Political Data Science

Lektion 7: Visualisering

Undervist af Jesper Svejgaard, foråret 2018
Institut for Statskundskab, Københavns Universitet
github.com/jespersvejgaard/PDS

I dag

1. Status: Boblejagt
2. Opsamling fra sidst
3. Appetizer: Quanteda
4. Dagens pensum: Visualisering
5. Midtvejsevaluering
6. Workshop
7. Opsamling og næste gang

Overblik

1. Intro til kurset og R
2. R Workshop I: Explore
3. R Workshop II: Import, tidy, transform
4. R Workshop III: Programmering & Git
5. Web scraping & API
6. Tekst som data
7. Visualisering
8. GIS & spatiale data
9. Estimation & prædiktion
10. Superviseret læring I
11. Superviseret læring II
12. Usuperviseret læring
13. Refleksioner om data science
14. Opsamling og eksamen

Boblejagt

Leaderboard

KPI: Mest XP inden for de seneste 90 dage

Leader på hold 1: Niclas Nordsted

Leader på hold 2: Christoffer Cappelen

Opsamling fra sidst

- Tekst som data
- Overordnede tilgange: Klassifikation + skalering
- Overordnet sondering: Superviseret / usuperviseret

Opsamling fra sidst

Opgave: Sentiment-analyse af bigrammer fra 1256 artikler om sexual harassment fra The Guardian i perioden 01-01-2013 til 01-01-2018

- Er der forskel på ladningen af de ord, som følger efter hhv. "he" og "she"?
- Er der forskel på forskellene før/efter Weinstein-skandalen breakede?

Opsamling fra sidst

```
# Loader pakker
library(tidyverse)
library(tidytext)

# Importerer data
bigrams <- read_csv("https://raw.githubusercontent.com/jespersvejgaard/PDS/master/data/bigrams.

# Tjekker data ud
head(bigrams, n = 4)
```

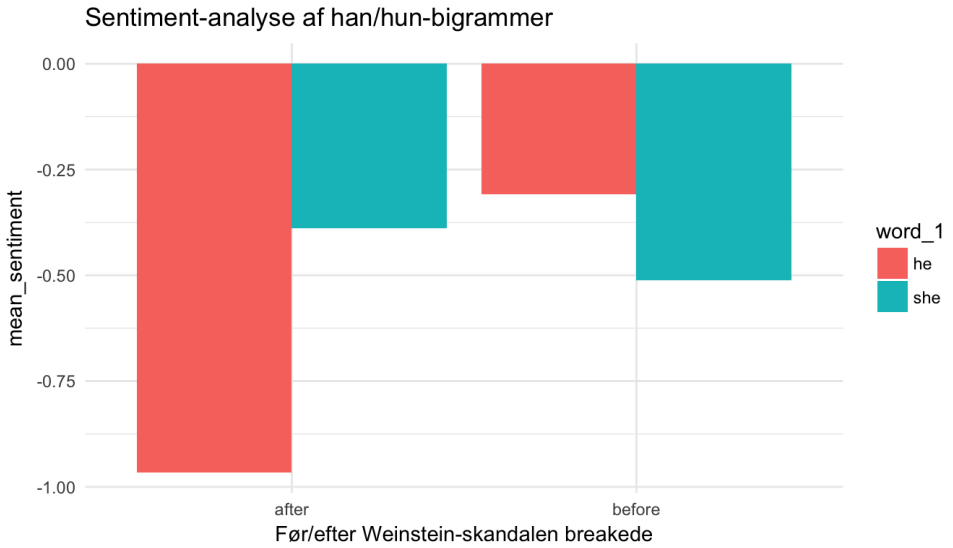
```
## # A tibble: 4 x 3
##   word_1 word_2 articles
##   <chr>  <chr>  <chr>
## 1 she    greets before
## 2 he     made   before
## 3 he     has    before
## 4 he     has    before
```

Opsamling fra sidst

Hvad sker der i koden her? Snak med sidemakkeren.

```
# Mulig løsning
bigrams %>%
  inner_join(get_sentiments("afinn"), by = c("word_2" = "word")) %>%
  group_by(articles, word_1) %>%
  summarise(mean_sentiment = mean(score)) %>%
  ggplot(aes(x = articles, y = mean_sentiment)) +
  geom_col(aes(fill = word_1), position = "dodge") +
  theme_minimal() +
  labs(x = "Før/efter Weinstein-skandalen breakede",
       title = "Sentiment-analyse af han/hun-bigrammer")
```

Opsamling fra sidst



Quanteda

Quanteda I

```
## Loader pakker
library(quanteda)
library(readtext)

## Definerer en vektor med tekstfiler
tekstfiler <- c("https://raw.githubusercontent.com/jespersvejgaard/PDS/master/data/rg1998.txt",
                "https://raw.githubusercontent.com/jespersvejgaard/PDS/master/data/rg2016.txt")

## Laver et korpus af tekster
korpus <- corpus(readtext(tekstfiler))

## Tjekker korpus
summary(korpus)
```

```
## Corpus consisting of 2 documents:
##
##   Text Types Tokens Sentences                     doc_id
##  text1  1854   7455       360 tmp.9X8Yol70qs/rg1998.txt
##  text2  4561  25238      1277 tmp.LMbJ84eLVq/rg2016.txt
##
## Source: /Users/jespersvejgaard/Desktop/Desktop/Akademiet/Political Data Science/PDS under co
```

Quanteda II

```
## Laver dfm: ændrer kapitaler, fjerner stopord, fjerner tegn, stemmer
tekster_dfm <- dfm(korpus,
                   tolower = TRUE,
                   remove = stopwords("danish"),
                   remove_punct = TRUE,
                   stem = TRUE)

## Tjekker et antal kolonner i dfm
tekster_dfm[, 20:26]
```

```
## Document-feature matrix of: 2 documents, 7 features (14.3% sparse).
## 2 x 7 sparse Matrix of class "dfm"
##        features
## docs    senest fem år præget fremdrift optimism fremgang
##   text1      2   1  9      1         2        1        7
##   text2     10   2 40      1         0        0        2
```

Quanteda III

```
## Visualiserer wordcloud
textplot_wordcloud(tekster_dfm, colors = RColorBrewer::brewer.pal(8,"Dark2"))
```

Visualisering

To formål - to tilgange

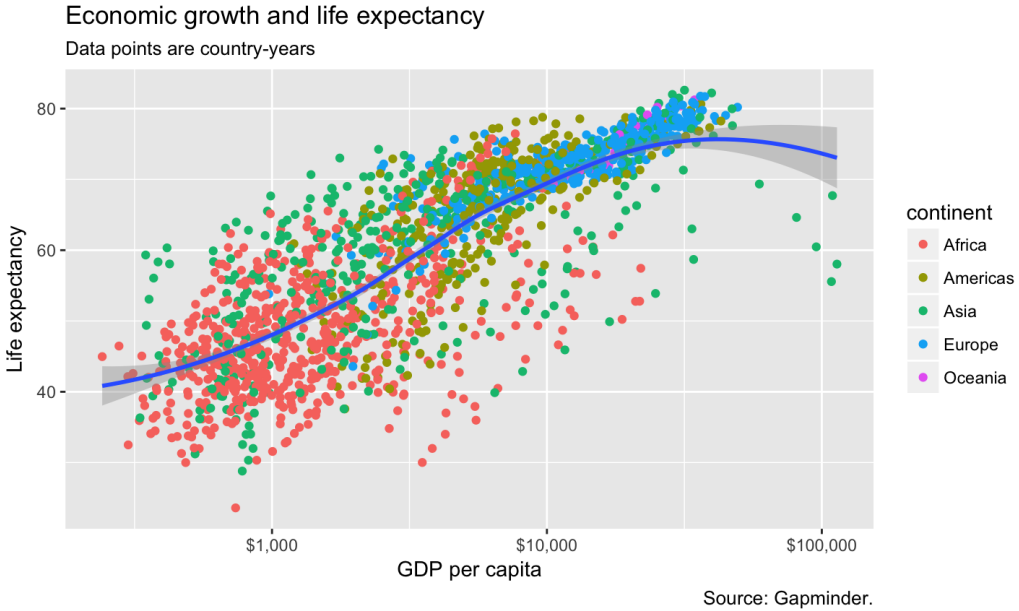
Exploratory

.

Explanatory

.

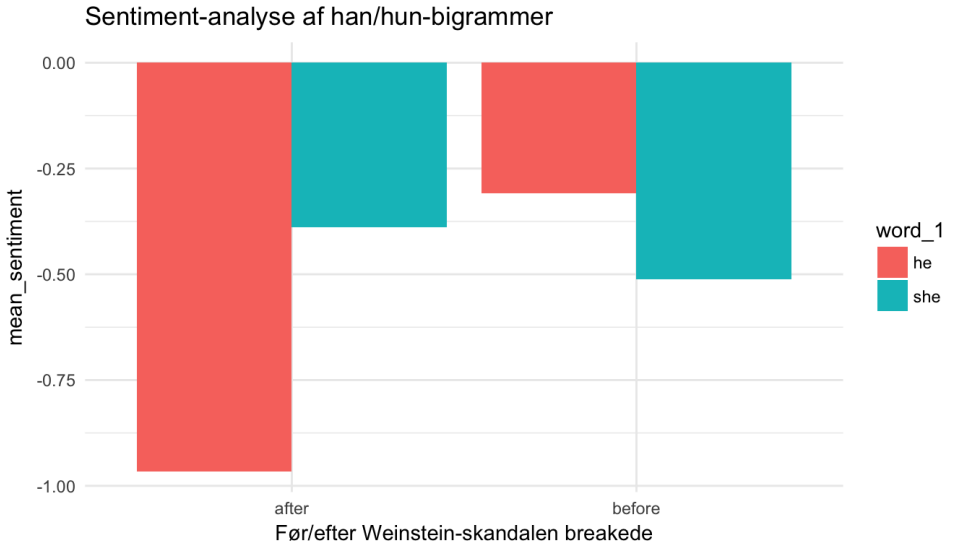
Der findes gode visualiseringer…



… Og der findes mindre gode

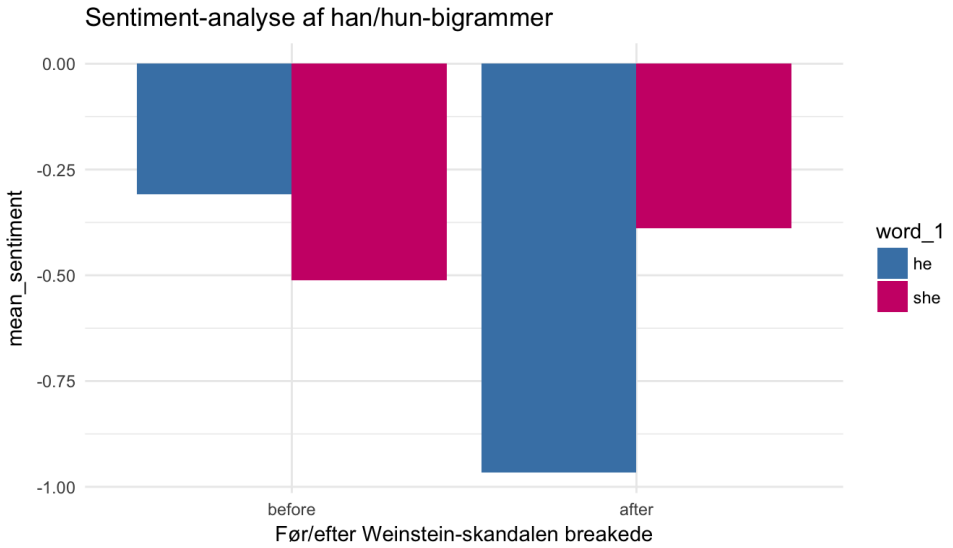
Eksempel

Hvad siger I til figuren fra før?



Eksempel

Nemmere at fortolke:



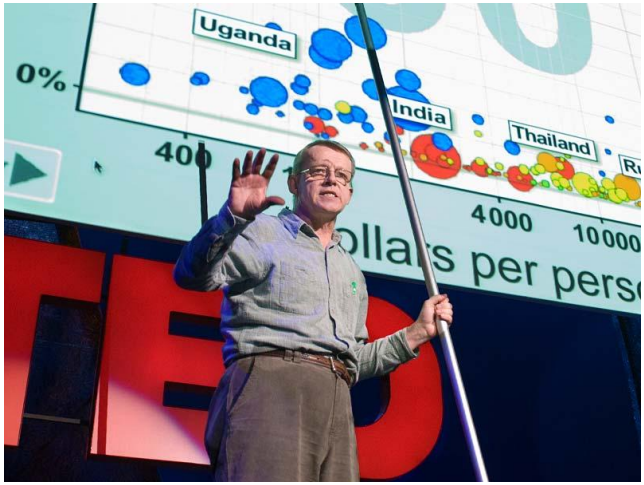
Motiverende eksempel

Trump's Fall From Grace in Hip-Hop

Tekst som data + visualisering

Fokus i dag

- Eksplorativ visualisering med `ggplot2`
- Eksemplificeret med gapminder

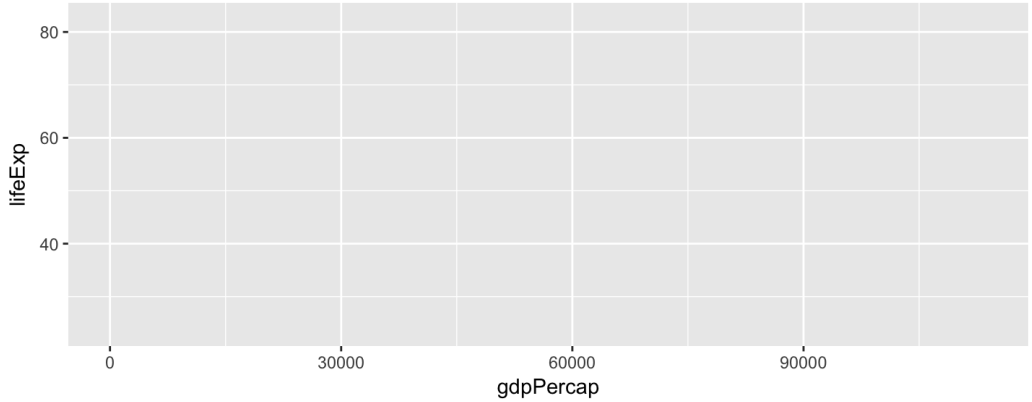


Opskrift: visualisering med `ggplot2`

1. Data: Fortæl ggplot() hvad vores data er (data =)
2. **Mapping:** Fortæl ggplot() hvilken relation vi vil se (mapping = aes())
3. **Geom:** Fortæl ggplot() hvordan vi vil se relationen i data (geom_*)
4. **Tilføj** evt. lag oven på geomet
5. **Juster** evt. plottet - skalaer, labels, titler, tick marks mm.

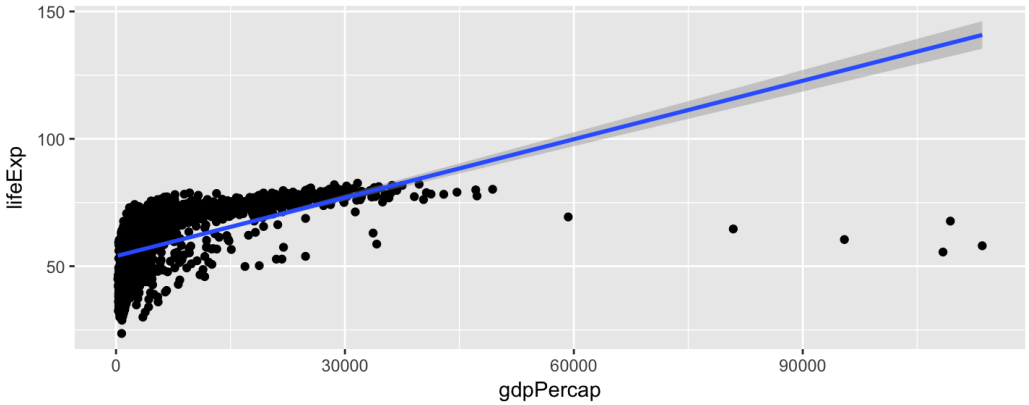
Data + mapping

```
## Kartesiansk koordinatsystem til plot
p <- ggplot(data = gapminder,
            mapping = aes(x = gdpPercap,
                          y = lifeExp))
```



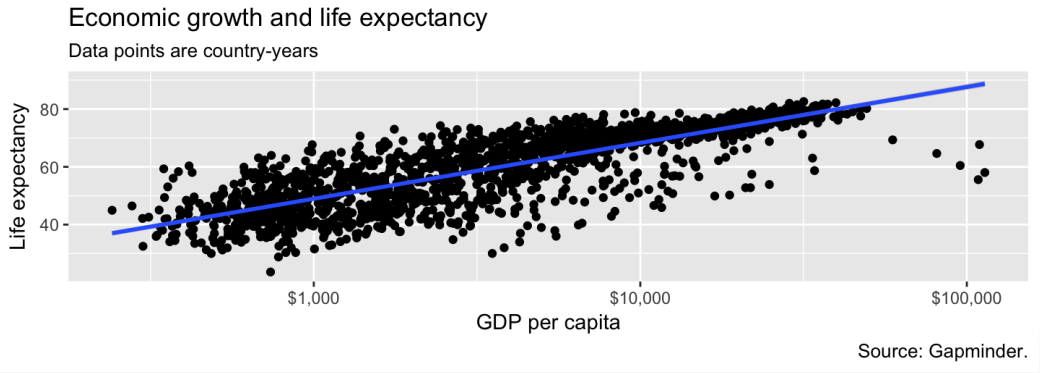
Geomer

```
## Tilføjer geomer
p + geom_point() +
    geom_smooth(method = "lm")
```



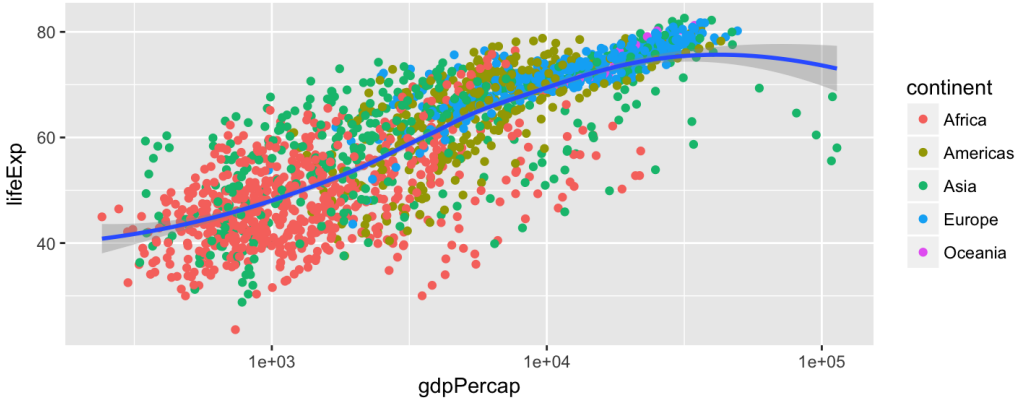
Justeringer

```
## Justerer skala og tilføjer labels
p + geom_point() +
    geom_smooth(method = "lm") +
    scale_x_log10(labels = scales::dollar) +
    labs(x = "GDP per capita",
         y = "Life expectancy",
         title = "Economic growth and life expectancy",
         subtitle = "Data points are country-years",
         caption = "Source: Gapminder.")
```



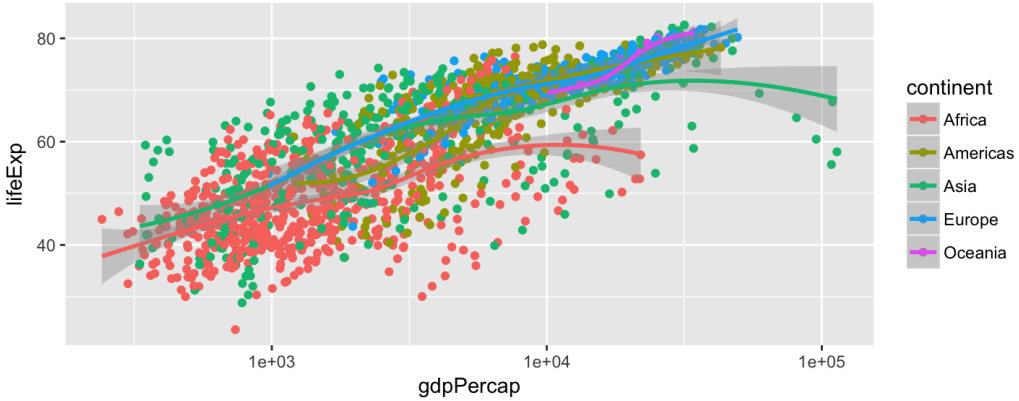
Color-æstetik (lokal)

```
## Lokal color-æstetik
ggplot(data = gapminder, mapping = aes(x = gdpPercap, y = lifeExp)) +
    geom_point(mapping = aes(color = continent)) +
    geom_smooth(method = "loess") +
    scale_x_log10()
```



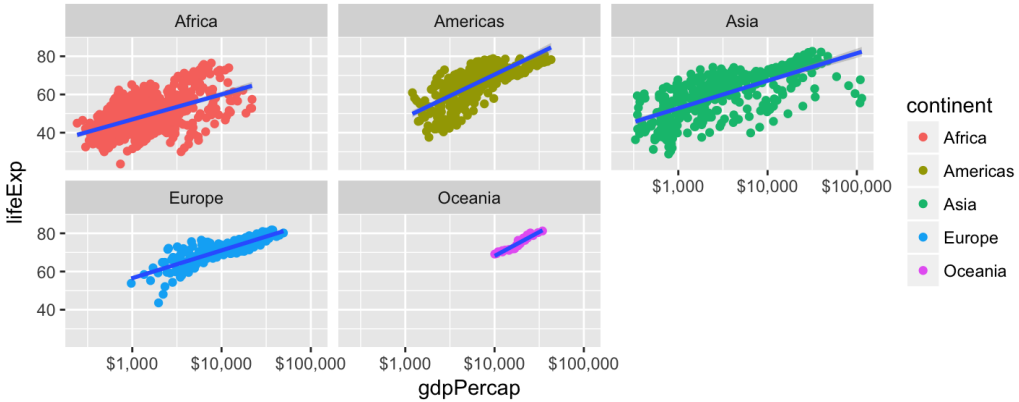
Color-æstetik (global)

```
## Global color-æstetik
ggplot(data = gapminder, mapping = aes(x = gdpPercap, y = lifeExp, color = continent)) +
    geom_point() +
    geom_smooth(method = "loess") +
    scale_x_log10()
```



Faceting

```
## Faceting
ggplot(data = gapminder, mapping = aes(x = gdpPercap, y = lifeExp)) +
    geom_point(mapping = aes(color = continent)) +
    geom_smooth(method = "lm") +
    scale_x_log10(labels = scales::dollar) +
    facet_wrap(~ continent)
```



Midtvejsevaluering

Midtvejsevaluering

Brug 5 minutter på at give Jesper feedback:

http://bit.ly/2FLdbCK

TAK <3

Workshop

Workshop

Find opgaverne i `07_opgaver.R` på Github under /PDS/opgaver.

Opsamling og næste gang

Vigtigste pointer fra i dag

- Quanteda, dfm og visualisering af tekster (også hip-hop-tekster)
- Gode og dårlige visualiseringer
- Opskrift på `ggplot2`: data, mapping, geomer, tilføjelser, justeringer
- Øvelser, hvor vi har brugt opskriften

```
ggplot(data = gapminder, mapping = aes(x = gdpPercap, y = lifeExp)) +
    geom_point(mapping = aes(color = continent)) +
    geom_smooth(method = "lm") +
    scale_x_log10(labels = scales::dollar) +
    facet_wrap(~ continent)
```

Næste gang - praktisk

Vi får fint besøg: Gæsteundervisning af Anders Woller

Sted: TBA

Tid: **Begge hold** tirsdag d. 27. marts **kl. 10 - 12**

Næste gang

- Indhold:
  - GIS og spatiale data
- Pensum:
  - DVSS (2017): kap 7 - læses til inspiration
  - Michalopoulos & Papaioannou (2013) - læs efter logikken/designet
- DataCamp:
  - Working with Geospatial Data in R

Tak for i dag!