4. Spridningsdiagram, korrelation och regression (bygger vidare på labb 4).

Vi kommer att använda datasetet **gapm** med länder och sex variabler från Gapminder[1], som ni träffat på i labb 4. Vi har dessutom lagt till vår variabel ”landlocked” från tidigare i kursen. Du kan ladda ner data här (högerklicka), eller från Githubsidan (labb 4) eller från Datafiler i Athena. Data är från 2022. Variablerna som finns i datasetet är[2]:

country – de länder som finns i Gapminderdata och för vilka det finns kompletta data
child_mort – antal barn som dör före fem års ålder, per 1000 barn födda
fertility – förväntat antal barn per kvinna
co2_cap – antal ton koldioxid som varje individ ”konsumerar”
gdp_cap – BNP per capita i dollar (köpkraftsjusterat)
life_exp – förväntad medellivslängd
landlocked – indikator för om ett land har kust eller inte

Börja med detta:

Sätt arbetskatalogen och ladda mosaicpaketet. Ladda ner data till arbetskatalogen. Läs in data till R, från arbetskatalogen, med read.csv-kommandot och skapa en data frame med era inlästa data, kalla den exv. **gapm**.

Bekanta er med hur data ser ut genom kommandona head(gapm) – titta på de första sex raderna, str(gapm) – vilka variabeltyper vi har, class(gapm) – vilken typ av dataobjekt vi har, summary(gapm) – sammanfattande mått för de olika variablerna. Gör också gärna exv. histogram över de enskilda variablerna för att se hur data är fördelade, exempelvis medellivslängd och koldioxidutsläpp i olika länder (detta behöver inte tas med i redovisningen).

4.1 Ta fram korrelationskoefficienten mellan barnadödlighet och övriga variabler (förutom landlocked)

Med vilken annan variabel är korrelationen högst?

4.2 Gör ett spridningsdiagram för sambandet mellan barnadödlighet och bnp per capita

Här behöver ni inte fixa till axlarna och rubrik men beskriv hur sambandet ser ut. Är sambandet linjärt?

4.3 Gör ett spridningsdiagram för sambandet mellan förväntat antal barn per kvinna och barnadödlighet

Beskriv hur sambandet ser ut. Är sambandet linjärt?

4.4 Gör en regression med förväntat antal barn per kvinna som responsvariabel och barnadödlighet som förklaringsvariabel. Plotta regressionslinjen i det spridningsdiagram ni gjorde i 4.3.

Hur starkt är sambandet mellan de två variablerna (förklaringsgraden R2)? Är sambandet signifikant? Tolka lutningskoefficienten. Kan vi säga något om kausalitet?

4.5 Till regressionen i 4.4, lägg till variabeln landlocked som en andra förklaringsvariabel.

Förändras R2 och lutningskoefficienten från 4.4 nämnvärt? Förändras lutningskoefficienten från 4.3? Är variabeln landlocked en signifikant förklaringsvariabel?

---

[1] Based on free material from GAPMINDER.ORG, CC-BY LICENSE.
[2] Mer exakta definitioner av vissa av variablerna finns på Gapminders hemsida men är inte viktiga för uppgiften.