dplyr - Huistaak

Neerslaggegevens Klemskerke

De dataset bevat neerslaggegevens per uur van 01/01/2012 tot en met 20/02/2017. In een eerste kolom zit de datum en tijd van de meting, de tweede kolom bevat de hoeveelheid neerslag. De volgende kolommen bevatten nog informatie die voor deze oefening niet relevant is.

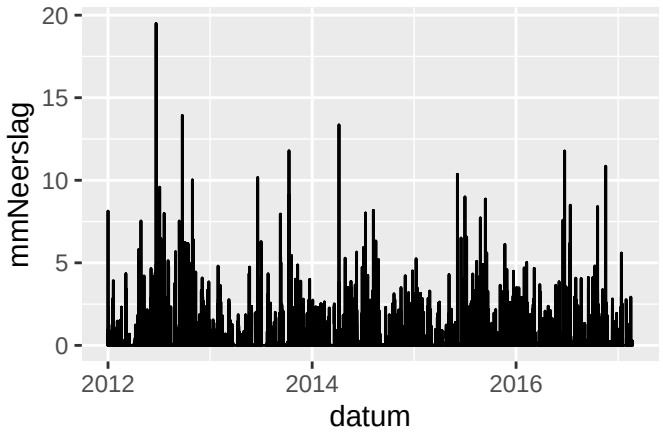
1. Lees de gegevens uit de file `20180123_rainfall_klemskerke.csv` in met het commando `read_csv2`.
2. Bekijk het resultaat van de vorige vraag. Is dit hetgeen je verwacht? Pas je code aan zodat de dataset proper ingelezen wordt.
3. Behoud enkel de kolommen met relevante informatie, en hernoem deze naar `datum` en `mmNeerslag`.
4. Welk formaat heeft de variabele `datum`? Aangezien we deze als datum willen gebruiken om te kunnen opsplitsen per jaar, maand, ... moeten we dit aanpassen naar een datum-formaat. Dat kan met het commando `as.POSIXct(datum)`. Met de commando's `year()` en `month()` uit het `lubridate` package kan je dan het jaar en de maand selecteren uit de datum.
   - Installeer het `lubridate` package
   - Laad het `lubridate` package
   - Verander het formaat van de variabele `datum` in een datum-formaat
   - Haal het jaar en de maand uit de variabele `datum`
5. Kopieer en run onderstaande functie in je script/Rmarkdown, en voeg daarna een nieuwe variabele `seizoen = weerkundig_seizoen(datum)` toe aan de dataset.

```
weerkundig_seizoen <- function(datum) {
  require(lubridate)
  md <- lubridate::month(datum)
  seizoen <-
    ifelse(md %in% c(12,1,2), "winter",
    ifelse(md %in% 3:5, "lente",
    ifelse(md %in% 6:8, "zomer",
    ifelse(md %in% 9:11, "herfst",
           NA))))
  #return
  factor(seizoen, levels = c("winter", "lente", "zomer", "herfst"))
}
```

   De structuur van je dataset zou er nu als volgt uit moeten zien:

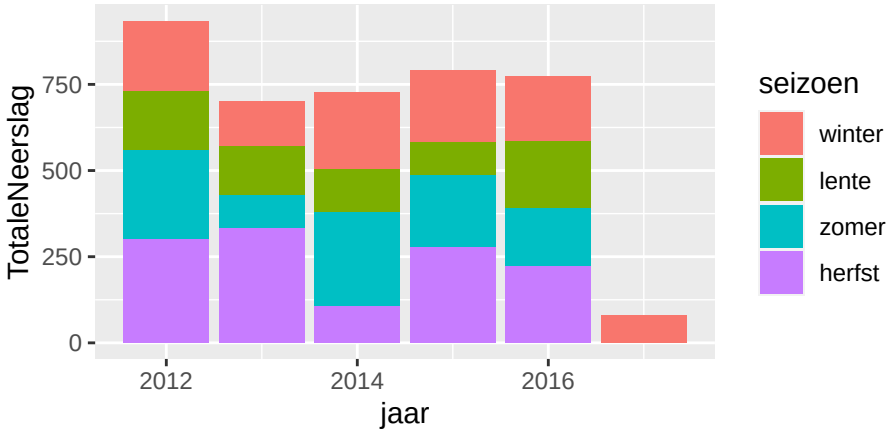
```
## tibble [45,064 x 5] (S3: tbl_df/tbl/data.frame)
##  $ datum     : POSIXct[1:45064], format: "2012-01-01" "2012-01-01" ...
##  $ mmNeerslag: num [1:45064] 0 0.95 0.72 0.2 0 0.27 0 0 0 0 ...
##  $ jaar      : num [1:45064] 2012 2012 2012 2012 2012 ...
##  $ maand     : num [1:45064] 1 1 1 1 1 1 1 1 1 1 ...
##  $ seizoen   : Factor w/ 4 levels "winter","lente",..: 1 1 1 1 1 1 1 1 1 1 ...
```

6. Vraag de `summary` op van de dataset. Wat valt op?
7. Maak de volgende lijnplot van de tijdreeks.



8. Op welk tijdstip (dag en uur) werd de meeste neerslag gemeten? Op basis van de grafiek kan je al een idee krijgen, maar geef ook een exacte datum, samen met de hoeveelheid neerslag.

9. Maak een barplot van de totale jaarlijkse neerslag, en kleur de balken volgens `seizoen`. Let op voor ontbrekende waarden in de variabele `mmNeerslag`. Kijk eens in de `help` van de functie `sum` hoe je dit kan oplossen. Raadpleeg ook zeker de `help` van `geom_bar()` bij problemen.

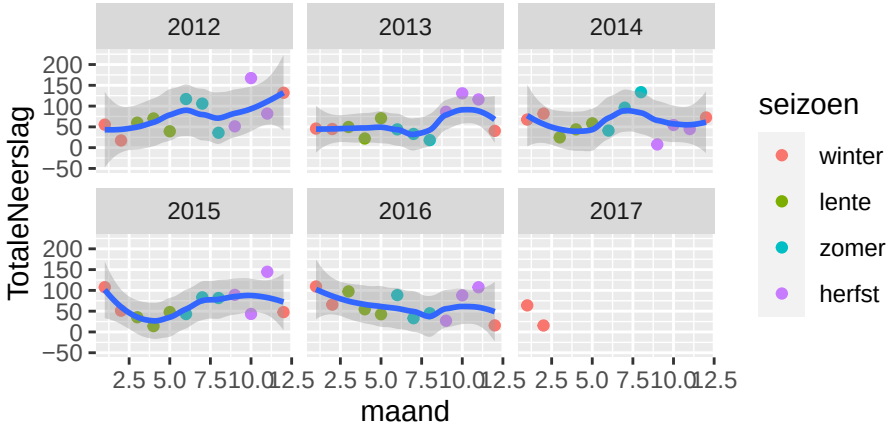


10. In welke 5 maanden viel de minste neerslag?

    - Bereken hiervoor de maandtotalen en -gemiddelden.
    - Vind je dezelfde 5 maanden terug op basis van het totaal als op basis van het gemiddelde? Verklaar.

11. Zijn er verschillen in gemiddelde neerslag tussen de seizoenen?

12. Maak een scatterplot van de totale neerslag per maand, opgesplitst over de verschillende jaren. Kleur de punten volgens seizoen en voeg een smoother toe.



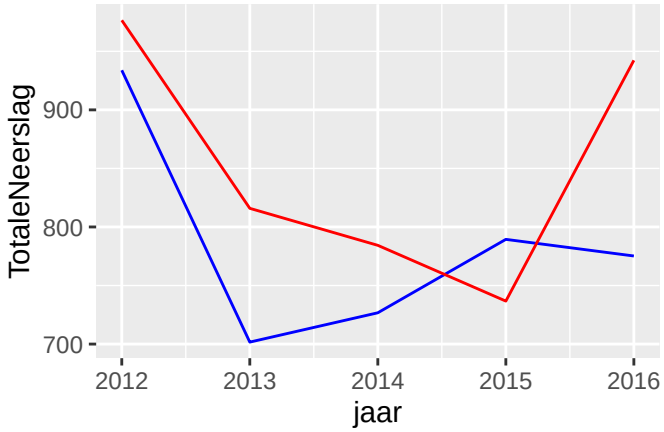
Neerslag België

In de dataset `Klimaatverandering - analyse neerslag jaar.xlsx` vind je de jaarlijkse neerslag in België terug voor de periode 1833 – 2017.

1. Lees deze dataset in met `read_excel`.
   - Specifieer zeker `sheet` en `range`.
   - Zorg daarna dat je enkel de lijn met `gemeten jaarneerslag` overhoudt.
   - Verwijder uiteindelijk de kolom met waarde `gemeten jaarneerslag`.
2. Deze dataset voldoet niet aan de vereisten van een *tidy* dataset. Pas deze aan zodat ze wel *tidy* wordt. Zorg dat het formaat van de variabelen correct is, en je de volgende structuur bekomt.

```
## tibble [185 x 2] (S3: tbl_df/tbl/data.frame)
##  $ jaar          : num [1:185] 1833 1834 1835 1836 1837 ...
##  $ TotaleNeerslag: num [1:185] 811 582 689 890 810 ...
```

3. Selecteer de jaren die ook in de Klemskerke dataset zitten.
4. Maak een grafiek met een lijn voor de totale jaarlijkse neerslag in België en in Klemskerke (tot en met 2016). Dit kan op verschillende manieren:
   - Gebruik de 2 datasets afzonderlijk



   - Plak de datasets onder mekaar met de functie `bind_rows()`. Je moet dan wel aan beiden eerst een variabele `locatie` (met de waarde `Klemskerke` of `Belgie`) toevoegen om te weten welke gegevens waarbij horen.