Løsningsforslag eksamen MET4 - HØST 25

Til sensor: Hver deloppgave teller likt og gir maksimalt 10 poeng. Selv om R-kode gis i løsningsforslaget under skal kandidaten i utgangspunktet ikke legge ved noe R-kode i sin besvarelse, men det skal ikke gis noe trekk dersom de har gjort det. Det er også mulig å bruke R i Del 2, men da mer som en kalkulator og til å finne kritiske verdier og sannsynligheter. Vi gir ikke "stilpoeng" på hverken figurer eller ligninger. Så lenge sensor fortstår hva som menes/illustreres (og her er vi romslige) skal det gis full uttelling på den aktuelle delen av oppgaven.

Del 1 - Dataanalyse med R

```
library(dplyr)
library(ggplot2)
load("claims.Rdata")
```

Oppgave 1

a)

```
claims %>%
 select(driver_age, years_license, car_age, mileage_km) %>%
 summary
```

```
##    driver_age years_license      car_age         mileage_km
##  Min.   :18   Min.   : 0.00   Min.   : 1.000   Min.   : 3036
##  1st Qu.:38   1st Qu.:18.00   1st Qu.: 5.000   1st Qu.:10123
##  Median :47   Median :27.00   Median : 7.000   Median :13721
##  Mean   :47   Mean   :26.87   Mean   : 6.993   Mean   :15130
##  3rd Qu.:56   3rd Qu.:36.00   3rd Qu.: 9.000   3rd Qu.:18575
##  Max.   :85   Max.   :60.00   Max.   :19.000   Max.   :50000
```

Fra tabellen ser vi at:

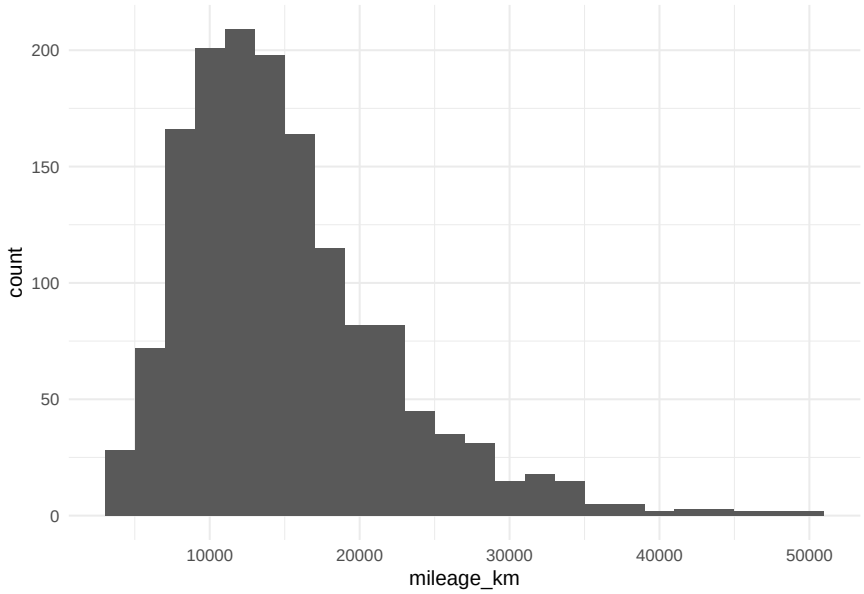
- Førernes alder har median og gjennomsnitt på 47 år, med en IQR på 38–56 år.
- Antall år med førerkort har median og gjennomsnitt på 27 år, med en IQR på 18–36 år.
- Bilalderen har median og gjennomsnitt på 7 år, med en IQR på 5–9 år.
- Den årlige kjørelengden har en median på 13 721 km, men et høyere gjennomsnitt på 15 130 km, som tyder på en høyreskjev fordeling. IQR er 10 123–18 575 km.

Til sensor: To og et halvt poeng per variabel. Kommentar om høyreskjev fordeling ikke nødvendig. Dersom kandidaten har presentert standardavvik istedet for IQR er dette også greit.

b)

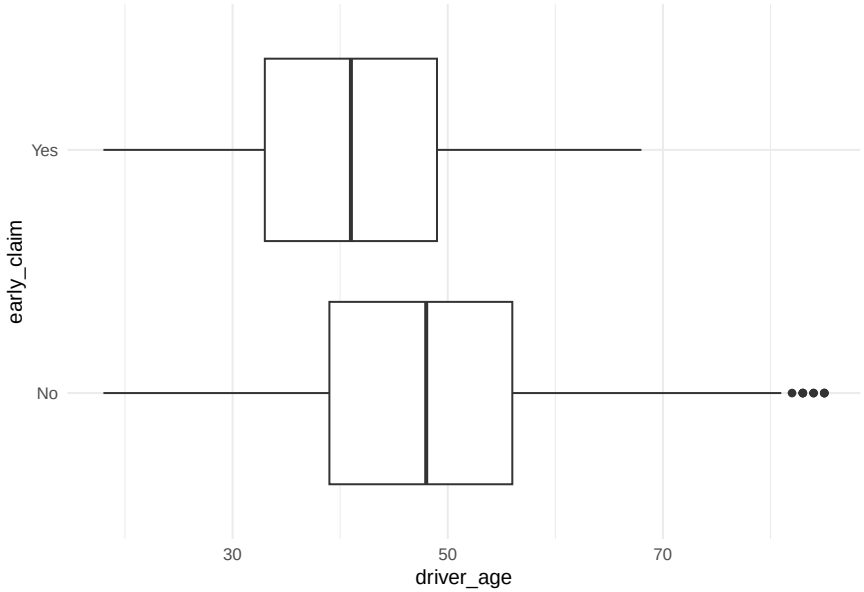
```
claims %>%
ggplot +
```

```
  geom_histogram(aes(x = mileage_km), binwidth = 2000) +
  theme_minimal()
```



Histogrammet viser at fordelingen av årlig kjørelengde er høyreskjev, der de fleste bilene har moderate kjørelengder mellom 10 000 og 20 000 km, mens noen få har svært høye verdier.

```
claims %>%
    ggplot +
  geom_boxplot(aes(x = driver_age, y = early_claim)) +
  theme_minimal()
```

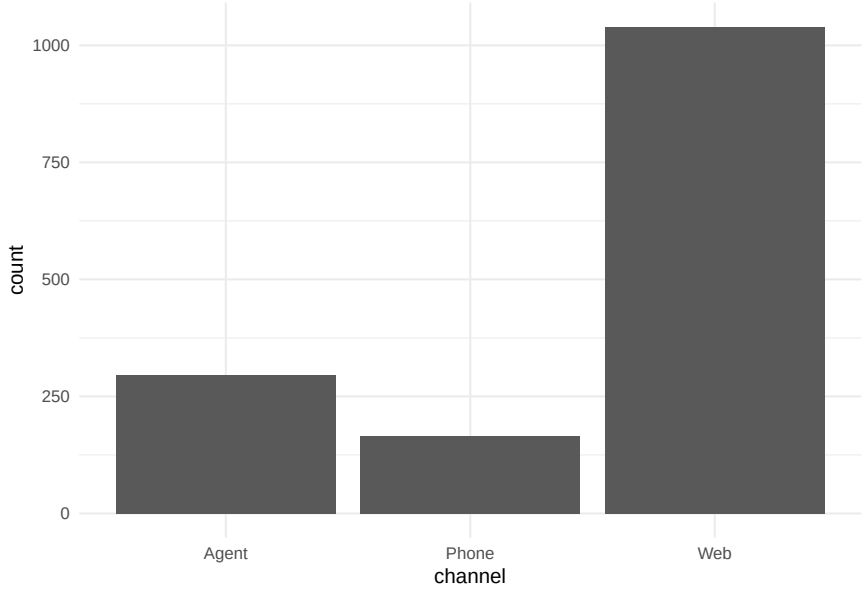


Boksplottet viser at medianalderen for de med tidlig skade er omtrent 41 år, mot ca. 48 år for de uten. Dette tyder på at yngre sjåfører har noe høyere risiko for tidlig skade. Det er ingen tydelig forskjell i spredning mellom de to gruppene.

Til sensor: 3 poeng per figur og 2 poeng per tilhørende kommentar. Kommentarene over er veiledende (Det er ikke så lett å se at medianalder er 41 og 48, så bare en kommentar om at den ene er større enn den andre er viktigst).

c)

```
claims %>%
   ggplot +
   geom_bar(aes(x = channel)) +
   theme_minimal()
```




```
table(claims$garage)
```

```
##
##  No Yes
## 746 754
```

```
table(claims$region)
```

```
##
## Rural Urban
##   375  1125
```

```
table(claims$multi_policy)
```

```
##
##  No Yes
## 949 551
```

Søylediagrammet viser at de fleste kundene kjøper forsikringen via internett (`Web`).

Frekvenstabellene viser videre at omtrent halvparten av kundene har garasje, og at de fleste bor i byområder (`Urban`) (1125 mot 375 i `Rural`). I tillegg har rundt en tredel av kundene flere forsikringer i selskapet.

Til sensor: 2.5 poeng per figur/tabell + tihørende kommentar.

Oppgave 2

a)

Vi skal her teste om forventet kjørelengde (`mileage_km`) er høyere for kunder som melder inn en tidlig skade enn for kunder som ikke gjør det. Vi skal altså utføre en ensidig test:

$$H_0 : \mu_{\text{Yes}} = \mu_{\text{No}} \quad \text{mot} \quad H_1 : \mu_{\text{Yes}} > \mu_{\text{No}}$$

der μ_{Yes} og μ_{No} er populasjonsgjennomsnittene til `mileage_km` for henholdsvis kunder som melder tidlig skade og kunder som ikke gjør det. Vi setter signifikansnivået til $\alpha = 0.05$.

```
mileage_yes <-
  claims %>%
  filter(early_claim == "Yes") %>%
  select(mileage_km) %>%
  pull

mileage_no <-
  claims %>%
  filter(early_claim == "No") %>%
  select(mileage_km) %>%
  pull
# To-sample t-test (mileage_km by early_claim)
t.test(mileage_yes, mileage_no,
       alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  mileage_yes and mileage_no
## t = 2.4353, df = 272.81, p-value = 0.007761
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  433.4979      Inf
## sample estimates:
## mean of x mean of y
##  16286.37  14941.23
```

Siden p-verdien er mindre enn 0.05 kan vi forkaste H_0. Det ser ut til at kunder som melder tidlig skade har høyere kjørelengde enn kunder som ikke gjør det.

Til sensor: 3 Poeng for riktig oppsatte hypoteser, 4 poeng for riktig utførelse av test og 3 poeng for riktig konklusjon. Det skal ikke trekkes poeng dersom kandidaten har valgt å anta lik varians. Merk: så lenge du notasjonsmessig forstår hva kandidaten mener skal det gis full pott. "H0: mu1 = mu2" er for eksempel ok. Se kommentar under for alternative fremgangsmåter.

Kommentar: Merk at i løsningsforslaget over vil `t.test()` teste om differansen mellom `mileage_yes` og `mileage_no` er `greater` enn 0, som samsvarer med alternativhypotesen $\mu_{\text{yes}} - \mu_{\text{no}} > 0$. Man kunne ekvivalent snudd argumentrekkefølgen på `mileage_yes` og `mileage_no` og testet med `alternative = "less"` som også ville samsvart med alternativhypotesen ($\mu_{\text{no}} - \mu_{\text{yes}} < 0$):

```
t.test(mileage_no, mileage_yes,  alternative = "less")
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  mileage_no and mileage_yes
## t = -2.4353, df = 272.81, p-value = 0.007761
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -433.4979
## sample estimates:
## mean of x mean of y
##  14941.23  16286.37
```

Oppgaven kan også løses med bruk av `~` men da må en ta hensyn til nivårekkefølgen til `early_claim` som er `no` og `yes` (Dette vil man se av R-utskriften fra `t.test()`). Det betyr at `t.test()` ser på differansen mellom `mileage_km` i `no` gruppa og `mileage_km` i `yes` gruppa. Av samme grunn som over må man derfor spesifisere 'alternative = "less" med denne fremgangsmåten dersom den skal samsvare med alternativhypotesen ($\mu_{\text{no}} - \mu_{\text{yes}} < 0$):

```
t.test(mileage_km ~ early_claim,  data = claims, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  mileage_km by early_claim
## t = -2.4353, df = 272.81, p-value = 0.007761
## alternative hypothesis: true difference in means between group No and group Yes is less than 0
## 95 percent confidence interval:
##       -Inf -433.4979
## sample estimates:
##  mean in group No mean in group Yes
##          14941.23          16286.37
```

Disse to fremgangsmåtene vil også gi full pott.

b)

Dersom de tre kjøpskanalene er like mye brukt, vil populasjonsandelen for hver av de tre kjøpskanalene være 1/3. Vi skal altså teste:

$$H_0 : p_{\text{Web}} = 1/3, \quad p_{\text{Agent}} = 1/3, \quad p_{\text{Phone}} = 1/3 \quad \text{mot} \quad H_1 : \text{Minst to sannsynligheter har andre verdier}$$

Vi setter signifikansnivået til $\alpha = 0.05$. Kji-kvadrat goodness-of-fit-test i R:

```
f  <- table(claims$channel)
p0 <- c(1/3, 1/3, 1/3)
chisq.test(x = f, p = p0)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  f
## X-squared = 888.2, df = 2, p-value < 2.2e-16
```

Siden p-verdien er mindre enn 0.05 kan vi forkaste H_0. Det er altså ikke støtte for at kanalene er like mye brukt. Dette samsvarer med figuren i oppgave 1c) der vi ser at det er en stor preferanse for å handle på nett.

Til sensor: 3 Poeng for riktig oppsatte hypoteser, 4 poeng for riktig utførelse av test og 3 poeng for riktig konklusjon.

Oppgave 3

```
treningssett <- claims[1:1050, ]          # La treningsett være de første 70 % av dataene
testsett <- claims[1051:nrow(claims), ]   # La testsettet være de resterende 30 %
```

a)

Utskrift av modell:

```
model1 <- glm(early_claim ~ driver_age + car_age + mileage_km +
    region + payment_method + garage,
  data = treningssett, family = "binomial")
summary(model1)
```

```
##
## Call:
## glm(formula = early_claim ~ driver_age + car_age + mileage_km +
##     region + payment_method + garage, family = "binomial", data = treningssett)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -1.402e+00  5.198e-01  -2.697  0.00700 **
## driver_age            -5.039e-02  7.767e-03  -6.488 8.72e-11 ***
## car_age                8.684e-02  3.436e-02   2.528  0.01148 *
## mileage_km             3.942e-05  1.260e-05   3.127  0.00176 **
## regionUrban            7.875e-01  2.564e-01   3.071  0.00213 **
## payment_methodFaktura  4.358e-01  1.875e-01   2.325  0.02007 *
## garageYes             -3.929e-01  1.926e-01  -2.040  0.04134 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 861.24  on 1049  degrees of freedom
## Residual deviance: 782.31  on 1043  degrees of freedom
## AIC: 796.31
##
## Number of Fisher Scoring iterations: 5
```

Tolkning via odds:

```
exp(coef(model1))
```

```
##           (Intercept)            driver_age               car_age
##             0.2461821             0.9508571             1.0907257
##            mileage_km           regionUrban payment_methodFaktura
##             1.0000394             2.1980014             1.5462636
##             garageYes
##             0.6751092
```

Vi ser at førerens alder har en signifikant negativ sammenheng med sannsynligheten for tidlig skade ved 0.001-nivået. En økning i alder på ett år er forbundet med at oddsen for tidlig skade reduseres med en multiplikativ faktor på 0.9508, tilsvarende en reduksjon i odds på

$$(1 - 0.9508) \cdot 100\% = 4.9\%.$$

Videre har garasje en signifikant negativ sammenheng med tidlig skade ved 0.05-nivået. Det å ha garasje er forbundet med en odds som er 0.6751 ganger lavere enn for kunder uten garasje, det vil si en reduksjon i odds på

$$(1 - 0.6751) * 100\% = 32.5\%.$$

Dette indikerer at både høyere alder og det å ha garasje henger sammen med lavere odds for tidlig skade, alt annet likt.

Til sensor: 4 Poeng for utskrift av riktig modell, 3 poeng for hver tolkning av koeffisient. Trekk 1 av 3 poeng dersom kommentar om signifikans mangler. Trekk 2 av 3 poeng dersom tolkning ikke er gjort via odds. Trekk 3 av 3 poeng for utsagn om garasje som "En enhets økning i garasje..." (tolkning av garasje som kontinuerlig variabel). Omregning til prosent er ikke nødvendig for full pott.

b)

```
# Klassifisering
  pred_model1 <- predict(model1, newdata = testsett, type = "response")
  klass_model1a <- ifelse(pred_model1 > 0.1, "Yes", "No")
  klass_model1b <- ifelse(pred_model1 > 0.3, "Yes", "No")


  # Kontigenstabeller
  sann <- testsett$early_claim
  tab_model1a <- table(sann, klass_model1a)
  tab_model1b <- table(sann, klass_model1b)

  tab_model1a
```

```
##      klass_model1a
## sann   No Yes
##   No  178 212
##   Yes  15  45
```

```
  tab_model1b
```

```
##      klass_model1b
## sann   No Yes
##   No  355  35
##   Yes  43  17
```

Ved grense 0.10 klassifiseres mange kunder som "Yes". Dette gir høy sensitivitet, altså flere faktiske "Yes" fanges opp (45 av 60), men også mange falske positive (212 "No"-kunder feilklassifiseres).

Ved grense 0.30 blir modellen mer konservativ. Den klassifiserer flere "No" korrekt (355 av 390), men fanger bare opp 17 av 60 faktiske "Yes". Altså får modellen høyere spesifisitet, men lavere sensitivitet.

Den lave grensen (0.10) er derfor bedre dersom målet er å fange opp flest mulig faktiske skadetilfeller, mens den høye grensen (0.30) er bedre dersom man ønsker å unngå feilvarsler siden den gjør færre feilklassifiseringer av trygge kunder.

Til sensor: 2 poeng for hver av tabellene, 6 poeng beskrivelse av fordeler og ulemper. Ord som "sensitivitet" og "spesifisitet" er ikke nødvendig for å få full pott. Kommentaren over er veiledende.

c)

```
library(caret)

# Tilpasser modellen
  model2 <- train(early_claim ~ driver_age + car_age + mileage_km +
                    region + payment_method + garage,
                  data = treningssett,
                  method = "knn",
                  tuneGrid = data.frame(k = 5))


  # Klassifisering
  klass_model2 <- predict(model2, newdata = testsett)
  tab_model2 <- table(sann, klass_model2)
  tab_model2
```

```
##      klass_model2
## sann   No Yes
##   No  381   9
##   Yes  56   4
```

```
  # Andeler riktige klassifiseringer
  tab_model2 %>% prop.table %>% diag %>% sum
```

```
## [1] 0.8555556
```

KNN-modellen klassifiserer de fleste kundene som `"No"`. Av totalt 390 kunder uten tidlig skade blir 381 korrekt klassifisert, mens kun 4 av 60 med tidlig skade blir riktig identifisert.

Den totale andelen riktige klassifiseringer er om lag 85.6 %. Dette høres høyt ut, men skyldes i hovedsak at de fleste kundene tilhører `"No"`-gruppen. Modellen fanger i liten grad opp kunder med tidlig skade og er derfor lite egnet dersom målet er å oppdage risikokunder, men den gjør få feilklassifiseringer av trygge kunder.

Til sensor: 3 poeng for tabell, 3 poeng for andel og 4 poeng om hvordan modellen presterer.

d)

Vi tenker oss en tilfeldig variabel G = «gevinstfor en vilkårlig kunde» når vi bruker en gitt modell.
Det finnes fire mulige utfall for G, avhengig av (faktisk, predikert):

- Riktig **Yes** (Yes, Yes): $G = +20\,000$
- Riktig **No** (No, No): $G = +300$
- **Falsk positiv** (No, Yes): $G = -500$
- **Falsk negativ** (Yes, No): $G = -5\,000$

Forventet nettogevinst per kunde er dermed forventningsverdien

$$\mathbb{E}[G] = 20\,000 \cdot P(\text{Yes, Yes}) + 300 \cdot P(\text{No, No}) - 500 \cdot P(\text{No, Yes}) - 5\,000 \cdot P(\text{Yes, No}).$$

De estimerte sannsynlighetene for disse fire utfallene hentes direkte fra den **normaliserte kontingenstabellen**, som gir andelen av alle kunder som havner i hver celle. Kall den normaliserte tabellen for P med rekkefølgen rader/kolonner = (No, Yes). Da er:

- $P(\text{Yes, Yes}) = P[2,2]$
- $P(\text{No, No}) = P[1,1]$
- $P(\text{No, Yes}) = P[1,2]$
- $P(\text{Yes, No}) = P[2,1]$

Utregning av forventingsverdier kan gjøres i R som følger:

```
# Lage normaliserte kontigenstabeller for å få sannsynlighetsfordelingen
 P1a <- tab_model1a %>% prop.table
 P1b <- tab_model1b %>% prop.table
 P2 <- tab_model2 %>% prop.table

 # Beregn forventet nettogevinst
 netto_1a <- 20000*P1a[2,2] + 300*P1a[1,1] - 500*P1a[1,2] - 5000*P1a[2,1]
 netto_1b <- 20000*P1b[2,2] + 300*P1b[1,1] - 500*P1b[1,2] - 5000*P1b[2,1]
 netto_2  <- 20000*P2 [2,2] + 300*P2 [1,1] - 500*P2 [1,2] - 5000*P2 [2,1]


# Presentert i en tabell:
 data.frame(
   Modell = c("Logistisk (0.10)", "Logistisk (0.30)", "KNN"),
   Nettogevinst_pr_kunde_kr = round(c(netto_1a, netto_1b, netto_2), 0)
 )
```

```
##             Modell Nettogevinst_pr_kunde_kr
## 1 Logistisk (0.10)                    1716
## 2 Logistisk (0.30)                     476
## 3              KNN                    -200
```

Vi ser at den logistiske modellen med lav klassifiseringsgrense (0.10) gir høyest forventet nettogevinst per kunde. Når gevinsten av å oppdage risikokunder og kostnaden ved å overse dem er høy, lønner det seg å bruke en modell som er god til å identifisere kunder med tidlig skade, selv om den da også klassifiserer noen flere feilaktig som risikokunder. Dette viser at den "beste" modellen kan avhenge av hva som faktisk verdsettes.

Til sensor: 7 poeng for riktige utregninger av forventede gevinster. 3 poeng for kommentar.

Oppgave 4

a)

Her er det rimelig å bruke en paret t-test hvor hvert par er en observasjon av kostnaden for en enkelt måned med- og uten Norgespris. Dette er fordi vi sammenligner to scenarioer (med og uten Norgespris) for den samme husholdningen over tid. Observasjonene er dermed avhengige, og variasjon mellom måneder (som temperatur og forbruk) påvirker begge ordninger likt. Ved å bruke en paret test fjerner vi denne felles variasjonen og tester direkte om gjennomsnittlig forskjell i kostnader er større enn null.

Det vil lønne seg for Sondre å tegne Norgesprisavtale viss strømkostnadene blir lavere med norgespris enn uten. Siden kostnadsdifferansen er regnet "Uten" minus "Med" Norgespris, vil positive verdier bety at Norgespris er billigere. Vi putter det vi ønsker å finne ut (lønner Norgespris seg?) i alternativhypotesen. Derfor blir hypotesene:

$$H_0 : \mu_Y = 0 \quad \text{vs} \quad H_A : \mu_Y > 0.$$

Der μ_Y er populasjonsgjennomsnittet til $Y_t = \texttt{stromstotte}_t - \texttt{norgespris}_t$.

For denne problemstillingen er det ikke er så farlig å være litt liberal (høyt signifikansnivå). Konsekvensen av en Type I feil er bare at en bytter uten at det strengt tatt er nødvendig. Vi bruker $\alpha = 10\%$.

Til sensor: 4 poeng for riktig hypotesetest og 4 poeng for riktig oppsatte hypoteser. Dersom kandidaten har satt opp en to-utvalgs t-test med riktig formulert (ensidig) hypoteser gis 4 poeng. 2 poeng for diskusjon rundt signifikansnivå, her godtas store signifikansnivå. Det ingen grunn til å sette det mindre enn 0.05. Det er argumentene som bør være fornuftige.

b)

Vi forkaster nullhypotesen for store verdier av testobservatoren, da dette vil favorisere alternativhypotesen:

$$t = \frac{\overline{y} - 0}{s_Y/\sqrt{n}} = \frac{275.49}{397.06/\sqrt{37}} = 4.22$$

Vi finner kritiske verdier for testobservatoren:

```
sign_niva <- 0.1
qt(p =sign_niva, df = 36, lower.tail = FALSE)
```

```
## [1] 1.305514
```

Vi ser at test-statistikken $t = 4.22 > 1.31$, og vi forkaster nullhypotsen. Det er altså grunn til å tro at Norgespris vil lønne seg for Sondre.

Til sensor: 7 poeng for utførelse av hypotesetest og 3 poeng er forbeholdt konkluderingen. Dersom kandidaten utfører en to-utvalgs t-test med `norgespris` og `stromstotte` kan de få full uttelling (10 poeng) dersom alt er gjort riktig (test + konklusjon). Ut fra standardavvikene i tabellen bør det antas ulik varians. Trekk 3 poeng hvis de har antatt lik varians.

Oppgave 5

a)

Den estimerte modellen er gitt ved

$$\texttt{differanse} = 809.566 - 58.988 \cdot \texttt{temperatur} + \epsilon.$$

Der ϵ er normalfordelt med forventning 0 og standardavvik 248.7.

Som vi var inne på i Oppgave 4a, tilsier positive verdier for differansen at Norgespris er mer gunstig, og her ser vi at differansen er avtagende for høyere temperaturer. Norgepris er altså gunstig hvis $\texttt{differanse} > 0$, dvs:

$$809.566 - 58.988 \cdot \texttt{temperatur} > 0 \quad \Longleftrightarrow \quad \texttt{temperatur} < \frac{809.566}{58.988} = 13.7.$$

I henhold til modellen, vil Norgespris lønne seg for måneder med gjennomsnittstemperatur $< 13.7°C$ og ikke lønne seg for måneder med gjennomsnittstemperatur $> 13.7°C$. En ca. verdi kan også ses fra figur.

Til sensor: 4 poeng for riktig oppsett av modell og 6 poeng for å finne intervallet. Full pott dersom de har brukt figur til å finne ca. samme intervall.

b)

Vi velger den modellen som har lavest AIC-verdi. Her er dette lm2 som bruker `temperatur` og `nedbor` som forklaringsvariable.

Vi ser at å legge til forbruk, både når en går fra modell `lm1` til `lm4` og fra `lm2` til `lm3`, ikke forbedrer AIC. Når du allerede har `temperatur` i modellen gir det mening at `forbruk` ikke gir noe særlig bedring i modellen. Som det pekes på i oppgaveteksten henger forbruk av energi tett sammen med temperatur (de er avhengige variabler), da norske husholdninger primært varmes opp med elektrisk energi.

Til sensor: 4 poeng for å finne modellen med lavest AIC og 6 poeng for riktig konklusjon om forbruk og tilhørende kommentar. En god besvarelse bør nevne noe om avhengighet mellom forbruk og temperatur.

c)

Setter vi inn `nedbor` $= 225.8$ og `temperatur` $= 10.4$ får vi

$$\widehat{\texttt{differanse}} = 1058.5930 - 62.9260 \cdot 10.4 - 1.0452 \cdot 225.8 = 168.1564 = 168.2.$$

Koeffisienten for Nedbør er etimstert til -1.0452. Det betyr at en ekstra millimeter med nedbør er assosisert med en nedgang på 1.0452 NOK i differansen mellom strømstøtteordning og Norgespris for Sondre sin strømkostnad.

At differansen avtar med økende nedbør (negativ koeffisient) kan forklares ved at økt nedbør gir fullere vannmagasiner (mer tilbud), som igjen gir lavere strømpris og gjør det mindre gunstig med Norgespris.

Til sensor: 4 poeng for korrekt prediksjon og 4 poeng for fortolkning av koeffisienten for nedbør. 2 poeng for en god kommentar om mekanismen bak.

Oppgave 6

a)

Som det står innledningsvis i Regnedelen av eksamen vil $Y_t > 0$ favorisere Norgespris og $Y_t < 0$ vil være i disfavør Norgespris. Av grafen i Vedlegg 3, ser vi at i) Januar 2025 er $Y_t > 0$ og derfor vil Norgespris være gunstig da. For ii) Juli 2025 er $Y_t < 0$ og det vil være fordelaktig å ikke ha Norgespris.

Til sensor: 5 poeng for riktig konklusjon for hver av månedene.

b)

En stasjonær tidsrekke har 1) konstant forventning, 2) konstant og endelig varians og 3) korrelasjonen mellom to variabler avhenger bare av avstanden i tid (ikke av tidspunktene). Det siste kravet klarer vi ikke sjekke av grafen, men de to første er brutt her: 1) Tidsrekken viser tydelige tegn til sesong (lav om sommeren og høy om vinteren). Dette er et brudd på stasjonaritet, siden forventningen ikke er konstant. 2) Det kan også se ut som spredningen er avtagende over tid, som bryter med konstant varians for stasjonære tidsrekker.

Til sensor: I utgangspunktet 5 poeng for å påpeke at sesongen bryter med stasjonaritetsantagelsen og 5 poeng for å påpeke at det ikke ser ut til å være konstant varians heller. De trenger ikke oppgi kravene for stasjonaritet, men trekk etter skjønn for å dra inn irrelevante elementer i begrunnelsen.

c)

Hvis vi tenker oss at svingningene i differansen skyldes endringer i temperaturen, kan vi si at sesongkomponenten S_t "modellerer" $\beta_1 \cdot \texttt{temperatur}_t$ i modellene i Oppgave 5. I stedet for å bruke data fra kilden til sesongen i differansen, så bare modellerer vi sesongen direkte.

Bakdelen med å bruke temperatur i en prediksjonsmodell hvor vi skal predikere langt frem i tid, er at vi må vite hva temperaturen vil bli langt frem i tid. Dette kan utfordrende siden vanlige værmeldinger stort sett ikke går så langt frem i tid.

Til sensor: 5 poeng for å drøfte påstanden om at sesongkomponenten "modellerer" temperatureffekten og 5 poeng for forklaringen mtp prediksjon frem i tid. En god besvarelse bør nevne dette med at vi trenger å vite eller kunne predikere temperaturen i fremtiden når vi bruker temperatur i modellene våre, mens sesongen kan vi bare la fortsette, siden den er kjent.

---

d)

$X_t = T_t + R_t$ er modellert med en ARIMA(0,1,1). Det vil si at den differensierte tidsrekken, ΔX_t, er en MA(1). Vi kan sette det opp slik:

$$\Delta X_t = X_t - X_{t-1}, \quad \text{der} \quad \Delta X_t = \widehat{\theta} u_{t-1} + u_t = -0.4823\, u_{t-1} + u_t.$$

der u_t er hvit støy med varians $\sigma^2 = 31.84$.

Til sensor: 10 poeng for riktig modell. Trekk 4 poeng for å sette opp en modell (uansett hvilken) for X_t. Trekk 4 poeng for feil ARMA-modell. Trekk 2 poeng dersom u_t (eller tilsvarende notasjon) ikke er definert (1 poeng for hvit støy, 1 poeng for tilhørende varians).

---

e)

Siden W er en sum, kan vi bruke sentralgrenseteoremet til å argumentere for at W er tilnærmet normalfordelt. Den røde linjen i histogrammet i Vedlegg 5 ser ut til å passe bra. Ved å bruke denne normalfordelingen, kan vi finne sannsynligheten for W er større enn null:

$$P(W > 0) = 0.9381.$$

Der vi har brukt:

```
pnorm(0, mean = 5909, sd = 3840, lower.tail = FALSE)
```

```
## [1] 0.9380737
```

Alternativt, kan en gjøre det på "gamle" måten:

$$P(W > 0) = P(\frac{W - 5909}{3840} > \frac{0 - 5909}{3840}) = P(Z > -1.54) = 1 - P(Z \leq -1.54) = 1 - 0.0619 = 0.9381.$$

hvor $P(Z \leq -1.54)$ kan letes opp i tabell eller regnes ut i R med:

```
pnorm(-1.54)
```

```
## [1] 0.06178018
```

Til sensor: 5 poeng for begrunnelse. Her bør sentralgrenseteoremet være nevnt. 5 poeng for korrekt utregning av sannsynligheten.

f)

Som det pekes på i teksten, vil innføringen av Norgespris gjøre at en stor andel av strømkundene ikke vil reagere på strømprissignaler. Dette vil kunne føre til enda høyere strømpriser når strømmen er en knapphetsgode. Det vil igjen kunne gjøre strømmen dyrere for de som ikke har tegnet Norgespris, og dermed blir differansen vi modellerer her høyere. Modellen er trent på data fra en periode hvor det ikke var noen Norgespris, så denne mekanismen er ikke innbakt i de historiske prisene. Hvis beskrivelsen her er korrekt, vil trolig modellen underdrive fordelen med å ha Norgespris i det påfølgende året. Hovedproblemet er altså at modellen vi tilpasset ikke generaliserer seg så godt til fremtiden.

Til sensor: 6 poeng for diskusjon av om disse aspektene vil ha innvirkning på prediksjon. Her ønsker vi primært at studenten skal se at innføringen Norgespris antagelig vil ha en effekt på strømmarkedet og dermed også prisen fremover, og at modellen(e) er trent på en periode uten Norgespris (kostnadene her er "kontrafaktiske"). 4 poeng for å si noe om hvordan det vil påvirke (forsterke gunstigheten av Norgesprisen). Her kan det være andre gode argumenter som går i en annen retning enn den løsningsforslaget antyder og det skal studenten ha uttelling for.