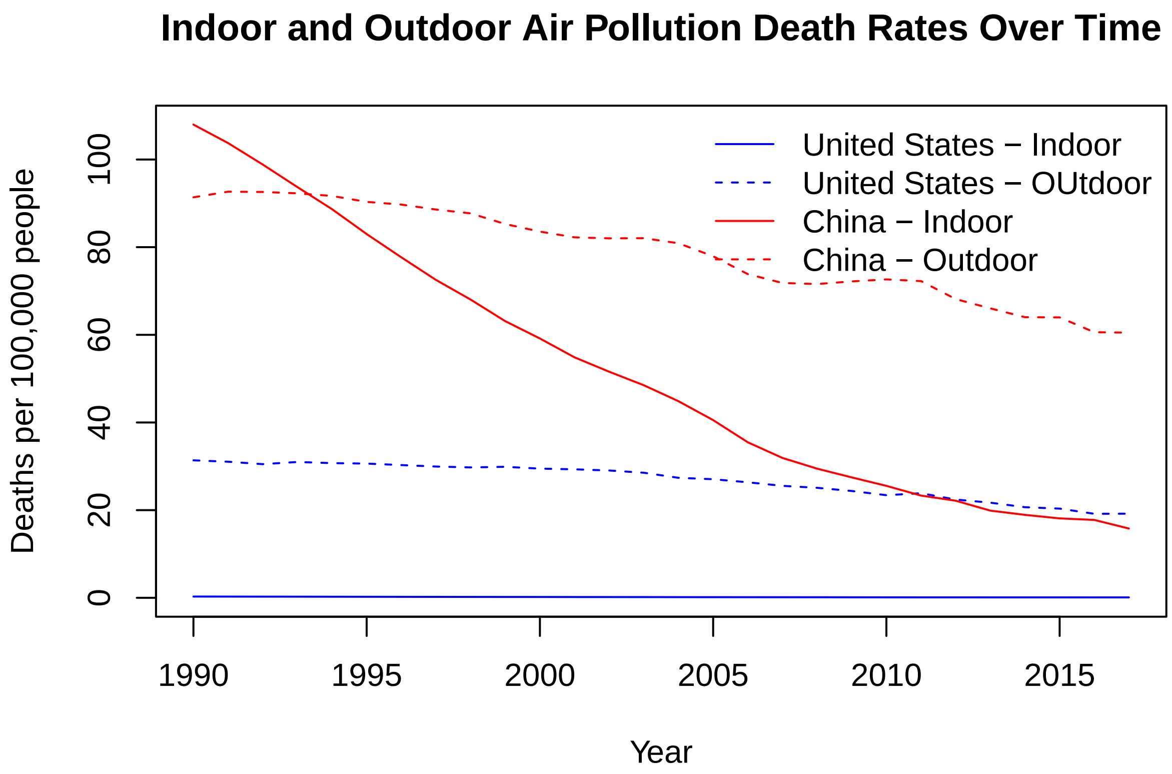
```
if(nrow(air_us) > 0){air_us$outdoor_total <- air_us$pm + air_us$ozone}
if(nrow(air_china) > 0){air_china$outdoor_total <- air_china$pm + air_china$ozone}

y_max <- max(air_us$indoor, air_us$outdoor_total, air_china$indoor, air_china$outdoor_total, na.rm = TRU

plot(air_us$Year, air_us$indoor, type = "l", col = "blue", lty = 1, ylim = c(0, y_max), xlab = "Year", 

lines(air_us$Year, air_us$outdoor_total, col = "blue", lty = 2)
lines(air_china$Year, air_china$indoor, col = "red", lty = 1)
lines(air_china$Year, air_china$outdoor_total, col = "red", lty = 2)

legend("topright", legend = c("United States - Indoor", "United States - OUtdoor", "China - Indoor", "C
```



3. Merge the air pollution data with the gdp data using merge()
   Merge is a function that combines data across two data frames by matching ID rows
   By default merge will identify ID rows as those where column names are the same between datasets, but it is safer to specify the columns you want to merge by yourself using “by”
   In our case, we want to merge both by country (either the “Entity” or “Code” columns) and year columns
   Note that by default, the merge function keeps only the entries that appear in both data frames - that is fine for this lab. If you need for other applications, you can change using the all.x or all.y arguments

to the function - check out the documentation at ?merge

```r
air_subset <- airpollution[,c("Entity", "Year", "indoor", "pm", "ozone")]
gdp_subset <- gdppercapiandgini[,c("Entity","Year","gini","GDP")]
air_gdp <- merge(air_subset, gdp_subset, by = c("Entity","Year"))
```

4. Make a plot with two subplots - one showing a scatter plot between log of per-capita GDP (x axis) and indoor air pollution death rate (y axis) and one showing log of per-capita GDP (x axis) and outdoor air pollution (y axis)
   Make sure to add appropriate titles to the plots and axes
   Use ylim to keep the range of the y axis the same between the two plots - this makes it easier for the reader to compare across the two graphs

STRECTH GOAL CHALLENGE - color the points based on continent. NOT REQUIRED FOR FULL POINTS - a challenge if you want to push yourself - continent info is included in the GDP dataset, but it is only listed for the year 2015
If you are trying this and getting stuck ASK FOR HELP - there are some tips and tricks for making it easier

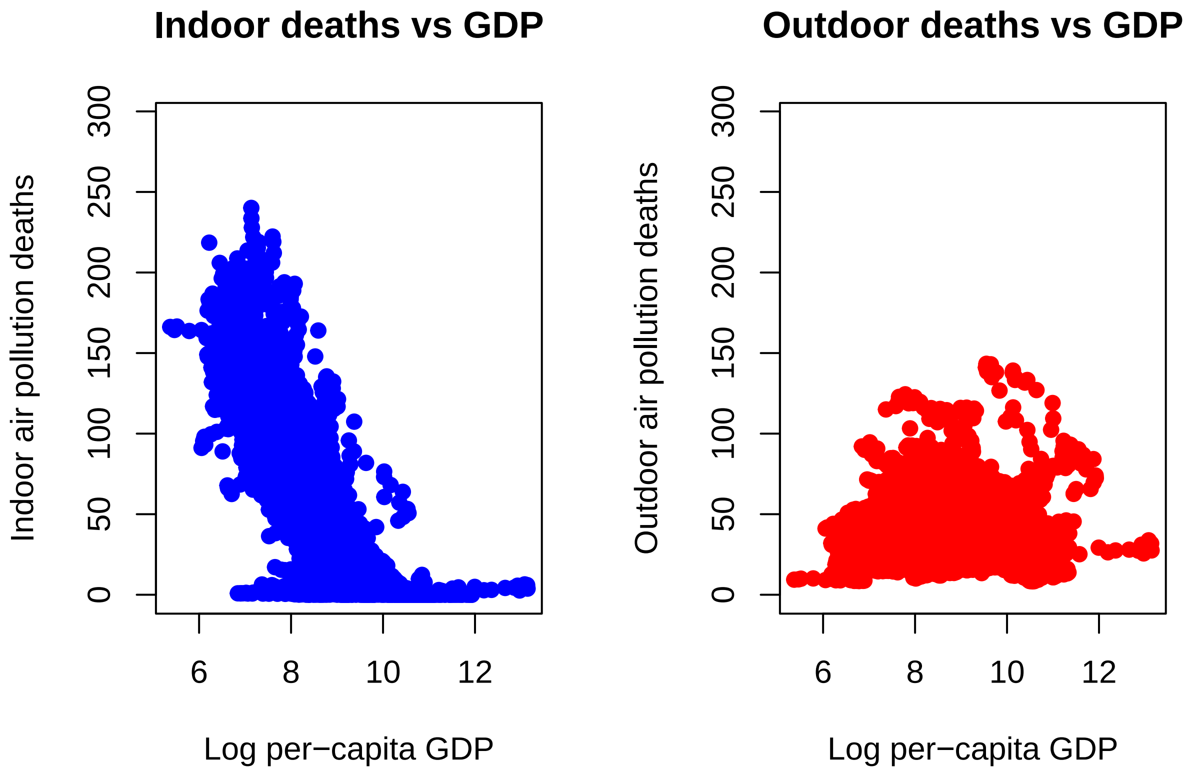
```r
air_gdp$log_GDP <- log(air_gdp$GDP)

y_range_indoor <- range(air_gdp$indoor, na.rm = TRUE)
y_range_outdoor <- range(air_gdp$pm + air_gdp$ozone, na.rm = TRUE)
y_limits <- range(y_range_indoor, y_range_outdoor)

par(mfrow = c(1, 2))

plot(air_gdp$log_GDP, air_gdp$indoor, ylim = y_limits, xlab = "Log per-capita GDP", ylab = "Indoor air

plot(air_gdp$log_GDP, air_gdp$pm + air_gdp$ozone, ylim = y_limits, xlab = "Log per-capita GDP", ylab =
```



```r
par(mfrow = c(1, 1))
```

5. Submission: Upload your Rmarkdown document and knitted PDF document to Canvas. Add your Rmarkdown file to your Github repository, commit your changes and push to your online repository (as we did Wednesday or last week)