Political Data Science

Lektion 2:

R Workshop I: Explore

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

I dag

1. Opsamling fra sidst
2. Overblik
3. Eksamensdatoer
4. Workflow i R
5. Styleguide
6. Highlights fra pensum
7. Workshop
8. Vigtigste pointer fra i dag
9. Næste gang

1. Opsamling fra sidst

- Hvad var de vigtigste pointer fra sidst?
- Gennemgang af opgaver (se `01_script.R` på GitHub)

2. 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

3. Eksamensdatoer

- 28/05-18: Frist - første indlevering
- 31/05-18: Her får man at vide, hvis man ikke er bestået første indlevering
- 05/06-18: Frist - anden indlevering
- 07/06-18: Her får man at vide, hvis man ikke er bestået anden indlevering
- 11/06-18: Frist - tredje indlevering

4. Workflow i R

- At kode: Scripts vs. konsollen
- At gemme: Scripts vs. environment
- Funktioner
- Musen :(
- Brug genveje:
  - run: cmd+enter / ctrl+enter
  - run, but stay: option+enter / alt+enter
  - rerun script: cmd+shift+s / ctrl+shift+s
  - pipe: cmd+shift+m / ctrl+shift+m
  - kommentér: cmd+shift+c / ctrl+shift+c
  - gem, fortryd, kopier, sæt ind etc
  - se flere her

5. Styleguide

– Phil Karlton

- Navngivning
  - alllowercase
  - period.separated
  - underscore_separated
  - lowerCamelCase, UpperCamelCase
  - vigtigst: konsistens!
- Dokumentation af kode
- Styleguides
  - Hadley Wickham's style guide
  - Google's R style guide

6. Highlights fra pensum

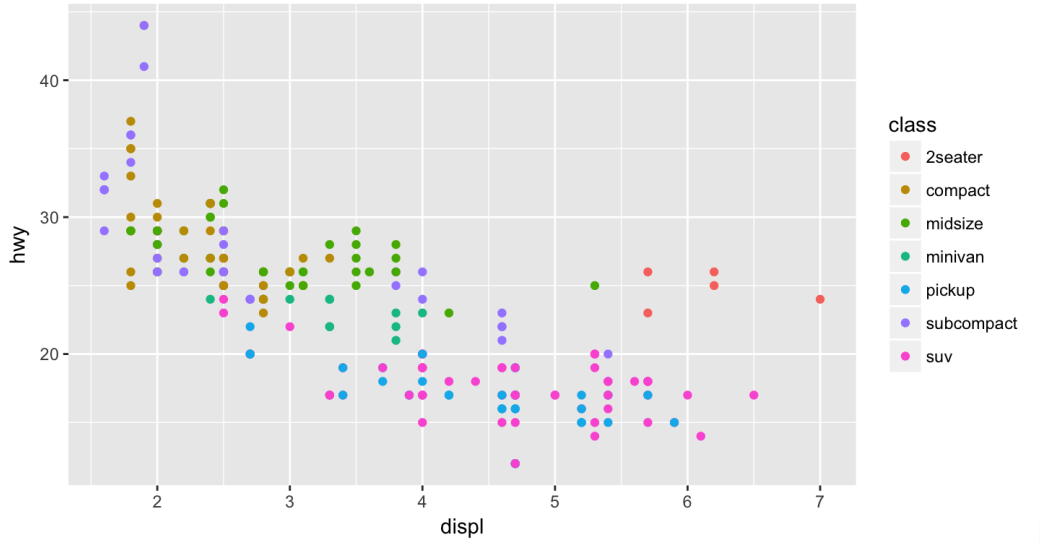
ggplot2

```
## # A tibble: 4 x 11
##   manufacturer model displ  year   cyl      trans   drv   cty   hwy    fl
##          <chr> <chr> <dbl> <int> <int>      <chr> <chr> <int> <int> <chr>
## 1         audi    a4   1.8  1999     4   auto(l5)     f    18    29     p
## 2         audi    a4   1.8  1999     4 manual(m5)     f    21    29     p
## 3         audi    a4   2.0  2008     4 manual(m6)     f    20    31     p
## 4         audi    a4   2.0  2008     4   auto(av)     f    21    30     p
## # ... with 1 more variables: class <chr>
```

6. Highlights fra pensum

ggplot2

```
ggplot(data = mpg, aes(x = displ, y = hwy, color = class)) +
  geom_point()
```



6. Highlights fra pensum

ggplot2

Hvad er forskellen mellem nedenstående?

```
ggplot(data = mpg, aes(x = displ, y = hwy)) +
  geom_point() +
  geom_smooth()

ggplot(data = mpg) +
  geom_point(aes(x = displ, y = hwy)) +
  geom_smooth(aes(x = displ, y = hwy))
```

6. Highlights fra pensum

filter()

```
mpg_1999 <- filter(mpg, year == 1999)

print(mpg_1999)
```

```
## # A tibble: 117 x 11
##    manufacturer                model displ  year   cyl      trans   drv
##           <chr>                <chr> <dbl> <int> <int>      <chr> <chr>
##  1         audi                   a4   1.8  1999     4   auto(l5)     f
##  2         audi                   a4   1.8  1999     4 manual(m5)     f
##  3         audi                   a4   2.8  1999     6   auto(l5)     f
##  4         audi                   a4   2.8  1999     6 manual(m5)     f
##  5         audi            a4 quattro   1.8  1999     4 manual(m5)     4
##  6         audi            a4 quattro   1.8  1999     4   auto(l5)     4
##  7         audi            a4 quattro   2.8  1999     6   auto(l5)     4
##  8         audi            a4 quattro   2.8  1999     6 manual(m5)     4
##  9         audi            a6 quattro   2.8  1999     6   auto(l5)     4
## 10    chevrolet c1500 suburban 2wd   5.7  1999     8   auto(l4)     r
## # ... with 107 more rows, and 4 more variables: cty <int>, hwy <int>,
## #   fl <chr>, class <chr>
```

6. Highlights fra pensum

filter()

```
manufac_japan <- c("honda", "toyota", "nissan")

mpg_japan <- mpg %>%
  filter(manufacturer %in% manufac_japan)

print(mpg_japan)
```

```
## # A tibble: 56 x 11
##    manufacturer  model displ  year   cyl      trans   drv   cty   hwy
##           <chr>  <chr> <dbl> <int> <int>      <chr> <chr> <int> <int>
##  1        honda  civic   1.6  1999     4 manual(m5)     f    28    33
##  2        honda  civic   1.6  1999     4   auto(l4)     f    24    32
##  3        honda  civic   1.6  1999     4 manual(m5)     f    25    32
##  4        honda  civic   1.6  1999     4 manual(m5)     f    23    29
##  5        honda  civic   1.6  1999     4   auto(l4)     f    24    32
##  6        honda  civic   1.8  2008     4 manual(m5)     f    26    34
##  7        honda  civic   1.8  2008     4   auto(l5)     f    25    36
##  8        honda  civic   1.8  2008     4   auto(l5)     f    24    36
##  9        honda  civic   2.0  2008     4 manual(m6)     f    21    29
```

6. Highlights fra pensum

arrange()

```
arrange(mpg, hwy)
```

```
arrange(mpg, desc(year))
```

```
## # A tibble: 234 x 11
##    manufacturer                model displ  year   cyl      trans   drv
##           <chr>                <chr> <dbl> <int> <int>      <chr> <chr>
##  1         audi                   a4   2.0  2008     4 manual(m6)     f
##  2         audi                   a4   2.0  2008     4   auto(av)     f
##  3         audi                   a4   3.1  2008     6   auto(av)     f
##  4         audi          a4 quattro   2.0  2008     4 manual(m6)     4
##  5         audi          a4 quattro   2.0  2008     4   auto(s6)     4
##  6         audi          a4 quattro   3.1  2008     6   auto(s6)     4
##  7         audi          a4 quattro   3.1  2008     6 manual(m6)     4
##  8         audi          a6 quattro   3.1  2008     6   auto(s6)     4
##  9         audi          a6 quattro   4.2  2008     8   auto(s6)     4
## 10    chevrolet c1500 suburban 2wd   5.3  2008     8   auto(l4)     r
```

6. Highlights fra pensum

select()

```
select(mpg, manufacturer:model, year, displ, contains("y"))
```

```
## # A tibble: 234 x 7
##    manufacturer      model  year displ   cyl   cty   hwy
##           <chr>      <chr> <int> <dbl> <int> <int> <int>
##  1         audi         a4  1999   1.8     4    18    29
##  2         audi         a4  1999   1.8     4    21    29
##  3         audi         a4  2008   2.0     4    20    31
##  4         audi         a4  2008   2.0     4    21    30
##  5         audi         a4  1999   2.8     6    16    26
##  6         audi         a4  1999   2.8     6    18    26
##  7         audi         a4  2008   3.1     6    18    27
##  8         audi a4 quattro  1999   1.8     4    18    26
##  9         audi a4 quattro  1999   1.8     4    16    25
## 10         audi a4 quattro  2008   2.0     4    20    28
## # ... with 224 more rows
```

6. Highlights fra pensum

mutate()

```
mpg_edt <- mpg %>%
  mutate(gallons_avg = (cty + hwy)/2,
         engine_size = ifelse(displ > 3, "big", "small"))
```

6. Highlights fra pensum

mutate()

```
glimpse(mpg_edt)
```

```
## Observations: 234
## Variables: 7
## $ manufacturer <chr> "audi", "audi", "audi", "audi", "audi", "audi", "...
## $ model        <chr> "a4", "a4", "a4", "a4", "a4", "a4", "a4", "a4 qua...
## $ displ        <dbl> 1.8, 1.8, 2.0, 2.0, 2.8, 2.8, 3.1, 1.8, 1.8, 2.0,...
## $ cty          <int> 18, 21, 20, 21, 16, 18, 18, 18, 16, 20, 19, 15, 1...
## $ hwy          <int> 29, 29, 31, 30, 26, 26, 27, 26, 25, 28, 27, 25, 2...
## $ gallons_avg  <dbl> 23.5, 25.0, 25.5, 25.5, 21.0, 22.0, 22.5, 22.0, 2...
## $ engine_size  <chr> "small", "small", "small", "small", "small", "sma...
```

6. Highlights fra pensum

summarize()

```
mpg_stats <- mpg %>%
  mutate(gallons_avg = (cty + hwy)/2) %>%
  group_by(manufacturer) %>%
  summarize(gallons_avg = mean(gallons_avg, na.rm = T),
            models = n_distinct(model)) %>%
  arrange(desc(gallons_avg))

print(mpg_stats)
```

```
## # A tibble: 15 x 3
##    manufacturer gallons_avg models
##           <chr>       <dbl>  <int>
##  1        honda    28.50000      1
##  2   volkswagen    25.07407      4
##  3      hyundai    22.75000      2
##  4       subaru    22.42857      2
##  5         audi    22.02778      3
##  6       toyota    21.72059      6
##  7      pontiac    21.70000      1
```

7. Workshop

- Find opgaverne i `02_opgaver.pdf` på GitHub

8. Vigtigste pointer fra i dag

- ggplot() - som tager tre argumenter: data + mapping + lag
- filter()
- arrange()
- select()
- mutate()
- summarize()
- group_by()

9. Næste gang

- Indhold:
  - R Workshop II: Import, tidy, transform
- Pensum:
  - R4DS: kap 9 - 13
  - CS: Data import
- DataCamp:
  - Cleaning data in R
  - Data manipulation in R with dplyr
  - Joining data in R with dplyr
- Supplerende:
  - R4DS: kap 14 - 16
  - Spachtholz (2017)