SOK-1004 Høst 2021

Prosjektoppgave for studenter på kull 2020 og tidligere.

Leveres i WiseFlow innen 15. desember 2021, kl. 16.00. Besvarelsen kan skrives individuelt, eller i grupper på maksimum 3 studenter.

Prosjektoppgaven består av to deler. Del A er et case fra Recogni. Recogni skal ikke kontaktes under eksamen. Del B er basert på inntektsdata i Troms fra 2015. Alle oppgaver skal besvares. Du skal benytte ggplot til alle figurer. Alle figurer skal ha meningsfulle titler, 'labels', benevninger og enheter på aksene.

Del A - Skipsdata

Dette er data fra et business case. Et norsk rederi eier et skip som opereres av to ulike mannskap med lik turnus. Hvert mannskap har sine egne arbeidsmetoder og tar i stor grad egne beslutninger om bord. Skipet er av nyere dato, og har et automasjonssystem som samler sensordata fra alle skipets systemer i sanntid. Rederiets ledelse mistenker at skipet ikke opereres optimalt, og har derfor tatt ut driftsdata med håp om å finne svar. De ønsker at du analyserer datamaterialet og bistår med råd om hvordan driften kan optimaliseres.

Datafilen «skipsdata.rdata» inneholder følgende variabler:

«dato» hver rad har en dato med tidskode, der hver observasjon (rad) har 1 minutts mellomrom.

«turnr» et tall mellom 1 og 9 som indikerer hvilken rekkefølge turene er gjennomført.

«mannskap» hvilket mannskap som er om bord, kodet som A eller B.

«hastighet» skipets hastighet målt i knop.

«forbruk» momentanforbruk på drivstoff målt i liter per time.

«stigevinkel» rullende standardavvik til 'pitch', skipets vinkel mot horisonten. Dette er et estimat på bølgehøyde. Høyere stigevinkel indikerer større bølger.

Dataene inneholder 48864 observasjoner fra 9 turer. Det er tatt ut data som måler båtens gange til og fra oppdrag. Hastigheten er derfor mellom 5 og 15 knop. Turene er valgt slik at det er mest mulig likhet mellom de to mannskapene med hensyn på gangtid og værforhold. Under en tur kan mannskapet på 'brua' selv velge fart. Økt fart gir økt forbruk men kortere gangtid. Rederiet har en fast kontrakt på bunkers i perioden på 4,90 per liter.

Ressurser som det kan være nyttig å se nærmere på:

http://www.sthda.com/english/wiki/ggplot2-density-plot-quick-start-guide-r-software-and-data-visualization

https://www.r-graph-gallery.com/2d-density-chart.html

https://ggplot2.tidyverse.org/reference/geom_density_2d.html

Tetthetsplot (eng. density plot) er et slags glattet histogram og benyttes ofte for å se på hvordan data fordeler seg rundt et gjennomsnitt.

Del 1: Optimalt forbruk

Følgende R kode laster «skipsdata» filen:

```
skipsdata <- readRDS(file = url("https://bit.ly/3FfJbKa"))
```

Gjør deretter følgende analyser:

a) Beregn gjennomsnittlig hastighet. Lag et tetthetsplot som viser fordelingen til hastighet, og vis gjennomsnittet som en vertikal linje i figuren. Beskriv fordelingen.

b) Beregn gjennomsnittlig hastighet til de to mannskapene. Lag et tetthetsplot som viser fordelingen til hastighet avhengig av hvilket mannskap som er om bord. Angi gjennomsnittshastigheten til mannskapene med to vertikale linjer. Hvilken ny informasjon vises sammenlignet med figuren i a)?

c) Lag et 2D-scatter plot med hastighet på x-aksen og forbruk på y-aksen. Punktene skal være små. Hvilke mønster vises? Hvordan tolker du dette?

d) Lag et plot med dato på x-aksen og forbruk på y-aksen. Legg til farge for hver tur, og informasjon om hvilket mannskap som er ombord. Hva er nytten av denne informasjonen?

e) Bølgehøyden har innvirkning på sammenhengen mellom skipets hastighet og momentanforbruk. Dette er estimert med variabelen «stigevinkel». Ta utgangspunkt i plottet fra oppgave c) og vis effekten av denne variabelen ved å legge til en fargegradient som funksjon av «stigevinkel» i plottet. Gradienten kan for eksempel være en regnbuegradient eller lignende. Hvordan tolker du dette?

f) Del datasettet i to, basert på «stigevinkel» høyere eller lik 0.5. Beregn gjennomsnittlig hastighet for mannskapene i de to gruppene av data. Beregn også forskjellen i gjennomsnittshastighet mellom mannskap.

Fyll ut følgende tabell:

Gjennomsnittshastighet

Data	Mannskap A	Mannskap B	A-B
Hele datasettet			
Stigevinkel < 0.5			
Stigevinkel ≥ 0.5			

g) Del datasettet i to, basert på «stigevinkel» høyere eller lik 0.5. Beregn gjennomsnittlig forbruk for mannskapene i de to gruppene av data. Beregn også forskjellen i gjennomsnittsforbruk mellom mannskap.

Fyll ut følgende tabell:

Gjennomsnittsforbruk

Data	Mannskap A	Mannskap B	A-B
Hele datasettet			
Stigevinkel < 0.5			
Stigevinkel ≥ 0.5			

Drøft dine funn over. Er det forskjell mellom mannskap på hastighet og forbruk? Har stigevinkel en forskjellig effekt på forbruk og hastighet mellom mannskap?

h) Som i c), lag et 2D-scatter plot med hastighet på x-aksen og forbruk på y-aksen. Legg til et lag som viser 2D-tetthetsfordelingen i plottet (`ggplot::stat_density2d()`) basert på mannskap.

På bakgrunn av din analyse skriv en kort oppsummering av dine funn til rederiets styre. Støtter dine resultater rederiets mistanke om at det er en forskjell mellom arbeidsmetodene til mannskap A og B? Gjennomsnittlig gangtid på båten er 1500 timer per år per mannskap. Benytt den oppgitte prisen på bunkers, og beregn hva rederiet kan spare pr år ved å ta i bruk arbeidsmetodene til det mest effektive mannskapet. Inkluder en figur som understøtter din konklusjon i oppsummeringen.

En skipsingeniør har også analysert andre deler av disse dataene. Hun mener at det er mulig å spare 5% på «forbruk» ved å optimalisere ballasten i båten ved gangtid. Hva er nettoeffekten av disse tiltakene?

Del 2: Motoreffektivitet

Forbrenningsmotorer konverterer den kjemiske energien i drivstoff til bevegelse. En motors effektivitet er lavest på tomgang og høyest nær maksimal belastning. Denne sammenhengen beskrives i detalj av en effektivitetskurve. Dieseleletriske skip har ofte flere store og små dieselmotorer. Mannskapet må velge den kombinasjonen av motorer som er mest hensiktsmessig. Du skal nå optimalisere valg av aktive motorer for et dieselelektrisk skip med hensyn på effektivitet.

Benytt datasettet «effektivitet.RData». Denne filen inneholder tre effektivitetskurver til et skip med to motorer. Én liten og én stor. Dataobjektene «liten», «stor» og «liten_stor» er effektivitetskurvene for henholdsvis en liten, en stor og begge motorene samtidig.

Datasettetene angir load og effektivitet.

«load» er motorens kapasitet til å produsere kraft (belastning), og er målt i kilowatt (kW).

«effektivitet» måles i kW per liter drivstoff.

Følgende R kode laster «effektivitet» filen:

```
load(url("https://bit.ly/3BcnSXA"))
```

a) Sett sammen «liten», «stor» og «liten_stor» i et nytt langt datasett. Benytt benevnelsen «motorkombinasjon» for å identifisere de originale datasettene.

b) Lag et plot som viser «liten», «stor» og «liten_stor» i det samme bildet, der «load» er på x-aksen og «effektivitet» på y-aksen. Sett farge etter «motorkombinasjon».

Tolk plottet. Hvilke motorvalg gir høyest effektivitet på de forskjellige load-nivåene?

c) Beregn «tapet» med følgende motorkombinasjoner:
   a. Kjøre 5000 timer på 300 kW med stor motor.
   b. Kjøre 7000 timer på 1600 kW med liten og stor motor.
   c. Kjøre 10000 timer på 2400 kW med liten og stor motor.

Del B - Skattetall

Datasettet «skattetall.RData» inneholder skattetall fra Troms fylke i 2015. Det er totalt 138521 personer med i datasettet som inneholder følgende variabler:

"kommnr" angir kommunenummer. Tromsø er 1902, mens Harstad er 1903.

"aldersgruppe" er en grupperingsvariabel for alder.

"fodt" er fødselsår.

"kjonn" er kodet «F» for kvinne og «M» for mann. Dataene inneholder også noen personer uten registrering, disse er kodet som «N».

"formue" angir formue i NOK i 2015.

"inntekt" angir inntekt i NOK i 2015.

"skatt" angir utlignet skatt i NOK i 2015.

Følgende R kode laster «skattetall» filen: `load(url("https://bit.ly/2YBntjg"))`

Benytt R pakken gglorenz for å lage figurer av ulikheter. Du finner pakken på CRAN:

https://cran.r-project.org/web/packages/gglorenz/index.html

Du finner repositoriet på github: https://github.com/jjchern/gglorenz

For å tegne Lorenz-kurven under 45 graders linja må du benytte `stat_lorenz(desc = FALSE)`. Alle oppgaver skal besvares. Du skal benytte ggplot til alle figurer. Alle figurer skal ha meningsfulle titler, 'labels', benevninger og enheter på aksene. For å beregne Gini kan du for eksempel benytte pakken ineq (https://cran.r-project.org/web/packages/ineq/).

a) Lag en Lorenz-kurve for inntekt i Troms i 2015 der Gini indeksen vises i figuren.

   Fyll ut følgende tabell, der du regner Gini om til prosent (x 100), og Inntekt er gjennomsnittlig inntekt.

Kommune	Gini (%)	Inntekt	Personer
Troms			138521

   Kommenter dine funn.

b) Lag en ny variabel «kommune», der Tromsø er kommunenr 1902, Harstad 1903, og Omegn er resten.

Fyll ut følgende tabell, der du regner Gini om til prosent (x 100), Inntekt er gjennomsnittlig inntekt og Andel er prosentandelen av skattebetalerne i kategorien Kommune.

Kommune	Gini (%)	Inntekt	Personer	Andel
Harstad				
Omegn				
Tromsø				

Kommenter dine funn, og sammenlign dem med det du fant i a).

c) Mange økonomer mener at det blir mer riktig å beregne Gini på disponibel inntekt, dvs inntekt fratrukket skatt. Beregn denne størrelsen. Dersom skatt er større enn inntekt vil disponibel inntekt bli negativ. Sett alle negative verdier lik 0. Lag en Lorenz-kurve for disponibel inntekt i Troms i 2015 der Gini indeksen vises i figuren. Kommenter dine funn, og sammenlign dem med det du fant i a).

d) Lag et plot av inntekt på y-aksen og skatt på x-aksen. Kommenter dine funn.

e) Lag et plot av formue på y-aksen og skatt på x-aksen. Kommenter dine funn, og sammenlign dem med det du fant i d).

f) Fyll ut følgende tabell, der du regner Gini om til prosent (x 100), Formue er gjennomsnittlig formue og Andel er prosentandelen av skattebetalerne i kategorien Kommune.

Kommune	Gini (%)	Formue	Personer	Andel
Harstad				
Omegn				
Tromsø				
Troms			138521	100%

Kommenter dine funn, og sammenlign dem med det du fant i a) og b).

g) Benytt variabelen «aldersgruppe» til å beregne Gini per aldersgruppe. Lag et plot der Gini er på y-aksen og aldersgruppe er på x-aksen. Kommenter dine funn.

h) Benytt variabelen «aldersgruppe» og «kjonn» til å beregne Gini per aldersgruppe per kjønn. Ta bort kategorien «N» i kjønn før du gjør din analyse. Lag et plot der Gini er på y-aksen og aldersgruppe er på x-aksen, sett farge etter kjønn. Kommenter dine funn.

Skriv et kort sammendrag om inntekt, formue og ulikheter i Troms fylke, basert på dine analyser.