

Testprøve

Martin Søyland

April 3, 2017

Instruksjon

- Oppgavene skal løses i et R-skript, og kopieres inn i Socrative.
- Skriptet 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, skal disse også oppgis som kommentarer i skriptet.
- Sørg for at koden er oversiktlig. For å skille oppgavene fra hverandre, anbefaler vi overskrifter av typen: ### Oppgave 1 #### i skriptet.

Oppgaver

Du har fått i oppgave å analysere reaksjoner på Donald Trumps statuser gjennom valgkampen i 2016. Dataene du får du fra http://folk.uio.no/martigso/stv1020/testprove/trump_face.csv. Variabelbeskrivelse kan du finne under oppgavene

1. Last inn data "trump_face.csv"
2. Opprett en ny variabel, "clinton_mentioned", som tar verdien 1 hvis variabelen "status_message" inneholder ordet "Clinton", og 0 ellers. Hvor mange statuser ble Clinton nevnt i? (Hint: Se oppgave 2a fra seminar 4).
3. Opprett to variabler som viser..
 - a. Prosent likes ("num_likes") av alle reaksjoner ("num_reactions")
 - b. Prosent sinte reaksjoner ("num_angrys") av alle reaksjoner ("num_reactions")
4. Lag et boxplot som viser proporsjonen likes (fra oppgave 3) på y-aksen og tidspunkt statusen ble delt på x-aksen ("time_of_day")
 - a. Hvilket tidspunkt viser boxblotet lavest median på?
5. Lag et subset av data (nytt datasett), der bare statuser som er publisert på natten ("night") er med
 - a. Lag en korrelasjonsmatrise fra det originale datasettet (fra oppgave 1) mellom variablene fra oppgave 3 (prosent likes og angry), antall negative ord ("negative_words") og antall positive ord ("positive_words"). Slett enheter list-wise.
 - b. Lag samme korrelasjonsmatrise som over, men på det nye datasettet med bare nattlige statuser. Slett enheter list-wise.
 - c. Signifikanstest en av korrelasjonene med begge datasettene. Kommenter kort resultatene.
6. Lag en regresjon fra det komplette datasettet, med antal sinte reaksjoner ("num_angrys") som avhengig variabel og "positive_words", "negative_words", om Clinton er nevnt, og dummysett av "time_of_day" som uavhengige variabler
 - a. Hvor mange flere sinte reaksjoner kan vi forvente Trump får om han nevner Clinton i statusen sin?
 - b. Sjekk om forutsetningen om normalfordelte restledd er oppfylt. Kommenter resultatet kort.
7. Gjør samme regresjon som oppgave 6, men nå med prosent sinte reaksjoner som avhengig variabel.
 - a. Sjekk om forutsetningen om normalfordelte restledd er oppfylt. Kommenter forskjellen fra oppgave 6a

Variabelforklaringer:

Variabelnavn	Beskrivelse
status_id	Id for Facebook status

Variabelnavn	Beskrivelse
status_message	Status tekst
link_name	Navn på den det linkes til
status_type	Type status
status_link	Link til status
status_published	Dato for statusens publisasjon
num_reactions	Totalt antall reaksjoner
num_comments	Antall kommentarer
num_shares	Antall delinger
num_likes	Antall likes
num_loves	Antall “love”-reaksjoner
num_wows	Antall “wow”-reaksjoner
num_hahas	Antall “haha”-reaksjoner
num_sads	Antall “sad”-reaksjoner
num_angrys	Antall “angry”-reaksjoner
positive_words	Antall positive ord i status
negative_words	Antall negative ord i status
total_words	Totalt antall ord i status
time_of_day	Tidspunkt for publisasjon av status