# Political Data Science

Lektion 2

*R Workshop I: Explore*

## Opgave 1

1. Installer og load pakken `nycflights13`. Du kan nu behandle data framen `flights`.
2. Er 'flights' tidy? Hvorfor / hvorfor ikke?
3. Hvad er udgør en enhed i datasættet?
4. Hvad er dimensionerne på datasættet? Hvad betyder det?
5. Vis dimensionerne på datasættet på to andre måder.

## Opgave 2 - arrange()

1. Hvornår afgik det sidste fly i datasættet?
2. Hvornår afgik det første?
3. Hvor forsinket var de 10 mest forsinkede fly tilsammen? [hint: head()]

## Opgave 3 - filter()

1. Hvornår har du fødselsdag? Tillykke! Hvor mange fly afgik der på din fødselsdag?
2. Find alle afgange der:
   - Var mere end 2 timer forsinkede
   - Blev opereret af selskaberne AA eller OO
   - Afgik i december
   - Afgik forsinket, men ankom til tiden
3. Hvor mange afgange har missing i `dep_time`?

## Opgave 4 - select()

1. Lav en dataframe, der indeholder alle variable fra `flights`, undtaget `tailnum`
2. Lav en dataframe, der kun indeholder variablene `year`, `month`, `day`, `dep_time og àrr_delay`
3. Hvad sker der, hvis du eksekverer koden `select(flights, contains("TIME"))`?

## Opgave 5 - mutate()

1. Brug `dep_time` til at lave to nye variable, `dep_hour` og `dep_minute` [hint: `%/%` og `%%`]
2. Lav en variabel, `gain`, som er `dep_delay` fratrukket `arr_delay` , dvs. den tid, der bliver indhentet i luften af de forsinkede fly.
3. Lav en variabel, `hours`, som er `àir_time` omregnet fra minutter til timer
4. Lav en variabel, `gain_per_hour`, som er `gain` delt med `hours`
5. Lav én mutate, hvor du konstruerer variablen `gain_per_hour` inden for den samme `mutate()`
6. Hvad sker der, hvis du bruger `transmute()` frem for `mutate()`?

## Opgave 6 - summarize()

1. Hvilket luftfartsselskab har den højeste ankomst-forsinkelse i gennemsnit? [hint: `group_by()`]
2. Hvilken måned er afgangs-forsinkelsen højest?
3. Hvilke destinationer kan man flyve til/fra med flest forskellige luftfartsselskaber? [hint: `n_distinct()`]

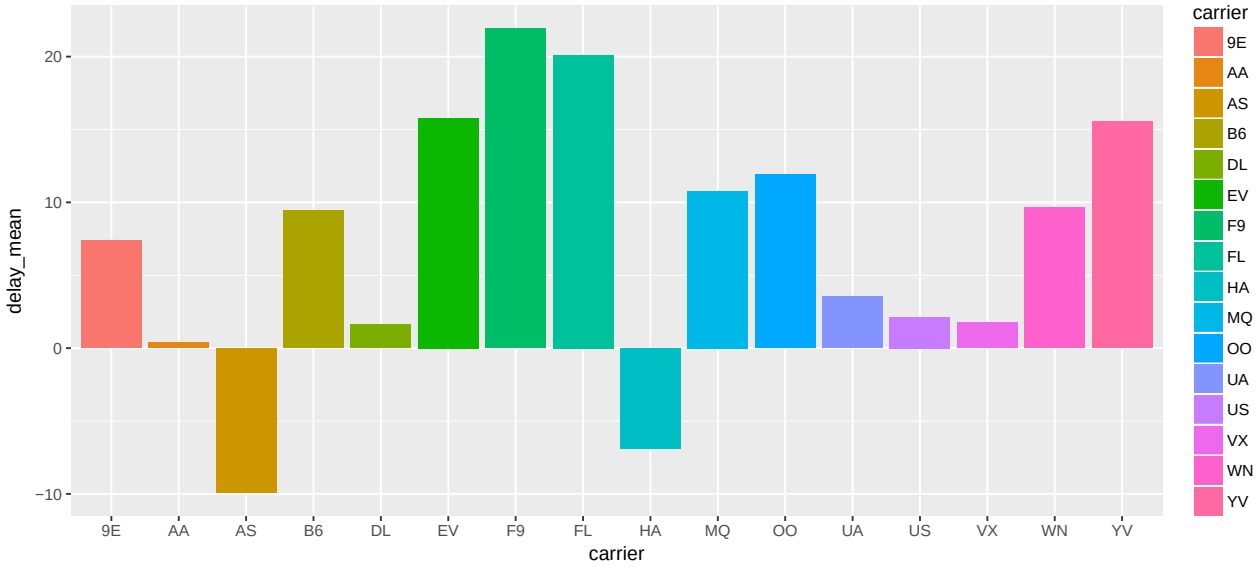
## Opgave 7

Tabellen herunder viser den gennemsnitlige ankomst-forsinkelse fordelt på årets måneder. Rekonstruer tabellen ved brug af `group_by()` og `summarize`.

```
## # A tibble: 12 x 4
## # Groups:   year [?]
##     year month delay_mean     n
##    <int> <int>      <dbl> <int>
##  1  2013     1  6.1299720 27004
##  2  2013     2  5.6130194 24951
##  3  2013     3  5.8075765 28834
##  4  2013     4 11.1760630 28330
##  5  2013     5  3.5215088 28796
##  6  2013     6 16.4813296 28243
##  7  2013     7 16.7113067 29425
##  8  2013     8  6.0406524 29327
##  9  2013     9 -4.0183636 27574
## 10  2013    10 -0.1670627 28889
## 11  2013    11  0.4613474 27268
## 12  2013    12 14.8703553 28135
```

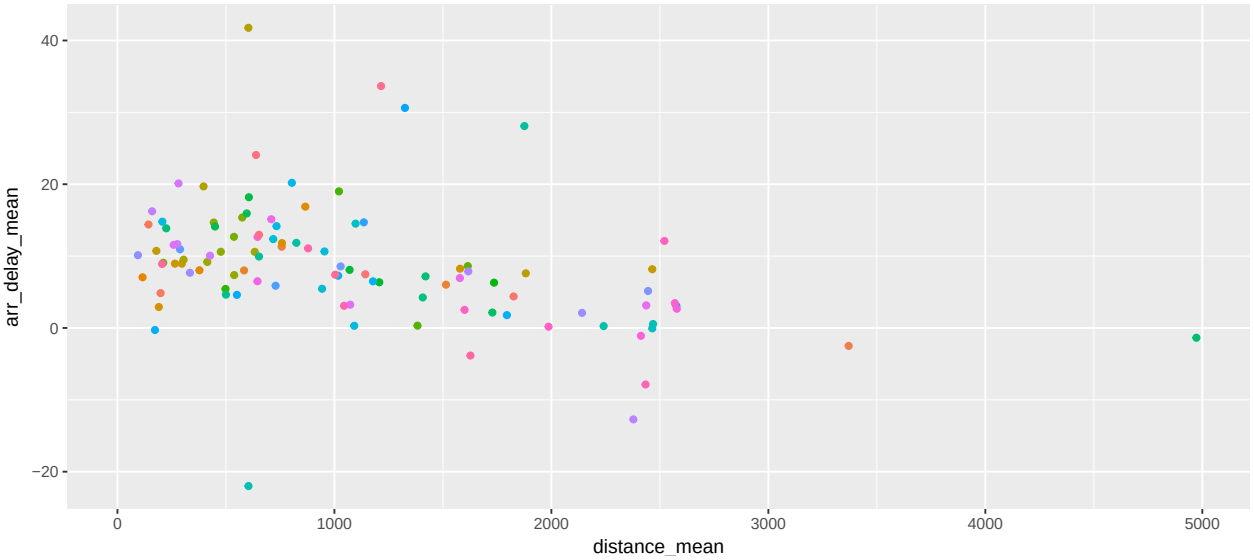
## Opgave 8

Figuren herunder viser flyselskabernes gennemsnitlige ankomstforsinkelse. Rekonstruer figuren.



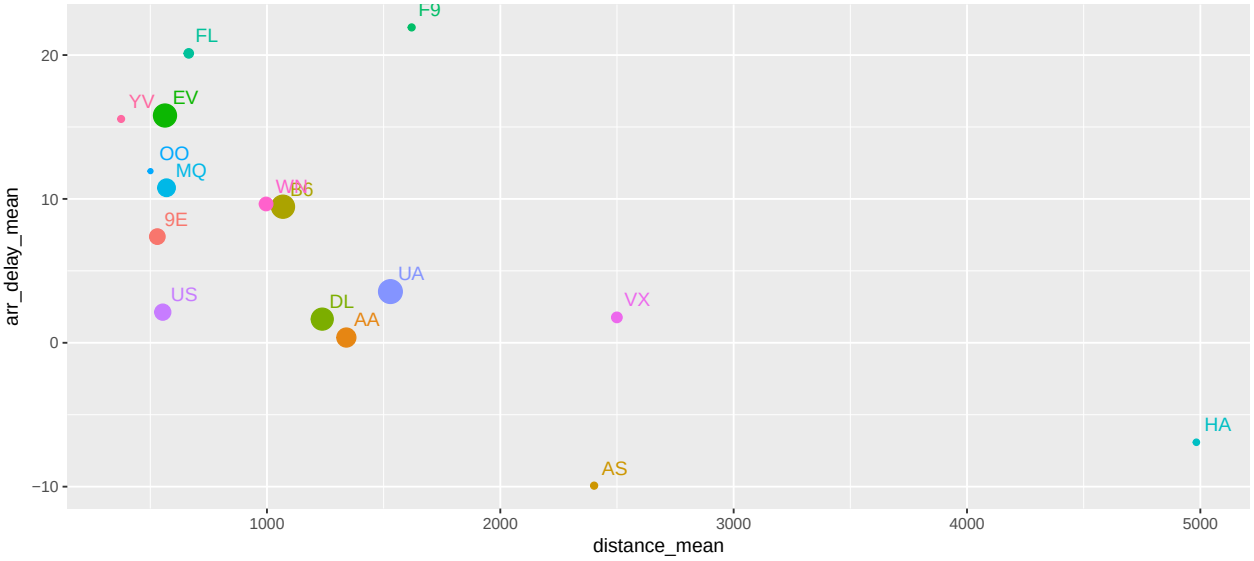
## Opgave 9

Figuren herunder viser sammenhængen mellem den gennemsnitlige distance og den gennemsnitlige ankomst-forsinkelse for flyvninger til de respektive lufthavne i `flights`. Hvilke lufthavne er det, der stikker ud i figuren herunder?




## Opgave 10

Figuren herunder viser den gennemsnitlige distance og den gennemsnitlige ankomst-forsinkelse for flyselskaberne i `flights`. Rekonstruer figuren.




## Opgave 11

Lav et plot, der kommunikerer en ny indsigt fra `flights`-datasættet, fx et histogram, et line-plot eller et box plot.