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.
- Husk at riktig kode er det viktigste; pass på å ikke bruk for lang tid på tolkninger.
- 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!

Variabelforklaringer:

respondent_id Unik id for individuelle respondenter

steak_prep Hvordan respondenten foretrekker biff stekt

hhold_income Husholdningsinntekten til respondenten (i dollar)

age Respondentens alder

smoke Om respondenten røyker (1 = ja, 0 = nei)

alcohol Om respondenten drikker alkohol (1 = ja, 0 = nei)

Oppgaver

1. Last inn data **steak_survey.csv**. Enhetene i datasettet er respondenter i en survey. Du kan enten laste ned data fra fronter, linken:
   `http://folk.uio.no/martigso/encrypt/`
   eller direkte inn i R med:
   `http://folk.uio.no/martigso/encrypt/steak_survey.csv`
2. Lag et stolpediagram av variabelen `steak_prep`. Kommenter hvilken verdi på variabelen som har høyest frekvens.
3. Lag en **ny variabel** – `steak_prep2` – som tar verdiene:

   - "Rare" når `steak_prep` er "Rare" **eller** "Medium rare"
   - "Medium" når `steak_prep` er "Medium"
   - "Well" når `steak_prep` er "Medium Well" **eller** "Well"

   Sjekk at variabelen ble kodet riktig med en tabell.
4. Gjør om den nye `steak_prep2` variabelen til en faktor, og sett kategorien med flest enheter til referansekategori.
5. Vis hvordan du finner korrelasjonen mellom variblene `smoke` og `alcohol`. Oppgi korrelasjonen i en kommentar.
6. Lag et boxplot med `steak_prep2` på x-aksen og `age` på y-aksen. Hvilken kategori på `steak_prep2` har lavest median?
7. Estimer en multinomisk logistisk regresjon med `steak_prep2` som avhengig variabel og `age`, `hhold_income`, `smoke` og `alcohol` som uavhengige variabler. Husk også å ta vare på informasjon om NA i regresjonen. Kommenter kort hva retningen for begge koeffisientene til `smoke` betyr.
8. Vis hvordan du sjekker konfidensintervallene på 5% nivå for effektene i regresjonen fra oppgave 6. Er effekten av `age` signifikant?
9. Legg inn predikerte **kategorier** (ikke sannsynligheter) i datasettet fra regresjonen i oppgave 6. Lag en tabell over predikerte (forventede) og faktiske verdier på `steak_prep2`. Kommenter kort hva tabellen viser.
10. Lag datasett (*test set*) der alder går fra 18:90, `hhold_income` er satt til median, `smoke` er satt til 0 og `alcohol` er satt til 1. Legg så inn predikerte sannsynligheter (løs regresjonsligningen) fra regresjonen (oppgave 7) i dette datasettet. Lag deretter et plot som har de forventede sannsynlighetene til *test settet* på y-aksen, alder på x-aksen og fargede linjer for hver av kategoriene på `steak_prep2`.