

## Instruksjoner

- Prøven skal besvares med et fungerende R-script som lastes opp i innleveringsmappen på Fronter.
- Scriptet skal inneholde nødvendig kode for å besvare oppgavene samt kommentarer markert med `#` som forklarer fremgangsmåten dere har valgt. Der oppgavene ber dere oppgi bestemte verdier eller tolkninger skal disse også oppgis som kommentarer i scriptet.
- Sørg for at koden er oversiktlig. For å skille oppgavene fra hverandre, anbefales overskrifter av typen:  
`### Oppgave 1 #####` i scriptet  
`# Oppgave 1:`  
eller lignende.
- Lykke til!

## Oppgaver

1. Last inn datasettet **trump\_tweets.csv**. Enhetene i datasettet twittermeldinger postet av Donal Trump i tidsrommet 2013-2016. Du kan enten laste inn data direkte fra linken:  
`http://folk.uio.no/martigso/baaap/trump_tweets.csv`  
eller laste ned fra Fronter og inn i R fra lokalt på din data.
2. Opprett en dikotom variabel – **obama\_nevnt** – som indikerer om Obama ble nevnt i tweeten (1) eller ikke (0).
3. Vis med et barplot hvor mange meldinger Trump nevnte og ikke nevnte Obama i. Kommenter kort eventuelle problemer med å bruke målet i en regresjonsanalyse.
4. Opprett en ny variabel – **sentiment** – som er summen av positive og negative ord (**positive\_ord** og **negative\_ord**) som proporsjon av totalt antall ord (**totalt\_antall\_ord**) i en twittermelding.
5. Lag et subset av data, der du fjerner radene med twittermeldinger som ikke har sentiment (0 på sentimentvariabelen).
6. Lag et bivariat regresjonsplot fra subsettet av data, med proposjon sentiment (**sentiment**) på x-aksen og antall personer som har merket tweeten som “favoritt” på y-aksen. Kommenter kort ett problem med å gjøre slutning fra dette plottet.
7. Estimér en OLS-modell fra det fulle datasettet, med **favoritter** som avhengig variabel, og **pr\_sentiment**, **obama\_nevnt** og **tidspunkt** som uavhengige variabler. Tolk effekten av natlige tweets substansielt.
8. Estimér samme OLS-modell som over, men kontroller også for fixed effects (ordinal/-factor) av år (**aar**). Kommenter kort forskjellen i forklart varians ( $R^2$ ) mellom de to modellene.

## Variabelforklaringer:

**id** Unik id for hver twittermelding

**tekst** Tekst for twittermelding

**dato** Dato for posting av tweet

**aar** År for posting av tweet

**tidspunkt** Tidspunkt på dagen for posting av tweet (morgen, dag, kveld, natt)

**favoritter** Antall personer som har trykket favorite på tweeten

**retweets** Antall personer som har retweetet tweeten

**positive\_ord** Antall positive ord i teksten til tweeten

**negative\_ord** Antall negative ord i teksten til tweeten

**totalt\_antall\_ord** Totalt antall ord i teksten til tweeten