Statistik og dataanalyse
Tællende aktivitet 1

Peter Heilbo Ratgen - perat17@student.sdu.dk

11. oktober 2020

1 Databehandling

Datafilen indeholder data om landes CO2-udledning per indbygger i metriske ton per indbygger. Der er to kolonner i datafilen, "Country"og "CO2". Først gemmes .xlsx-filen som en .csv fil. Denne data er dog ikke komplet, det indeholder også de lande eller territorier der ikke har opgivet data. Det giver ikke mening at betragte disse lande i statistikken, derfor fjernes disse fra datasættet. Dette gøres ved at fjerne alle datapunkter der har ".."som værdi. Også datapunkter der ikke er lande skal også fjernes fra datasættet, derfor fjernes alle datapunkter hvor "income"indgår i navnet med kommandolinjeværktøjet `sed`. Her er der fjernet 10 datapunkter. Derudover fjernes datapunktet "world"også, siden det noget der kan regnes frem til. Til sidst i databehandlingen skal `,` ændres til `.`, for at R importerer data som den korrekte type.

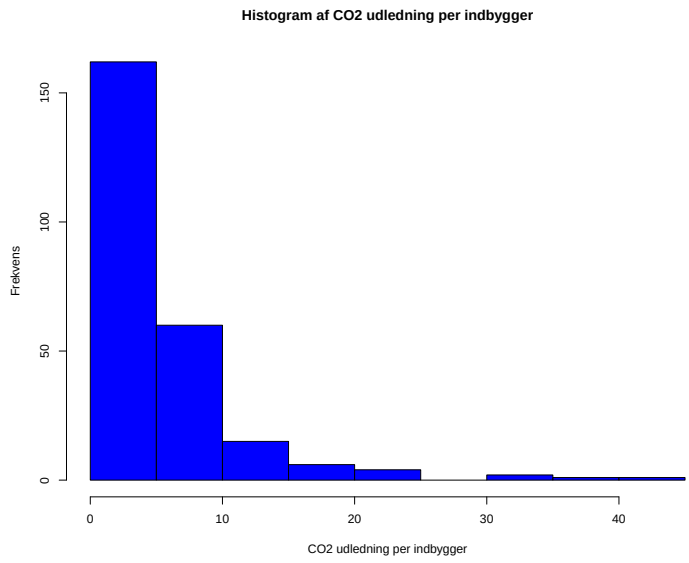
Nu kan data indporteres til R med:

```
1 co2  <- read.csv("CO2.csv", sep = ";", row.names = NULL)
```

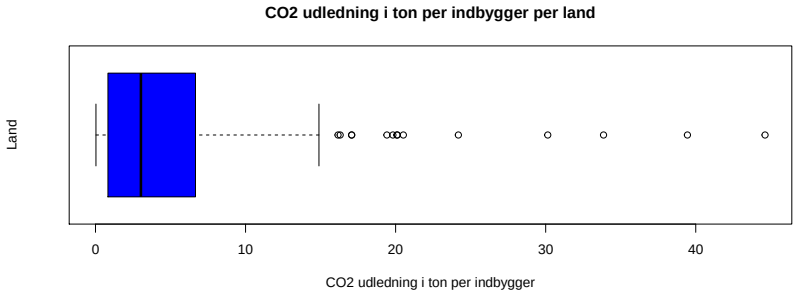
Her bruges `sep = ";"` til at indikere at værdierne i .csv formatet ikke er separeret med et komma, men med et semikolon.

2 Grafiske illustrationer

2.1 Histogram



2.2 Boxplot



Boxplottet fortæller at langt de flest lande (75%, eller dem inden for det 3. kvartil) ikke udleder over 10 ton CO2 per indbygger. Det fortæller også at der er nogle lande der udleder langt mere CO2 per indbygger, disse er de lande der ligger uden for 1.5 gange den interkvatile afstand.

Danmark ligger indenfor det 3. kvartil med sine 6.51 ton CO2 udledt per indbygger. Det vil også sige at Danmark ligger over medianen.

3 Deskriptiv analyse

Gennemsnit	4.9489
Median	3.0259
Typeinterval	0-5
Standardafvigelse	6.1881
Standard error	0.0246
Varians	38.2931
Minimum	0.0303
Maximum	44.6179
Q_1	0.8280
Q_3	6.6646

4 Normalfordeling

Som det ses på ovenstående histogram er data er højreskæv. Normalfordelingen har en symmetrisk klokkeform. En anden indikator er at gennemsnittet og medianen ikke er tæt på at være den samme værdi.

Derfor er data ikke normalfordelt.

5 R-kode

```r
co2  <- read.csv("CO2.csv", sep = ";", row.names = NULL)

h <- hist(co2$CO2)
pdf("co2plot.pdf", width = 10, height = 8)
plot(h, xlab = "CO2 udledning per indbygger", ylab = "Frekvens"
     , main = "Histogram af CO2 udledning per indbygger",
         col = "blue")
dev.off()

pdf("co2boxplot.pdf", width = 10, height = 4)
boxplot(co2$CO2, main = "CO2 udledning i ton per indbygger per land"
        , col = "blue", horizontal = TRUE
        , xlab = "CO2 udledning i ton per indbygger"
        , ylab = "Land", boxwex = 1.5)
dev.off()

#lande der ligger uden for 1.5 IRQ (potentielle outliers)
for (i in 1:length(co2$CO2)) {
  if (co2$CO2[i] > 6.6646 + (6.6646 - 0.828) *  1.5) {
    print(co2$Country[i])
    print(co2$CO2[i])
  }
}

# Deskriptiv analyse
#standardafvigelse
sd(co2$CO2)
summary(co2$CO2)

#standard error
sd(co2$CO2) / length(co2$CO2)
var(co2$CO2)
```