Group Project 2

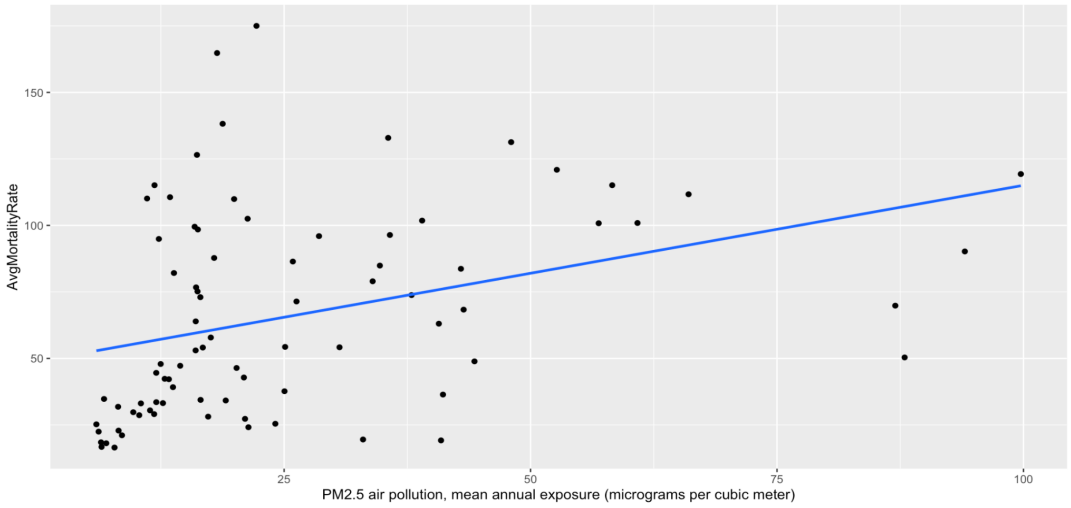
- What data set you plan to use
    - We wrangleed and merged 4 data sets that include the following:
        - Mortality rate attributed to household and ambient air pollution (per 100 000 population): link
        - Medical doctors (per 10 000 population): link
        - PM 2.5 air pollution, mean annual exposure
        - Designation of income status to country: link

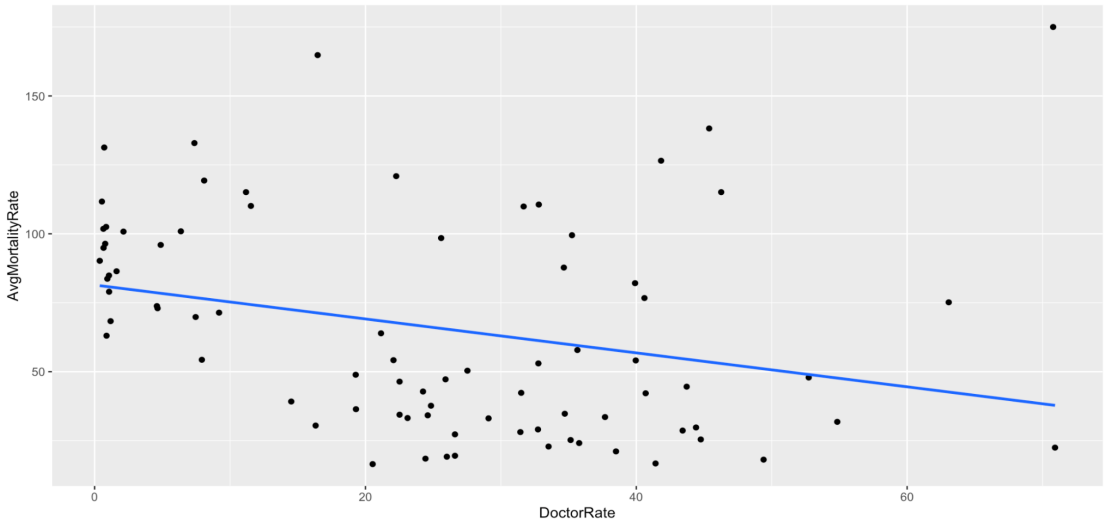
- What question(s) you want to answer with these data
    - Our research question/problem is: What is the effect of the number of doctors and the existing pollution level on the mortality rate of an income level status of a country?
    - We intend to treat mortality rate as the outcome variable, and the remaining variables as the explanatory variables. Medical doctors and PM 2.5 air pollution levels will be numerical variables, whereas income status of a country will be a categorical variable with levels: high income, middle income and low income.

- What model(s) do you plan to use to address this questions
    - You do not need to have *fit* these models, just explain in words or with an equation what their structure will be
    - We will fit a multiple regression model with two numerical explanatory variables (Doctor Rate and Air Pollution) for the outcome variable (mortality rate) and how those differ for each level of the categorical variable.

- Demonstration that you can import this data set into R
    - In our project repo (linked here): SDS291_GP_wrangling.qmd contains the code that successfully imports this dataset into R

- Include an exploratory visualization (e.g., a scatter plot, or side-by-side box plots, showing the relationship between your outcome and and one of your explanatory variables)


AvgMortalityRate
150
100
50
25
50
75
100
PM2.5 air pollution, mean annual exposure (micrograms per cubic meter)


AvgMortalityRate
150
100
50
0
20
40
60
DoctorRate