SOK-2009,høst 2022, Gruppeeksamen

Gruppenummer 1, Kandidatnummer 1, 35, 133

Buktafestivalen

I denne rapporten skal vi svare på noen spørsmål til verdiene vi finner i dataene til Buktafestivalen. Blandt annet skal vi se på inntekter, besøkende og salget av drikke og om andre faktorer kan ha innvirkning på besøkende og/eller salget av drikkevarer. Vi tar for oss totalt 7 oppgaver med spørsmål som vi skal svare på.

Totale Inntekter og deltagelse

(Oppgave 1)

Det første vi ser i den første tabellen under her er den totale inntekten pr år og flest inne på området samtidig det året og hvor mye hver gjest da bidro til totalinntekten

År	Gjester	Inntekt	Inntekt_per_deltager
2016	5723	1868516	326.4924
2017	0	3372432	0.0000
2018	5077	3098624	610.3258
2019	3561	3226524	906.0725

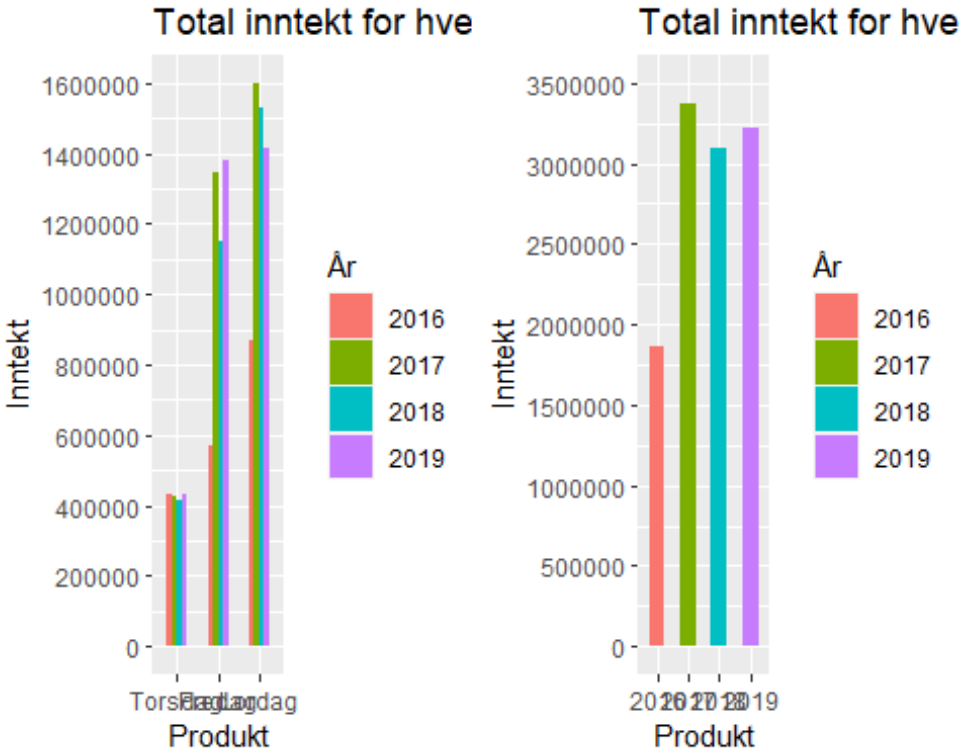
I tabellen finner vi at det var flest gjester i 2016, men at de pr deltaker brukte minst, gjester for 2017 er ikke registert og når deltakere eller inntekt pr deltaker er i sammenligning må vi se bort dra dette året. Samtidig var det færrest deltakere i 2019 men til gjengjeld brukte de mest pr deltaker. Året de hadde høyest samlet inntekt var i 2017. Deltakerne bruker mer for hvert år, inntektene øker selv om deltakerantallet har gått ned. Og hvis vi bryter disse tallene opp i dager får vi disse resultatene:

År	Dag	Gjester	Total_inntekt_per_dag
2016	Torsdag	2655	431304.0
2016	Fredag	4017	568272.0
2016	Lordag	5723	868940.0
2017	Torsdag	0	426940.0
2017	Fredag	0	1346096.0
2017	Lordag	0	1599396.0
2018	Torsdag	2114	414888.0
2018	Fredag	3261	1153132.0
2018	Lordag	5077	1530604.0
2019	Torsdag	1956	431574.4

```
2019    Fredag          3411                1378507.2
2019    Lordag          3561                1416442.4
```

I liket på dagene kan vi se at lørdag er den dagen alle årene det er flest besøkende og størst inntekt med fredagene rett bak og tordagen med lavet besøkende og lavest inntekt, men en økende stigning i inntekt og besøkende fra torsdag til lørdag er å se.

I diagrammene under kan vi se dette visuelt:



Her ser vi tallene i fra tabellene visuelt og bekrefter at torsdag er den dagen med færrest besøkende og inntekt over alle årene og økende til fredag og videre lørdag. Og på årsbasis ser vi at 2017 er året med høyest samlet inntekt og 2016 den med lavest.

Produkttyper

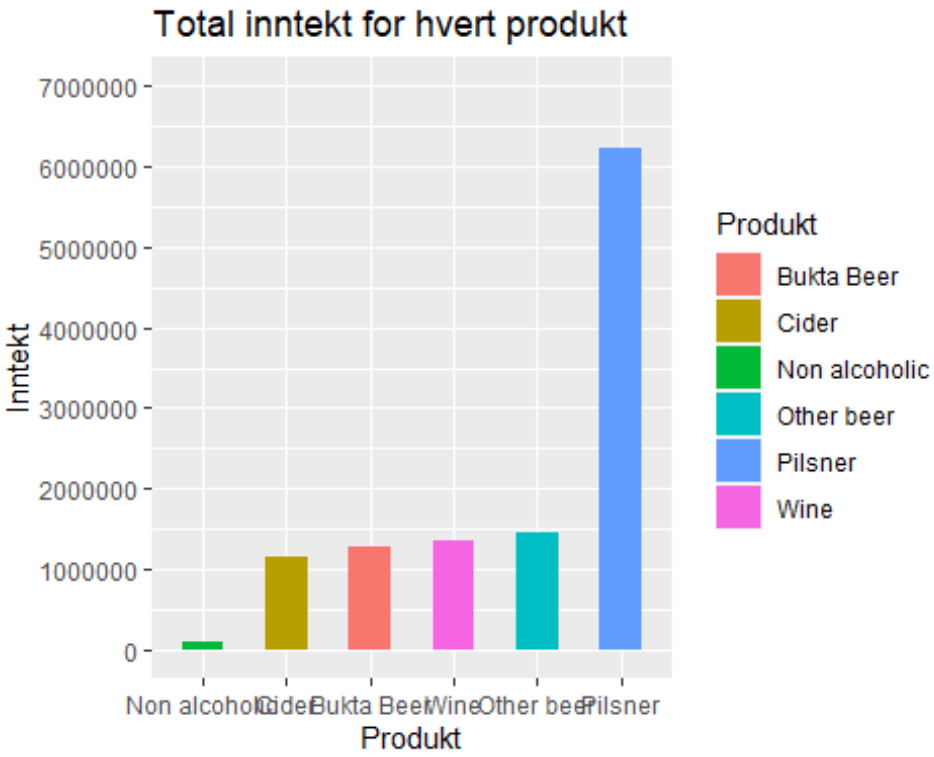
(oppgave 2.1)

Hvis vi ser på drikkevarene som selges så ser man totalt sett at Pilsner er det som selges mest, mens ikke-alkoholholdige varer selger minst.

```
Produkt            Inntekt
--------------  ----------
Bukta Beer       1284348.0
Cider            1147168.8
Non alcoholic     112655.2
```

```
Other beer       1446688.8
Pilsner          6223619.2
Wine             1351616.0
```

Dette ser vi også tydelig på dette diagrammet at Pils er mest populært.

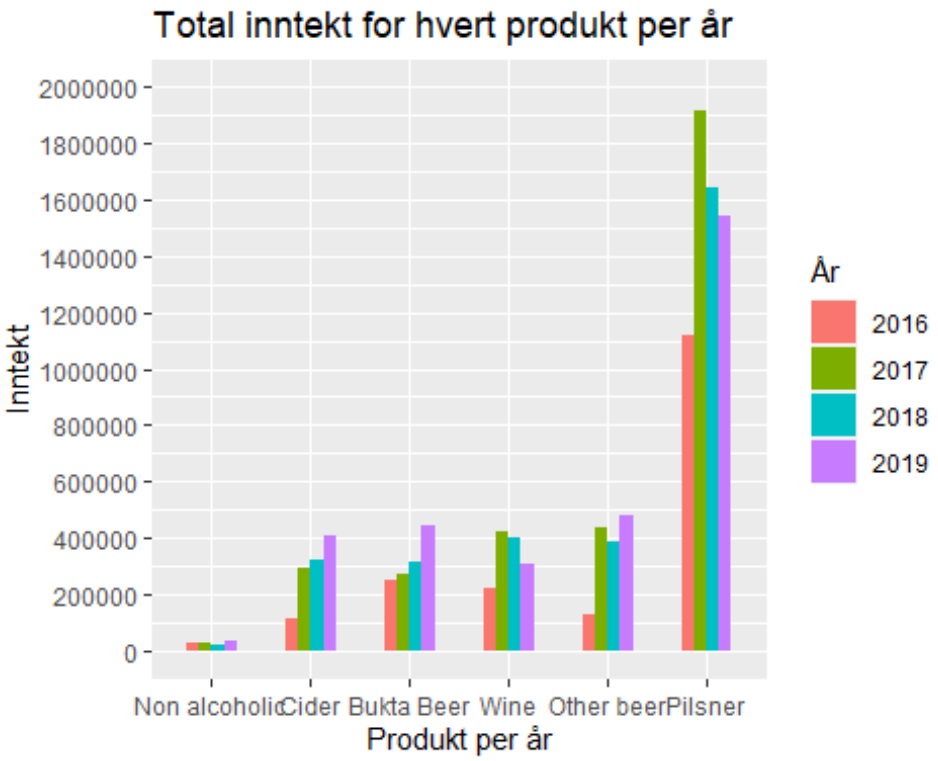


I tabellen under får vi opp hva som er mest og minst populært pr dag hvert år.

```
Produkt          År        Inntekt
--------------  -----  ----------
Bukta Beer       2016     252300.0
Bukta Beer       2017     271728.0
Bukta Beer       2018     318168.0
Bukta Beer       2019     442152.0
Cider            2016     116772.0
Cider            2017     296068.0
Cider            2018     322380.0
Cider            2019     411948.8
Non alcoholic    2016      25488.0
Non alcoholic    2017      27852.0
Non alcoholic    2018      23228.0
Non alcoholic    2019      36087.2
Other beer       2016     132532.0
Other beer       2017     439484.0
Other beer       2018     390624.0
```

```
Other beer      2019     484048.8
Pilsner         2016    1118872.0
Pilsner         2017    1917720.0
Pilsner         2018    1645056.0
Pilsner         2019    1541971.2
Wine            2016     222552.0
Wine            2017     419580.0
Wine            2018     399168.0
Wine            2019     310316.0
```

Vi ser et pils er fortsatt den som gir mest inntekt og non alcohol minst inntekt.



Dette ser vi i diagrammet over også, pils er mest populært mens non alcoholic er mist populær, samtidig er også ikke alhoholholdige drikker rimeligere enn alkoholholdige drikker så selv om man selger flere enheter ikke alkohol så vil dette alikevel gi mindre inntekt enn om man solgte samme mengde pils. Men siden inntekten er såpass liten så vil ikke prisforskjellen utgjøre et veldig stort utslag på inntekten og vil fortsatt være minst innbrignede, og er liten variasjon på for hvert år. Bukta beer viser stigende trend over årene i popularitet selv om denne er forskjellig fra år til år. Cider er også økende i trend, mens annen type øl er lite bevegelig med unntak fra 2016. Vin er den eneste typen som viser en nedgående trend, samme med pils unntak av 2016, men er likevel den som bringer inn mest inntekter hvert år.

Parvis t-test og Holm

(oppgave 2.2)

Kjører først summary, derreter parvis t-test for så justere med holm til sist.

```
`summarise()` has grouped output by 'Produkt', 'Time', 'År'. You can override
using the `.groups` argument.


Call:
lm(formula = Time ~ Total_inntekt_per_time, data = Time)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5525 -1.2353 -0.1268  1.1627  3.2156

Coefficients:
                          Estimate  Std. Error t value             Pr(>|t|)
(Intercept)            18.75578423  0.11350189 165.246 < 0.0000000000000002
***
Total_inntekt_per_time  0.00006199  0.00001198   5.173           0.00000039
***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.606 on 348 degrees of freedom
Multiple R-squared:  0.0714,    Adjusted R-squared:  0.06873
F-statistic: 26.76 on 1 and 348 DF,  p-value: 0.0000003901


    Paired t-test

data:  Time$Time and Time$Total_inntekt_per_time
t = -16.119, df = 349, p-value < 0.00000000000000022
alternative hypothesis: true mean difference is not equal to 0
99 percent confidence interval:
 -7171.191 -5185.672
sample estimates:
mean difference
      -6178.431


    Pairwise comparisons using t tests with pooled SD

data:  Total_inntekt_per_time and Produkt

             Bukta Beer           Cider                Non alcoholic
Cider        1.00000              -                    -
Non alcoholic 0.00143             0.00078              -
Other beer   1.00000              1.00000              0.000074
Pilsner      < 0.0000000000000002 < 0.0000000000000002 < 0.0000000000000002
Wine         1.00000              1.00000              0.00296
             Other beer           Pilsner
```

```
Cider          -                      -
Non alcoholic  -                      -
Other beer     -                      -
Pilsner        < 0.0000000000000002   -
Wine           1.00000                < 0.0000000000000002

P value adjustment method: holm
```

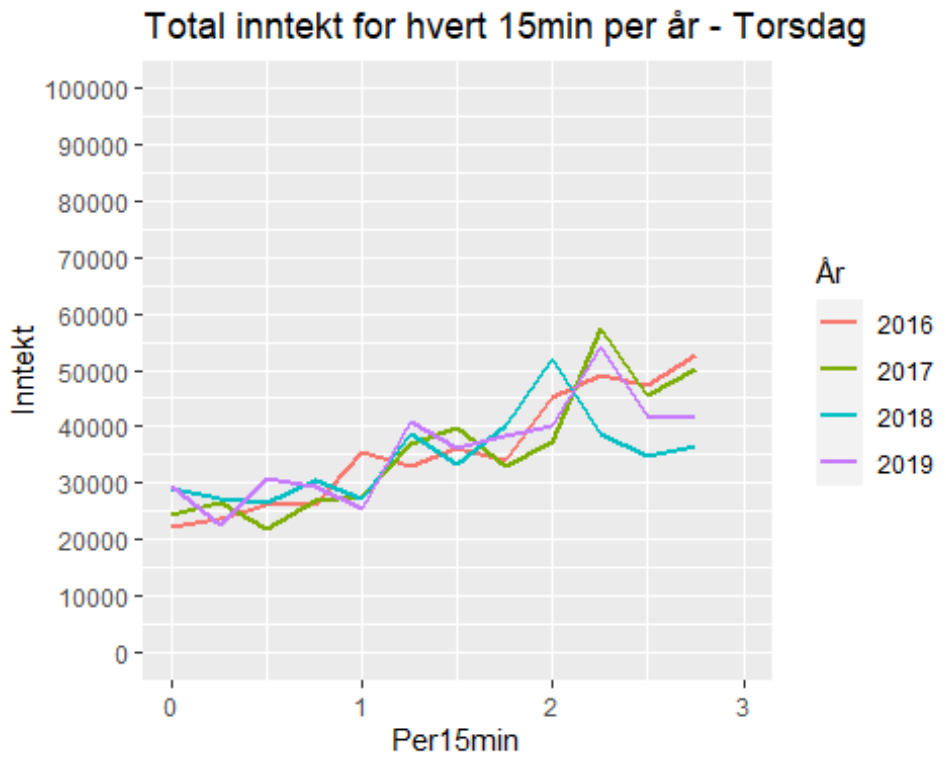
Vi bruker parvis t-test metoden med holm-justering for å teste om gjennomsnittsforskjellen mellom målpar er null eller ikke. I denne sammenhengen ønsker vi å finne ut hvilke produkter bukta festivalen tjener mer eller mindre på og hvilke som er relativt lik. Både fra den parvise t-testen og metoden holm får vi at p-verdien er 0.00000000000000022, som er veldig signifikant siden den er så nær 0. Så resultatet er altså signifikant. Vi er 99 % sikker på at disse to gruppene er forskjellige.

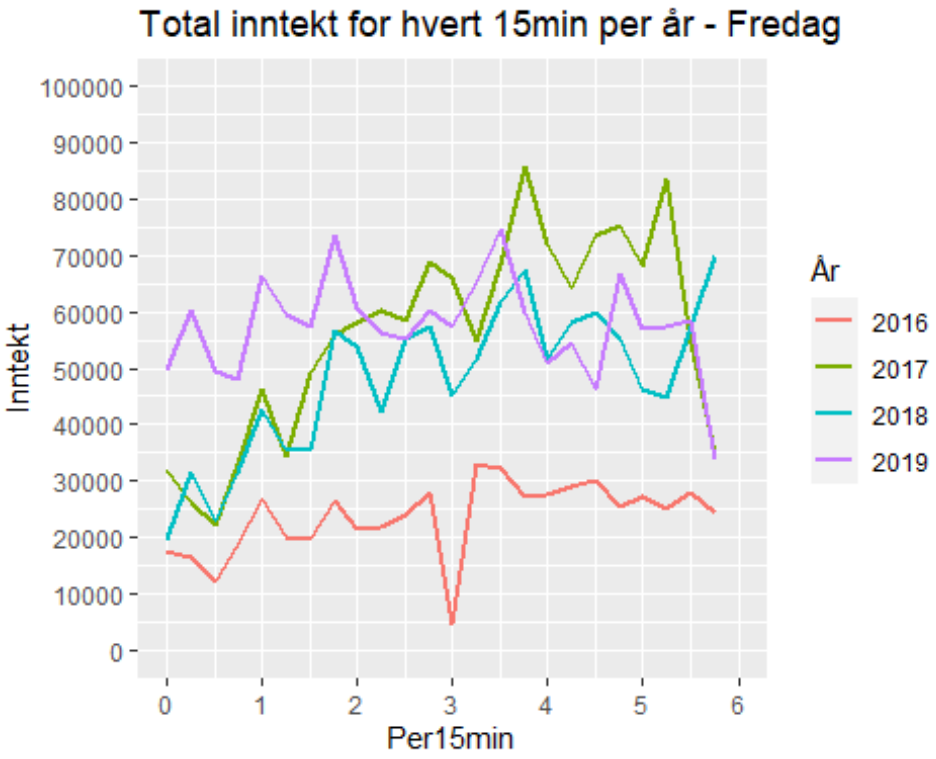
Til konklusjon ser vi at Bukta festivalen tjener mest på Pilsner siden gjennomsnittsforskjellen mellom produktet Pilsner er «nær nok» til null, derfor kan vi konkludere i praksis at pilser er mest ettertraktet. Bukta Beer og Cider er relativt like.

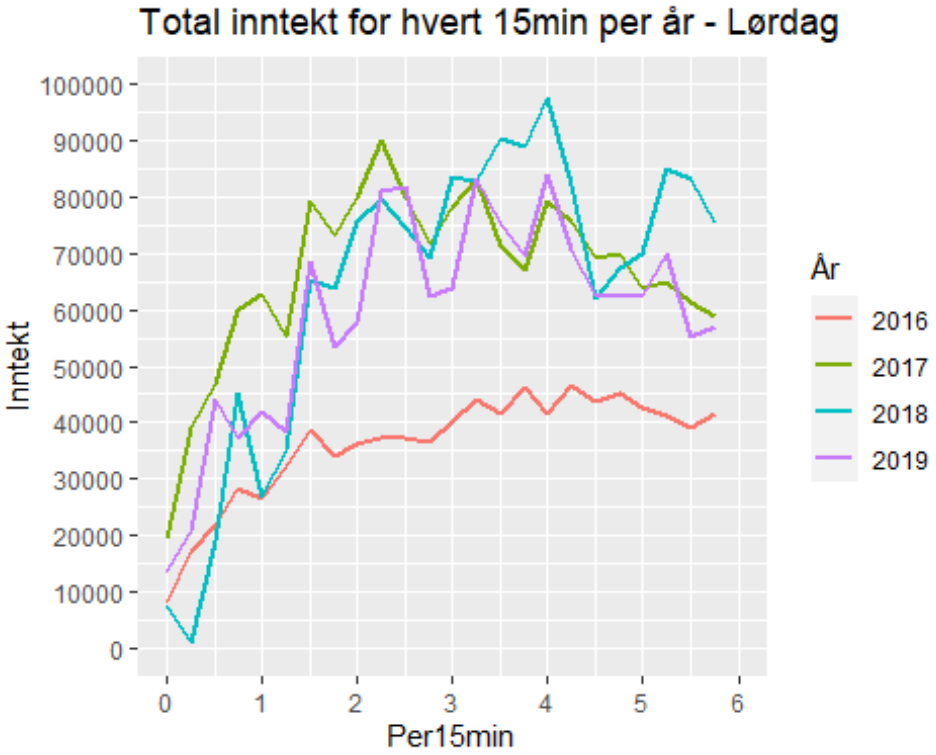
Total inntekt og inntekt pr 15 min

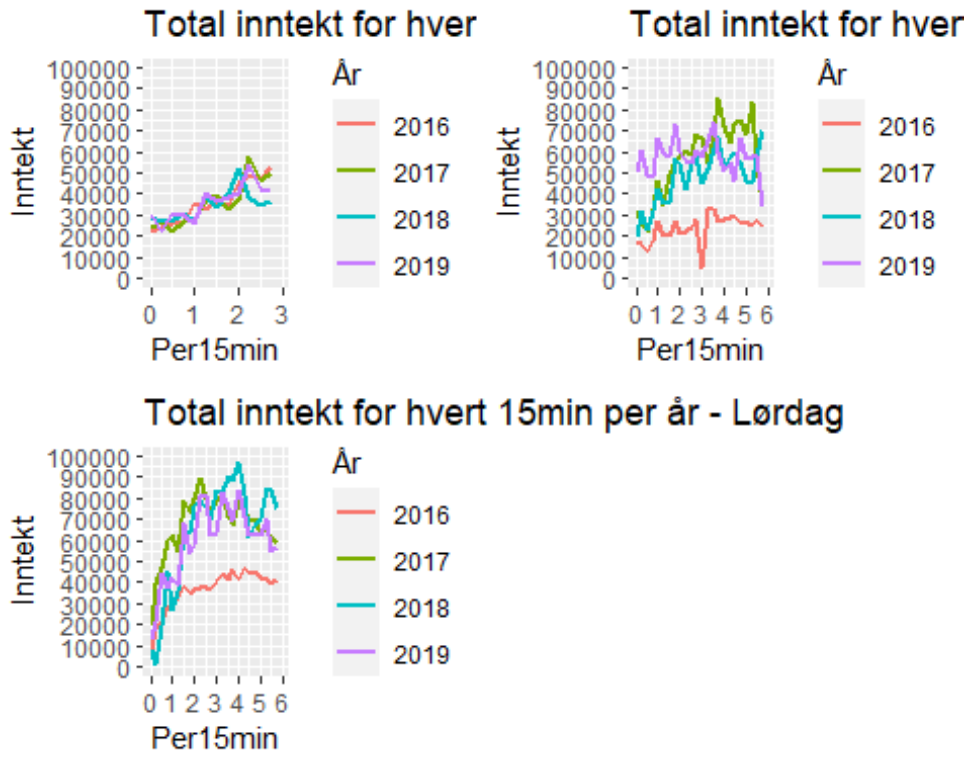
(Oppgave 3.1)

Først får du hver figur for seg og derreter samlet for bedre sammenligning.




Total inntekt for hvert 15min per år - Fredag
Inntekt
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
0
1
2
3
4
5
6
Per15min
År
2016
2017
2018
2019


Total inntekt for hvert 15min per år - Lørdag
Inntekt
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
0
1
2
3
4
5
6
Per15min
År
2016
2017
2018
2019



Her er 3 grafer som viser sammenhengen mellom total inntekt og hver 15 min av festivalen. En graf for hver dag, torsdag, fredag og lørdag, hvor hver viser en linje for hvert år. Vi skal nå se nærmere på hvilke trender vi ser i de ulike grafene og om det er noe forskjell mellom årene og dagene.

Vi starter med å analysere torsdagen. Det første vi ser er at per15min går kun til 2.75, altså antall timer fra start av konsert per 15 min. Dette betyr at 2 timer og 45 minutter etter konserten startet endte den, mens på fredager og lørdager varer konserten i 5 timer og 45 minutter. Det er mest sannsynlig slik fordi mennesker jobber i ukedagene og å feste på torsdag for så deretter dra på jobb på fredagen bakfull er ikke særlig populært i arbeidsmarkedet.

På fredag i 2016 har grafen et enormt bunnpunkt 3 timer inn i festivalen, dette er pga. problemer med betalingsterminalen, så vi ser derfor bort ifra det.

Lørdag er ukedagen som er mest innbringende fordi folk har fri før festivalen og dagen etter. Også her ser vi at år 2016 ligger lavest på grafen. Dette kan også skyldes videre problemer med betalingsterminal fra fredagen eller at prisen for drikke var rett og slett lavere enn årene etter. Ser man på prisstigningen gjennom de fire årene ser man at prisen på pilsner har steget med kr 4.- fra 2016 til 2017 og 2018 og kr 6,4.- fra 2018 til 2019. Det vil si en økning på kr 10.4 ,- fra år 2016 til 2019.

Under er tabellen for pris pr enhet for pils pr år.

```
 Pris  År     Produkt
-----  -----  --------
 68.0  2016   Pilsner
 72.0  2017   Pilsner
 72.0  2018   Pilsner
 78.4  2019   Pilsner
```

Alt i alt var 2016 det minst innbringende året, mens 2017 var det mest innbringende, etterfulgt av 2019 og deretter 2018, selv om det er snakk om minimale forskjeller.

Linær regresjon av inntekten

(oppgave 3.2)

Sammenligne inntekt mellom år og dager pr 15 min.

Kode for hva som er sammenlignet og resultatene ut under og forklaring til sist.

```
# Oppgave 3.2
try <- aggregate(Inntekt ~ Per15min + År + Dag, data = Bukta_data, sum)

try2 <- lm(Inntekt ~ Per15min + År + Dag, data = try)
summary(try2)


Call:
lm(formula = Inntekt ~ Per15min + År + Dag, data = try)

Residuals:
   Min     1Q Median     3Q    Max
-45330  -7509   -721   9497  31985

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept)    11418       2566   4.450  0.00001333129648765 ***
Per15min        5125        544   9.420 < 0.0000000000000002 ***
År2017         25065       2455  10.211 < 0.0000000000000002 ***
År2018         20502       2455   8.352  0.0000000000000602 ***
År2019         22634       2455   9.221 < 0.0000000000000002 ***
DagFredag       3111       2513   1.238                0.217
DagLordag      13209       2513   5.256  0.00000033193658708 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13440 on 233 degrees of freedom
Multiple R-squared:  0.5639,    Adjusted R-squared:  0.5527
F-statistic: 50.21 on 6 and 233 DF,  p-value: < 0.00000000000000022

coef(try2)
```

```
(Intercept)    Per15min      År2017      År2018      År2019   DagFredag
  11418.284    5124.581   25065.267   20501.800   22633.467    3110.986
  DagLordag
  13208.645
```

```
linearHypothesis(try2,'DagFredag=DagLordag')
```

```
Linear hypothesis test

Hypothesis:
DagFredag - DagLordag = 0

Model 1: restricted model
Model 2: Inntekt ~ Per15min + År + Dag

  Res.Df         RSS Df  Sum of Sq      F       Pr(>F)
1    234 47011438484
2    233 42117228701  1 4894209783 27.076 0.0000004288 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(try2,'År2017=År2018')
```

```
Linear hypothesis test

Hypothesis:
År2017 - År2018 = 0

Model 1: restricted model
Model 2: Inntekt ~ Per15min + År + Dag

  Res.Df         RSS Df Sum of Sq      F  Pr(>F)
1    234 42741985541
2    233 42117228701  1 624756841 3.4563 0.06427 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(try2,'År2018=År2019')
```

```
Linear hypothesis test

Hypothesis:
År2018 - År2019 = 0

Model 1: restricted model
Model 2: Inntekt ~ Per15min + År + Dag

  Res.Df         RSS Df Sum of Sq      F Pr(>F)
1    234 42253548784
2    233 42117228701  1 136320083 0.7541 0.3861
```

```
linearHypothesis(try2,'År2017=År2019')

Linear hypothesis test

Hypothesis:
År2017 - År2019 = 0

Model 1: restricted model
Model 2: Inntekt ~ Per15min + År + Dag

  Res.Df         RSS Df Sum of Sq      F Pr(>F)
1    234 42294638238
2    233 42117228701  1 177409537 0.9815 0.3229
```

Vi har inntekt som uavhengig variabel. Plusstegnet inkluderer per15min, år og dag i modellen som en prediktorer (uavhengige) variabler. Resultatet til modellen er at det er kun mulig å si noe om inntekten er forskjellig fra 2016 og de andre årene, man kan altså ikke sammenligne år 2018 og 2019 for eksempel. Vi ser også at alle variablene er veldig signifikante (som indikert med "***"), utenom for dag fredag. Vi kan dermed forkaste nullhypotesen siden p-verdien blir under signifikansnivået ofte 0,05 og 0,01.

"Residual standard error" eller residulaer på norsk er avsnittet som oppsummerer residualene og viser feilen mellom prediksjonen av modellen og de faktiske resultatene. Jo mindre resultat jo bedre. I denne sammenhengen fikk vi et ganske stort tall som vil si at regresjonsmodellen passer dårlig til et datasett.

I den linærene hypotesen mellom dag fredag og lørdag får vi en p-verdi som er mindre enn 0,001, derfor flagges den med tre stjerner (***) og det finnes dermed ingen samvarians mellom de to variablene.

Neste er den linære hypotesen mellom år 2017 og 2018 hvor man får en femprosents signifikansnivå, som vil si at man aksepterer så mye som fem prosent sannsynlighet (p=0,06427). En p-verdi på 0,06 vil i så fall indikere at det er 6% sannsynlighet for at det vi har observert bare er tilfeldig.

I den linære hypotesen mellom år 2018 og 2019 får man en trettiprosents signifikansnivå, som vil si at man aksepterer så mye som tretti prosent sannsynlighet (p=0,3861) for at nullhypotesen stemmer og at de tendensene man har observert er tilfeldige. Dersom p-verdien er innenfor et visst signifikansnivå, sier man at den sammenhengen som er observert «er signifikant». så i denne sammenhengen sier man at «sammenhengen er signifikant innenfor et trettiprosents signifikansnivå».

Mellom år 2017 og 2019 er p-verdien så godt som lik den linære hypotesen mellom år 2018 og 2019. En p-verdi på 0.3229 betyr at det er mer enn 30 prosent sjanse for å se disse resultatene i verden når nullhypotesen er sann.

Været og salg av drikkevarer

(Oppgave 4)

```
   Per15min         Nedbør       Luft_temperatur    Solskin
 Min.   :0.000   Min.   :0.0000   Min.   : 9.10   Min.   : 0.00
 1st Qu.:1.188   1st Qu.:0.0000   1st Qu.:12.57   1st Qu.: 0.00
 Median :2.375   Median :0.0000   Median :14.70   Median :34.50
 Mean   :2.575   Mean   :0.1383   Mean   :15.30   Mean   :32.47
 3rd Qu.:4.000   3rd Qu.:0.0000   3rd Qu.:17.20   3rd Qu.:60.00
 Max.   :5.750   Max.   :2.7000   Max.   :25.00   Max.   :60.00
     Vind          Gjester           Dag          År
 Min.   :2.500   Min.   :   0.0   Torsdag:48   Length:240
 1st Qu.:3.875   1st Qu.: 128.2   Fredag :96   Class :character
 Median :4.800   Median :2035.0   Lordag :96   Mode  :character
 Mean   :4.548   Mean   :2063.2
 3rd Qu.:5.100   3rd Qu.:3330.2
 Max.   :7.700   Max.   :5723.0
   Inntekt
 Min.   : 1080
 1st Qu.:32184
 Median :46230
 Mean   :48192
 3rd Qu.:62863
 Max.   :97612

NULL
```

Fra disse dataene kan de se ut til at nedbør, luft temperatur og solskinn har påvirkning på hvor mange enheter som blir solgt. Men ikke vind.

Når vi legger til variabler i regresjonsanalyse vil adjusted R-squared alltid gå oppover. Så en kan ikke bruke det for å se hvor sikker vi kan være på disse dataene og derfor er vi ikke helt sikre på om alle disse variablene har særlig effekt på salget.

Forberedelse til pils-salg

(Oppgave 5)

Koder for hvordan utregnet og resultat med forklaring under.

```
#Lager pils-verdi fra Bukta_data
Pils <- Bukta_data %>%
  #Grupperer for år
  group_by(År) %>%
  #Filtrerer for hvilken dag, hvilken time og hvilket produkt
  filter(Dag == "Lordag",
         Time == 20,
         Produkt == "Pilsner") %>%
  #Lager en pils-verdi hvor vi summerer antall pils
  summarise(pils = sum(Antall)) %>%
  #Tar så gjennomsnittet av antall pils som blir kjøpt hver lørdag klokken 20
  summarise(mean(pils)) %>%
  pull()
```

```r
poisson <- qpois(0.95, Pils, lower.tail = TRUE)

ppois(2144, lambda = Pils)
```

```
[1] 0.9508895
```

Det må gjøres klart 2144 pils for at de skal være 95% sikre på at det ikke blir bestilt mer enn dette.

Oppsummering

Pils er drikkevaren som gir størst inntekt, 2017 var det beste året ved samlet inntekt, og været har egentlig lite å si for salget så fremst gjestene ikke blir blåst bort. Og for å være forberedt til salgstoppen kl 20 til 21 bør det gjøres klart 2144 pils.