ESP 106Lab 3

Fran Moore

2025-01-21

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

```
#gdp data
gdp_data <- read.csv("gdp.csv")

#Air-Pollution data
air_pollution_data <- read.csv("airpollution.csv")
head(gdp_data)
```

```
##        Entity     Code Year Total.population..Gapminder..HYDE...UN. Continent
## 1    Abkhazia OWID_ABK 2015                                      NA      Asia
## 2 Afghanistan      AFG 1800                                 3280000
## 3 Afghanistan      AFG 1801                                 3280000
## 4 Afghanistan      AFG 1802                                 3280000
## 5 Afghanistan      AFG 1803                                 3280000
## 6 Afghanistan      AFG 1804                                 3280000
##   Gini.coefficient..World.Bank..2016..
## 1                                   NA
## 2                                   NA
## 3                                   NA
## 4                                   NA
## 5                                   NA
```

```
## 6                                     NA
##   Output.side.real.GDP.per.capita..gdppc_o...PWT.9.1..2019..
## 1                                                         NA
## 2                                                         NA
## 3                                                         NA
## 4                                                         NA
## 5                                                         NA
## 6                                                         NA
```

```
head(air_pollution_data)
```

```
##        Entity Code Year
## 1 Afghanistan  AFG 1990
## 2 Afghanistan  AFG 1991
## 3 Afghanistan  AFG 1992
## 4 Afghanistan  AFG 1993
## 5 Afghanistan  AFG 1994
## 6 Afghanistan  AFG 1995
##   Deaths...Ambient.particulate.matter.pollution...Sex..Both...Age..Age.standardized..Rate.
## 1                                                                                 46.44659
## 2                                                                                 46.03384
## 3                                                                                 44.24377
## 4                                                                                 44.44015
## 5                                                                                 45.59433
## 6                                                                                 45.36714
##   Deaths...Household.air.pollution.from.solid.fuels...Sex..Both...Age..Age.standardized..Rate.
## 1                                                                                     250.3629
## 2                                                                                     242.5751
## 3                                                                                     232.0439
## 4                                                                                     231.6481
## 5                                                                                     238.8372
## 6                                                                                     239.9066
##   Deaths...Ambient.ozone.pollution...Sex..Both...Age..Age.standardized..Rate.
## 1                                                                   5.616442
## 2                                                                   5.603960
## 3                                                                   5.611822
## 4                                                                   5.655266
## 5                                                                   5.718922
## 6                                                                   5.739174
##   Deaths...Air.pollution...Sex..Both...Age..Age.standardized..Rate.
## 1                                                          299.4773
## 2                                                          291.2780
## 3                                                          278.9631
## 4                                                          278.7908
## 5                                                          287.1629
## 6                                                          288.0142
```

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

```
us_data <- air_pollution_data[air_pollution_data$Entity == "United States", ]
argentina_data <- air_pollution_data[air_pollution_data$Entity == "Argentina", ]
 head(us_data)
```

```
##             Entity Code Year
## 6077 United States  USA 1990
## 6078 United States  USA 1991
## 6079 United States  USA 1992
## 6080 United States  USA 1993
## 6081 United States  USA 1994
## 6082 United States  USA 1995
##      Deaths...Ambient.particulate.matter.pollution...Sex..Both...Age..Age.standardized..Rate.
## 6077                                                                                  28.08404
## 6078                                                                                  27.70024
## 6079                                                                                  27.10677
## 6080                                                                                  27.44725
## 6081                                                                                  27.12268
## 6082                                                                                  26.93429
##      Deaths...Household.air.pollution.from.solid.fuels...Sex..Both...Age..Age.standardized..Rate.
## 6077                                                                                    0.2833959
## 6078                                                                                    0.2712254
## 6079                                                                                    0.2570071
## 6080                                                                                    0.2523433
## 6081                                                                                    0.2412800
## 6082                                                                                    0.2302462
##      Deaths...Ambient.ozone.pollution...Sex..Both...Age..Age.standardized..Rate.
## 6077                                                                   3.281703
## 6078                                                                   3.348164
## 6079                                                                   3.383141
## 6080                                                                   3.541285
## 6081                                                                   3.606160
## 6082                                                                   3.690748
##      Deaths...Air.pollution...Sex..Both...Age..Age.standardized..Rate.
## 6077                                                          31.19507
## 6078                                                          30.85611
## 6079                                                          30.27920
## 6080                                                          30.75236
## 6081                                                          30.47439
## 6082                                                          30.35046
```

```
head(argentina_data)
```

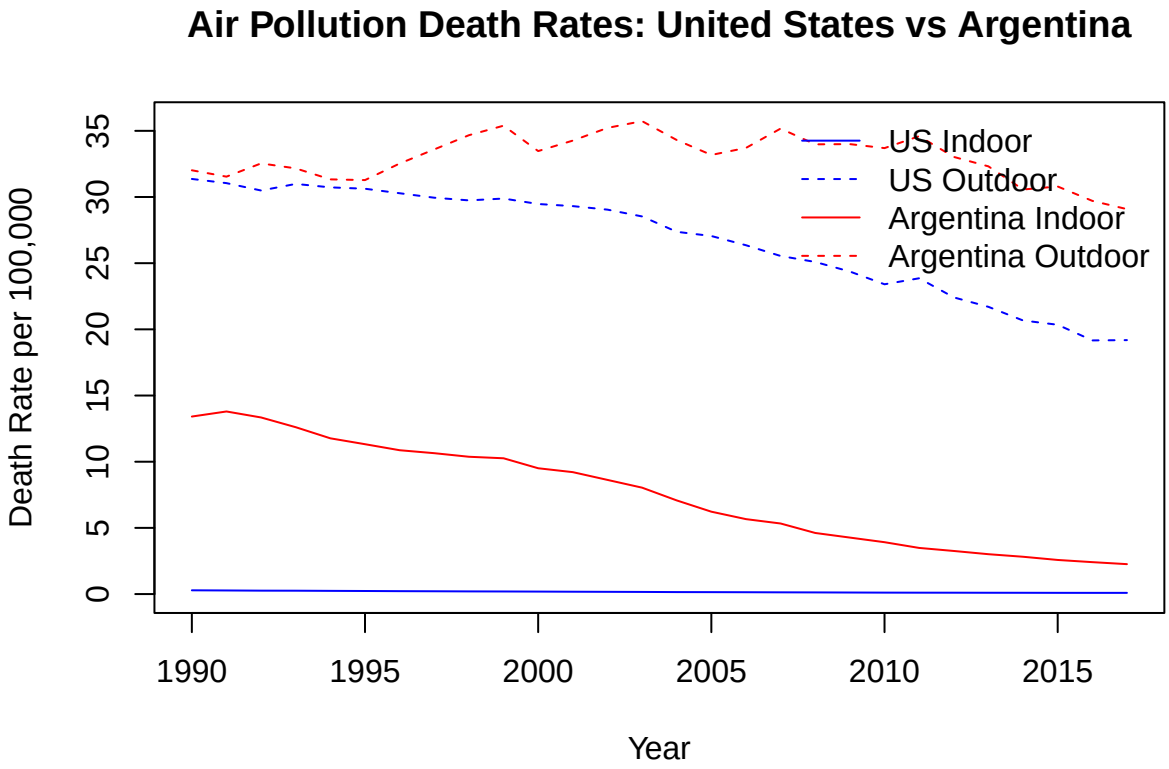
```
##        Entity Code Year
## 225 Argentina  ARG 1990
## 226 Argentina  ARG 1991
```

```
## 227 Argentina  ARG 1992
## 228 Argentina  ARG 1993
## 229 Argentina  ARG 1994
## 230 Argentina  ARG 1995
##     Deaths...Ambient.particulate.matter.pollution...Sex..Both...Age..Age.standardized..Rate.
## 225                                                                                  31.16498
## 226                                                                                  30.68255
## 227                                                                                  31.64959
## 228                                                                                  31.26103
## 229                                                                                  30.42736
## 230                                                                                  30.32275
##     Deaths...Household.air.pollution.from.solid.fuels...Sex..Both...Age..Age.standardized..Rate.
## 225                                                                                     13.41293
## 226                                                                                     13.79430
## 227                                                                                     13.34271
## 228                                                                                     12.60725
## 229                                                                                     11.76547
## 230                                                                                     11.32351
##     Deaths...Ambient.ozone.pollution...Sex..Both...Age..Age.standardized..Rate.
## 225                                                                  0.8544604
## 226                                                                  0.8545864
## 227                                                                  0.8903259
## 228                                                                  0.9017075
## 229                                                                  0.9055820
## 230                                                                  0.9559308
##     Deaths...Air.pollution...Sex..Both...Age..Age.standardized..Rate.
## 225                                                          45.26229
## 226                                                          45.15719
## 227                                                          45.70004
## 228                                                          44.58503
## 229                                                          42.91370
## 230                                                          42.40998
```

```r
# adding total for US
us_data$Total_Outdoor <- us_data$`Deaths...Ambient.particulate.matter.pollution...Sex..Both...Age..Age.s
                         us_data$`Deaths...Ambient.ozone.pollution...Sex..Both...Age..Age.standardized.

#addinh total Outdoor for Argentina
argentina_data$Total_Outdoor <- argentina_data$`Deaths...Ambient.particulate.matter.pollution...Sex..Bo
                                argentina_data$`Deaths...Ambient.ozone.pollution...Sex..Both...Age..Age
```

```r
plot(us_data$Year, us_data$`Deaths...Household.air.pollution.from.solid.fuels...Sex..Both...Age..Age.sta
     type = "l", col = "blue", lty = 1,
     ylim = c(0, max(c(us_data$`Deaths...Household.air.pollution.from.solid.fuels...Sex..Both...Age..Ag
                       us_data$Total_Outdoor,
                       argentina_data$`Deaths...Household.air.pollution.from.solid.fuels...Sex..Both...
                       argentina_data$Total_Outdoor))),
     xlab = "Year", ylab = "Death Rate per 100,000",
     main = "Air Pollution Death Rates: United States vs Argentina")

# US total Outdoor
lines(us_data$Year, us_data$Total_Outdoor, col = "blue", lty = 2)

# add Argentina indoor pollution
```

```r
lines(argentina_data$Year, argentina_data$`Deaths...Household.air.pollution.from.solid.fuels...Sex..Bot
      col = "red", lty = 1)

# Argentina total Outdoor
lines(argentina_data$Year, argentina_data$Total_Outdoor, col = "red", lty = 2)

# legend for easier readability
legend("topright", legend = c("US Indoor", "US Outdoor", "Argentina Indoor", "Argentina Outdoor"),
       col = c("blue", "blue", "red", "red"), lty = c(1, 2, 1, 2), bty = "n")
```



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
# Merging air pollution data with GDP data
merged_data <- merge(air_pollution_data, gdp_data,
                     by = c("Entity", "Year"))
head(merged_data)
```

```
##        Entity Year Code.x
## 1 Afghanistan 1990    AFG
## 2 Afghanistan 1991    AFG
## 3 Afghanistan 1992    AFG
## 4 Afghanistan 1993    AFG
## 5 Afghanistan 1994    AFG
## 6 Afghanistan 1995    AFG
##   Deaths...Ambient.particulate.matter.pollution...Sex..Both...Age..Age.standardized..Rate.
## 1                                                                                  46.44659
## 2                                                                                  46.03384
## 3                                                                                  44.24377
## 4                                                                                  44.44015
## 5                                                                                  45.59433
## 6                                                                                  45.36714
##   Deaths...Household.air.pollution.from.solid.fuels...Sex..Both...Age..Age.standardized..Rate.
## 1                                                                                    250.3629
## 2                                                                                    242.5751
## 3                                                                                    232.0439
## 4                                                                                    231.6481
## 5                                                                                    238.8372
## 6                                                                                    239.9066
##   Deaths...Ambient.ozone.pollution...Sex..Both...Age..Age.standardized..Rate.
## 1                                                                   5.616442
## 2                                                                   5.603960
## 3                                                                   5.611822
## 4                                                                   5.655266
## 5                                                                   5.718922
## 6                                                                   5.739174
##   Deaths...Air.pollution...Sex..Both...Age..Age.standardized..Rate. Code.y
## 1                                                          299.4773    AFG
## 2                                                          291.2780    AFG
## 3                                                          278.9631    AFG
## 4                                                          278.7908    AFG
## 5                                                          287.1629    AFG
## 6                                                          288.0142    AFG
##   Total.population..Gapminder..HYDE...UN. Continent
## 1                                12412000
## 2                                13299000
## 3                                14486000
## 4                                15817000
## 5                                17076000
## 6                                18111000
##   Gini.coefficient..World.Bank..2016..
## 1                                   NA
## 2                                   NA
## 3                                   NA
## 4                                   NA
## 5                                   NA
## 6                                   NA
##   Output.side.real.GDP.per.capita..gdppc_o...PWT.9.1..2019..
## 1                                                          NA
## 2                                                          NA
## 3                                                          NA
## 4                                                          NA
```

```
## 5                                                                    NA
## 6                                                                    NA
```

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
# Check the column names in merged_data
colnames(merged_data)
```

```
##  [1] "Entity"
##  [2] "Year"
##  [3] "Code.x"
##  [4] "Deaths...Ambient.particulate.matter.pollution...Sex..Both...Age..Age.standardized..Rate."
##  [5] "Deaths...Household.air.pollution.from.solid.fuels...Sex..Both...Age..Age.standardized..Rate."
##  [6] "Deaths...Ambient.ozone.pollution...Sex..Both...Age..Age.standardized..Rate."
##  [7] "Deaths...Air.pollution...Sex..Both...Age..Age.standardized..Rate."
##  [8] "Code.y"
##  [9] "Total.population..Gapminder..HYDE...UN."
## [10] "Continent"
## [11] "Gini.coefficient..World.Bank..2016.."
## [12] "Output.side.real.GDP.per.capita..gdppc_o...PWT.9.1..2019.."
```

```r
# Renaming the GDP per capita column to a simpler name since its complicated
colnames(merged_data)[colnames(merged_data) == "Output.side.real.GDP.per.capita..gdppc_o...PWT.9.1..2019
colnames(merged_data)
```

```
##  [1] "Entity"
##  [2] "Year"
##  [3] "Code.x"
##  [4] "Deaths...Ambient.particulate.matter.pollution...Sex..Both...Age..Age.standardized..Rate."
##  [5] "Deaths...Household.air.pollution.from.solid.fuels...Sex..Both...Age..Age.standardized..Rate."
##  [6] "Deaths...Ambient.ozone.pollution...Sex..Both...Age..Age.standardized..Rate."
##  [7] "Deaths...Air.pollution...Sex..Both...Age..Age.standardized..Rate."
##  [8] "Code.y"
##  [9] "Total.population..Gapminder..HYDE...UN."
## [10] "Continent"
## [11] "Gini.coefficient..World.Bank..2016.."
## [12] "GDP.per.capita"
```

```r
#remove rows where GDP.per.capita is NA or <= 0
merged_data <- merged_data[!is.na(merged_data$GDP.per.capita) & merged_data$GDP.per.capita > 0, ]
nrow(merged_data)
```

```
## [1] 4788
```

```
sum(is.na(merged_data$GDP.per.capita))
```

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

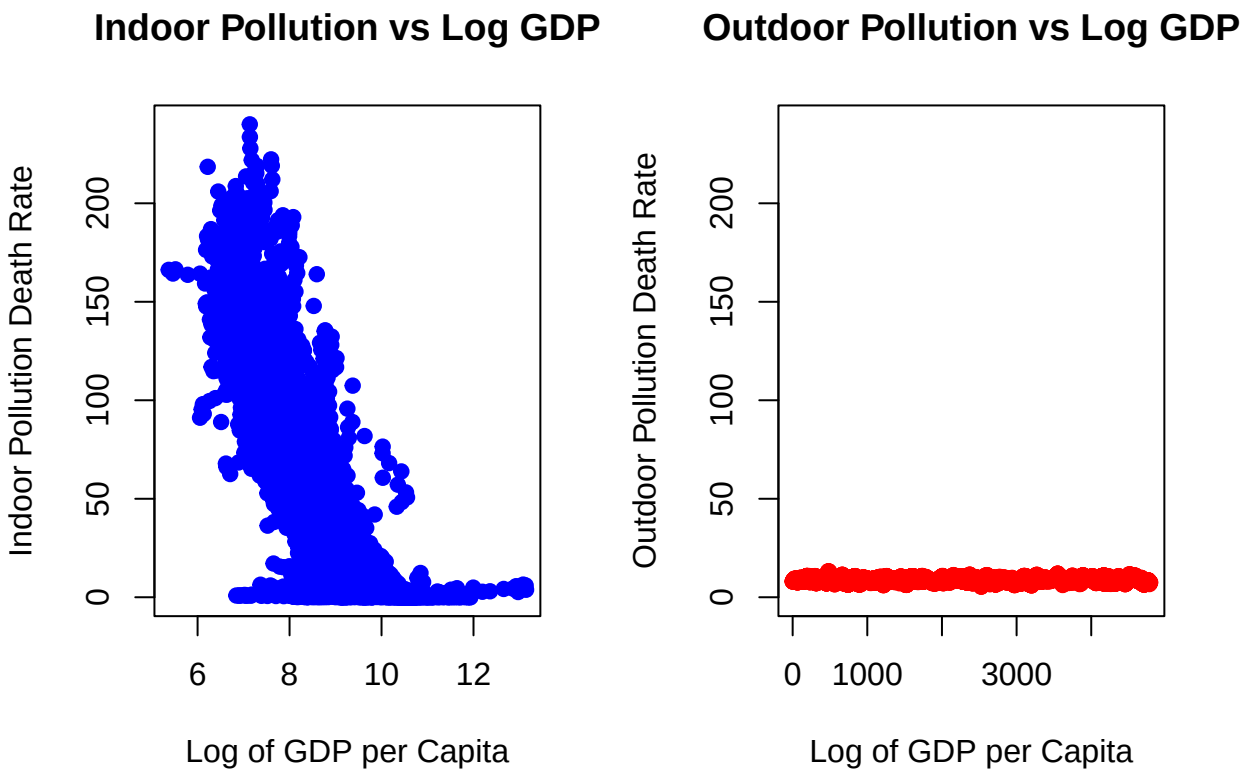
```
#calculate the log of GDP per capita
merged_data$Log_GDP <- log(merged_data$GDP.per.capita)
head(merged_data$Log_GDP)
```

```
## [1] 8.113854 8.000385 7.907174 8.064054 8.202534 8.288904
```

```
#calculate range of indoor and outdoor pollution death rates
y_range <- range(c(
  merged_data$`Deaths...Household.air.pollution.from.solid.fuels...Sex..Both...Age..Age.standardized..Ra
  merged_data$Total_Outdoor
), na.rm = TRUE)
y_range
```

```
## [1] 3.858714e-03 2.400830e+02
```

```
# side-by-side plotting
par(mfrow = c(1, 2))  # Divide the plotting area into 1 row and 2 columns

# Plot 1
plot(merged_data$Log_GDP,
     merged_data$`Deaths...Household.air.pollution.from.solid.fuels...Sex..Both...Age..Age.standardized
     xlab = "Log of GDP per Capita",
     ylab = "Indoor Pollution Death Rate",
     main = "Indoor Pollution vs Log GDP",
     ylim = y_range,
     pch = 19, col = "blue")

# Plot 2
plot(merged_data$Log_GDP,
     merged_data$Total_Outdoor,
     xlab = "Log of GDP per Capita",
     ylab = "Outdoor Pollution Death Rate",
     main = "Outdoor Pollution vs Log GDP",
     ylim = y_range,
     pch = 19, col = "red")
```



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)