Titanic

Prosjekt econ 3170/4170

Pakker

I prosjetet kommer vi hovedsaklig til å benytte oss av to pakker: tidyverse og tidymodels. Begge pakkene er samlinger av mange ulike pakker. Tidyverse skal vi hovedsaklig bruke til å manipulere data og visualisering. Tidymodels skal vi bruke til maskinlæring.

```
library(tidyverse)
```

```
Warning: package 'tidyverse' was built under R version 4.3.3
```

```
Warning: package 'ggplot2' was built under R version 4.3.3
```

```
Warning: package 'tidyr' was built under R version 4.3.3
```

```
Warning: package 'forcats' was built under R version 4.3.3
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```r
library(tidyverse)
```

Data

```r
titanic <- read_csv("train.csv")
```

```
Rows: 891 Columns: 12
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (5): Name, Sex, Ticket, Cabin, Embarked
dbl (7): PassengerId, Survived, Pclass, Age, SibSp, Parch, Fare

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Vi benytter oss av data fra kaggel: https://www.kaggle.com/competitions/titanic/data.

Datasettet bester av 12 kolonner. Her kommer en liten oversikt over hva de betyr:

- Survival: Om vedkommende overlevde: 1 = Ja, 0 = Nei.
- Pclass: 1, 2 eller 3.klasse.(1 er best)
- Sex: Kjønn
- Age: Alder
- Sibsp: Antall søsken/ ektefeller på Titanic
- Parch: Antall foreldre/ barn på Titanic
- Ticket: Bilettnummer
- Fare: Pris
- Cabin: Lugarnummer
- Embarked: Hvor de gikk ombord: C = Cherbourg, Q = Queenstown, S = Southampton.
- Name
- PassengerId:

```r
head(titanic)
```

```
# A tibble: 6 x 12
  PassengerId Survived Pclass Name    Sex     Age SibSp Parch Ticket  Fare Cabin
        <dbl>    <dbl>  <dbl> <chr>   <chr> <dbl> <dbl> <dbl> <chr>  <dbl> <chr>
1           1        0      3 Braund~ male     22     1     0 A/5 2~  7.25 <NA>
2           2        1      1 Cuming~ fema~    38     1     0 PC 17~ 71.3  C85
3           3        1      3 Heikki~ fema~    26     0     0 STON/~  7.92 <NA>
4           4        1      1 Futrel~ fema~    35     1     0 113803 53.1  C123
5           5        0      3 Allen,~ male     35     0     0 373450  8.05 <NA>
6           6        0      3 Moran,~ male     NA     0     0 330877  8.46 <NA>
# i 1 more variable: Embarked <chr>
```

Pclass og Fare virker veldig spennede, de fleste har nok sett filmen og der blir tydligfremstilt at hvis du er i første klasse har du en større sjanse for å overleve. Samme med Fare, hvis du har betalt mye for biletten indikerer dette en høyere klasse som kan hinte til en større sannsynlighet for å overleve.

Cabin og Ticket er nok også verdier som har en sterk tilknytting med hvilken klasse du er i og hvor mye du betalte. Likevel blir det vannskelig å skulle si om noen overlever p.ga hvilket bilettnummer man har. Ved første øyekast ser det ut som at lugar variablen har mange manglende verdier. Hva den faktiske lugaren var er nok kanskje ikke altfor interessant, men hvilket dekk den ligger på kunne vært av stor interesse. Ettersom at dekket har mye å si for om du klarte å komme deg ut i tide.

Alder og kjønn er igjen ganske interessante variabler. Vi kjenner jo alle til "Kvinner og barn først". Her vil det nok være interessant å manipulere dataen for å fremmheve om vedkommende er et barn eller ikke. En annen dimensjon kan også være om personen blir ansett som gammel eller ikke, ettersom at disse gruppen antaglivis får prioritet under evakuering av skipet. Samt også kombinere disse verdiene med hvilken klasse de var i.

Navn er litt "tricky". Man kan gjerne tenke seg at naven til noen ikke er av stor betydning når det kommer til overlevelse som forsåvidt kan stemme. Likevel inkluderer navene tittler so f.eks mrs og master, og dette kan igjen være interessant.

Embarked er igjen variabel det blir litt vannskelig å si noe om, ettersom at det er kun tre forskjellige steder, som egentlig ikke burde ha særlig stor betydning. På den andre siden kan det være en korrelasjon mellom hvor man gikk om bord og hvilken klassen man er i.

Parch og Sibsp er en ganske interessante variabler. Kan det være noe sammenheng mellom hvor stor familie du har ombord og overlevelse.

PassengerId vikrer ikke veldig relevant, den virker litt mer som en variabel for å kunne referere til passasjer og den har nok mest sannsynlig ikke noe å gjøre med overlevelsen.

Inspirasjon: https://www.kaggle.com/code/allohvk/titanic-advanced-eda?scriptVersionId=77739368

Utforsking av dataen

Behandling av manglende verdier(NA's)

Bruker skim fra pakken skimr for å få en oversikt over dataen:

```
skimr::skim(Titanic)
```

Table 1: Data summary

Name	Titanic
Number of rows	32
Number of columns	5
_____	
Column type frequency:	
factor	4
numeric	1
_____	
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
Class	0	1	FALSE	4	1st: 8, 2nd: 8, 3rd: 8, Cre: 8
Sex	0	1	FALSE	2	Mal: 16, Fem: 16
Age	0	1	FALSE	2	Chi: 16, Adu: 16
Survived	0	1	FALSE	2	No: 16, Yes: 16

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Freq	0	1	68.78	136	0	0.75	13.5	77	670	

Her får vi vite at det 891 rander og at 5 av de tolv kolonnene er av typen "character". Det som er av mest interesse i outputen er kolonnen med "n_missing". Kolonnen "Cabin" har 687 manglende verdier. Det er rimelig å tenke seg at lugarnummeret ikke har så mye å si for overlevelsen, og at heller "Pclass" er viktigere for analysen. Med så mange manglende verdier blir det også vanskelig å lage nye variabler med hvilket dekk man var på. Da måtte man eventuelt gjort noen antagelser om at første klasse var på dekk A og B, men det blir jo nesten

det samme som pclass og derfor ikke hensiktsmessig. Ettersom at datasettet ikker er så stort er det mer hensiktsmessig å fjerne hele kolonnen fremfor å fjerne alle randene med manglende verdier.

Alder er også en variabel som mangler 177. Denne kolonnen er litt mer problematisk. Som drøftet ovenfor kan man tenke seg at alder har mye å si for analysen. Datasettet er lite så det blir dumt å fjerne disse radene. En muligheter er å kunne fylle disse verdiene med gjennomsnittsalderen. En annen metode som blir diskutert i denne artikkelen: https://www.kaggle.com/code/allohvk/titanic-missing-age-imputation-tutorial-advanced, er f.eks. å se på tittel som "Master" for å indentifisere unge gutter.

```
mean_age_master <- titanic |>
  filter(grepl("Master", Name)) |>
  summarise(mean(Age, na.rm = T)) |>
  pull()


titanic <- titanic |>
  mutate(Age = ifelse(is.na(Age) & (grepl("Master", Name)),
                      yes = round(mean_age_master,0),
                      no = Age))
```

Her finner vi gjennomsnittlig alder for personer med "Master" som tittel og implemterer dette i datasettet med en ifelse-statment.

```
titanic |>
  select(Age) |>
  summarise(sum(is.na(Age))) |>
  pull()
```

```
[1] 173
```

Likevel har vi fortsatt 173 manglende verdier. Andre fremmgangsmetoder som også blir nevnt i artikkelen er å bruke gjennomsnittsalderen for gitte klasser og kjønn.

```
mean_age_menn_p <- rep(0,3)
mean_age_kvinne_p <- rep(0,3)
for (i in 1:3)
{
  mean_age_menn_p[i] <- titanic |>
    filter(Pclass == i &  Sex == "male") |>
```

```
    summarise(round(mean(Age, na.rm = T),0)) |>
    pull()

  mean_age_kvinne_p[i] <- titanic|>
    filter(Pclass == i &  Sex == "female") |>
    summarise(round(mean(Age, na.rm = T),0)) |>
    pull()
}

mann_p3 <- "Moran, Mr. James"
mann_p2 <- "Williams, Mr. Charles Eugene"
mann_p1 <- "Woolner, Mr. Hugh"

kvinne_p3 <- "Moran, Miss. Bertha"
kvinne_p1 <- "Thorne, Mrs. Gertrude Maybelle"
kvinne_p2 <- "Keane, Miss. Nora A"

navn_med_na <- c(mann_p1, mann_p2, mann_p3,
                 kvinne_p1, kvinne_p2, kvinne_p3)

aldere <- c(mean_age_menn_p, mean_age_kvinne_p)

for (i in 1:3)
{
titanic <- titanic |>
   mutate(Age = ifelse(is.na(Age) & Sex == "male" & Pclass == i,
                       mean_age_menn_p[i],
                       Age)) |>
  mutate(Age = ifelse(is.na(Age) & Sex == "female" & Pclass == i,
                      mean_age_kvinne_p[i],
                      Age))
}


for (i in 1:6)
{
  person_ald <- titanic |>
    filter(Name == navn_med_na[i]) |>
    pull(Age)

  stopifnot(person_ald == aldere[i])
}
```

Her finner vi først gjennomsnittsalderne på for kjønn gitt pclass, deretter finner vi seks tilfeldige navn for en gitt pclass som vi vet har manglende verdier. Så fyller vi datarammen med disse verdiene. For å sjekke at verdiene ble riktige kjører vi en løkke med en "stopifnot" metode for å sjekke at personene fikk riktig verdi.

Nå har vi kun 2 manglende verdier for embarked. Det burde nok gå greit å fjerne dem, men vi kan også prøve å utforske dataen før vi gjør det.

```
titanic |>
  filter(is.na(Embarked))
```

```
# A tibble: 2 x 11
  PassengerId Survived Pclass Name          Sex     Age SibSp Parch Ticket  Fare
        <dbl>    <dbl>  <dbl> <chr>         <chr> <dbl> <dbl> <dbl> <chr>  <dbl>
1          62        1      1 Icard, Miss.~ fema~    38     0     0 113572    80
2         830        1      1 Stone, Mrs. ~ fema~    62     0     0 113572    80
# i 1 more variable: Embarked <chr>
```

Det er her snakk om to kvinner som begge overlevde og var i første klasse. Datasette er ikke kjempe stort og det ville vært dumt å skulle miste rader med folk som overlevde. Vi kan utforske litt mer.

```
titanic |>
  filter(Sex == "female" & Pclass == 1) |>
  group_by(Embarked, Pclass) |>
  summarise(mean_survived = mean(Survived), antall = n())
```

```
`summarise()` has grouped output by 'Embarked'. You can override using the
`.groups` argument.
```

```
# A tibble: 4 x 4
# Groups:   Embarked [4]
  Embarked Pclass mean_survived antall
  <chr>     <dbl>         <dbl>  <int>
1 C             1         0.977     43
2 Q             1         1          1
3 S             1         0.958     48
4 <NA>          1         1          2
```

Det vi kan tolke fra output er at de fleste kvinner som er i førseklasse overlevde, uavhengig av hvor de kom ombord. Derfor tenker vi det er rimelig å fylle de manglende verdiene med tilfeldig trekk mellom "C" of "S", ettersom at de er de verdien som er mest sannsylige.

```r
tilfeldig_embarked <- sample(c("C", "S"), size = 1)

titanic <- titanic |>
  mutate(Embarked = ifelse(is.na(Embarked),
                           yes = tilfeldig_embarked,
                           no = Embarked))
```

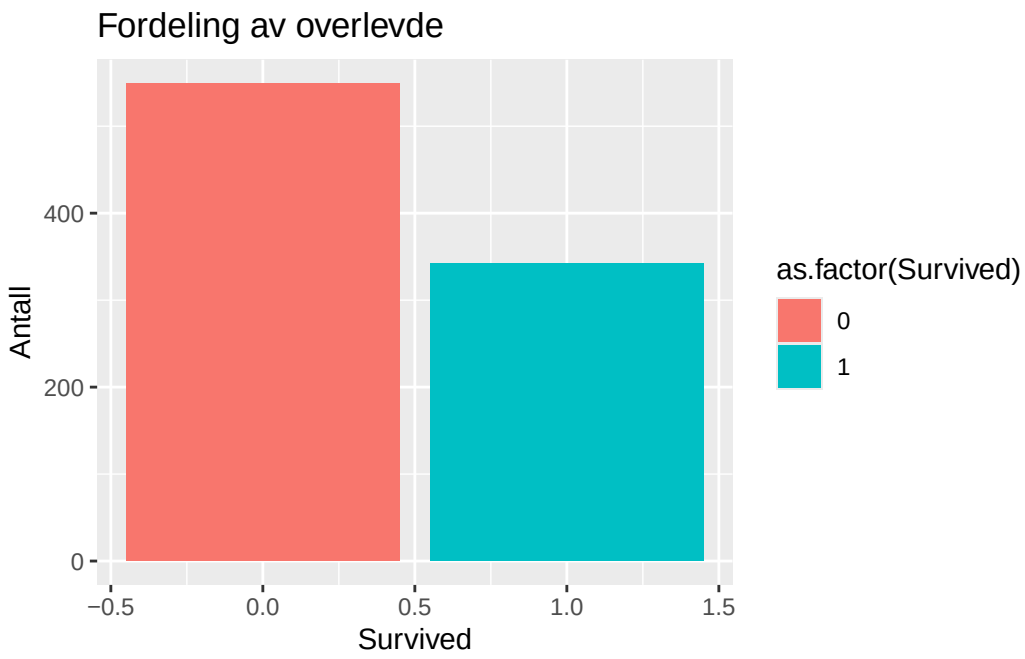
Visualisering av data

Nå som vi har bearbeidet datarammen har vi lyst til å utforske dataen litt mer. Dette gjør vi for å få en bedre forståelse for ulike sammenhenger i dataen.

```r
titanic |>
  group_by(Survived) |>
  summarise(Antall = n())
```

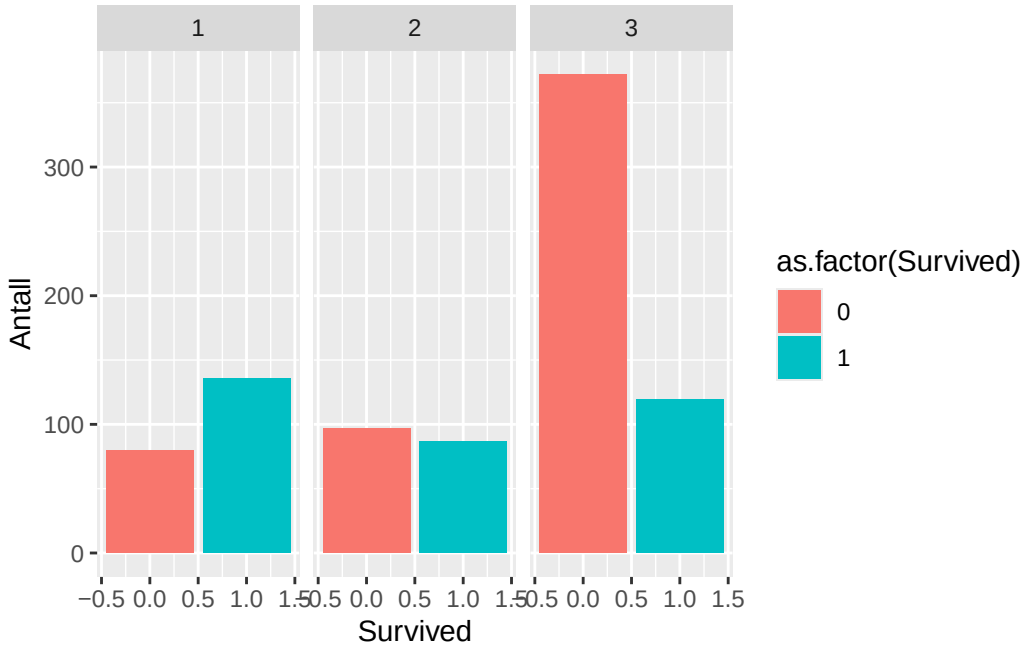
```
# A tibble: 2 x 2
  Survived Antall
     <dbl>  <int>
1        0    549
2        1    342
```

```r
titanic |>
  ggplot(aes(x = Survived, fill = as.factor(Survived))) +
  geom_bar() +
  ylab("Antall") +
  ggtitle("Fordeling av overlevde")
```



I følge dataen vi har er det kun 342 person som overlevde, mens det var 549 som ikke gjorde det. Har dette noe å si for f.eks. Pclass.

```
titanic |>
  ggplot(aes(x = Survived, fill = as.factor(Survived))) +
  geom_bar() +
  ylab("Antall") +
  facet_wrap(vars(Pclass))
```

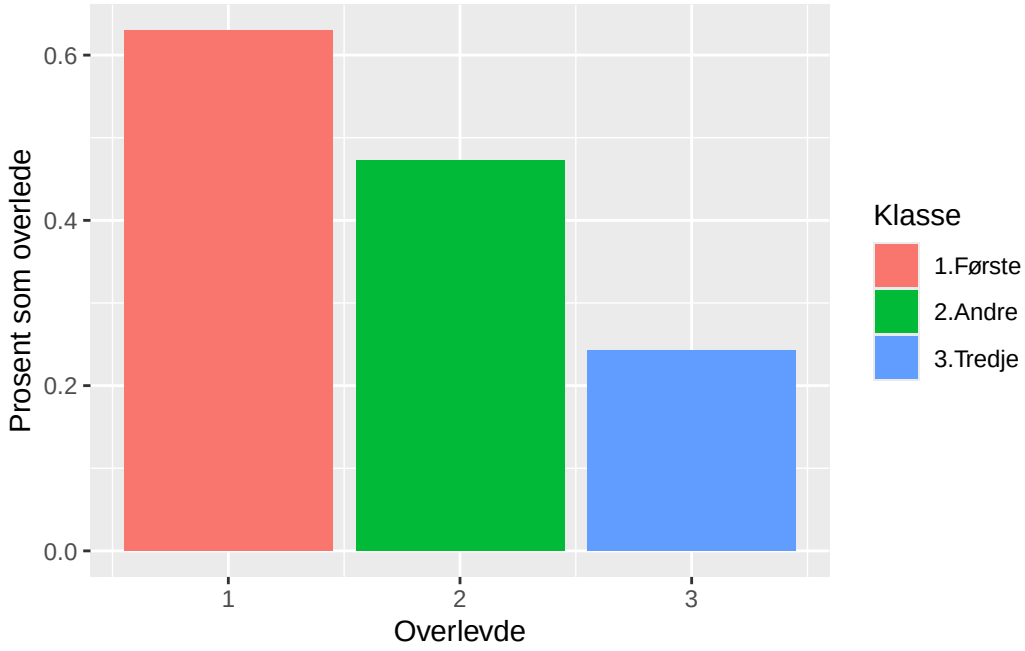


Antallet som overlevde ser ut til å være ganske likt mellom klassene. Det er nok mer interesant å se på ovelevelsen som en andel av antallet.

```
titanic |>
  group_by(Pclass) |>
  summarise(Survived_share = sum(Survived)/n())
```

```
# A tibble: 3 x 2
  Pclass Survived_share
   <dbl>          <dbl>
1      1          0.630
2      2          0.473
3      3          0.242
```

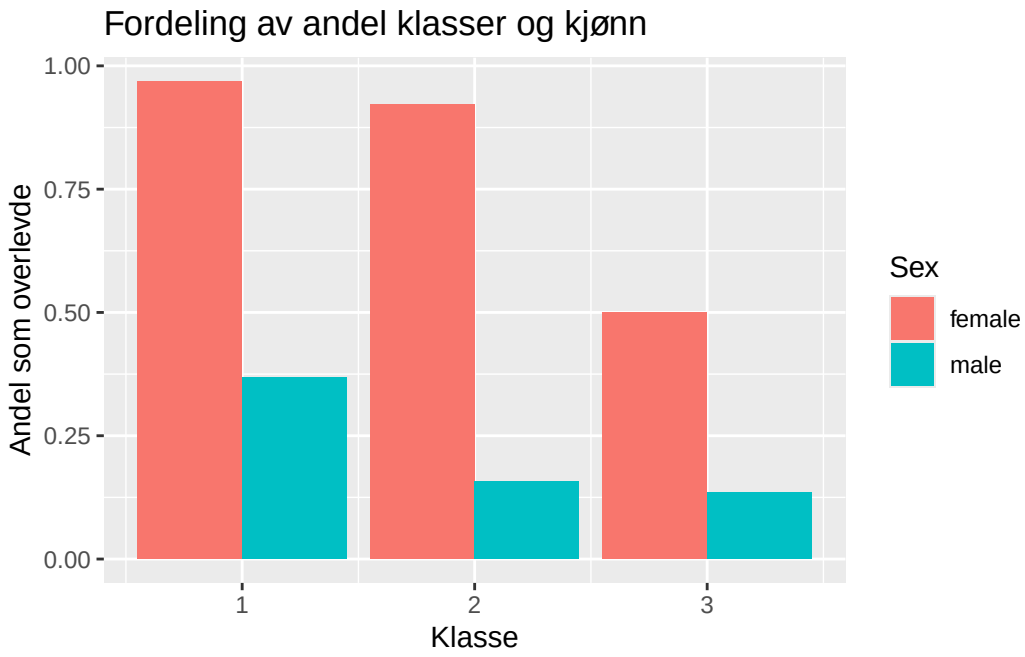
```
titanic |>
  group_by(Pclass) |>
  summarise(Survived_share = sum(Survived)/n()) |>
  mutate(Klasse = c("1.Første", "2.Andre", "3.Tredje")) |>
  ggplot(aes(x = Pclass,y = Survived_share, fill = Klasse)) +
  geom_col() +
  ylab("Prosent som overlede") +
  xlab("Overlevde")
```



Her får vi et tydligere bilde over overlevelse og hvilken klasse du er i. Vi ser at andel som overlevde blir betraktlig større desto høyre klasse klasse man er i. Kan kjønn ha noe betydning?

```
titanic |>
  group_by(Pclass,Sex) |>
  summarise(Survived_share = sum(Survived)/n())|>
  ggplot(aes(x = Pclass, y =Survived_share, fill = Sex)) +
  geom_col(position = position_dodge())+
  ylab("Andel som overlevde") +
  xlab("Klasse") +
  ggtitle("Fordeling av andel klasser og kjønn")
```

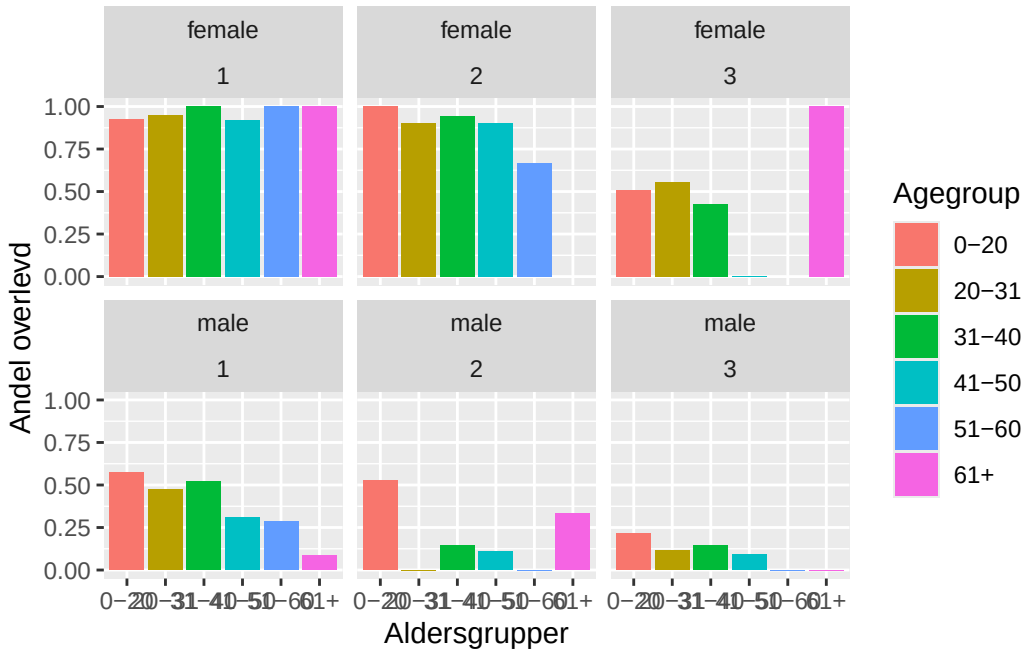
```
`summarise()` has grouped output by 'Pclass'. You can override using the
`.groups` argument.
```



Her ser vi at det er en tydelig sammenheng mellom klasse og kjønn for overlevelse. Det ser ut som at kvinner i første og andre klasse har realtivt stor sannsynlighet for å overleve. Kvinner i tredje klasse ligger på rundt femti prosent. Andelen menn som overlevde er betraktlig lavere i alle klasser samenlignet med kvinner. Andel menn som overlevde i andre og tredje klasse er omtrent halvparten av det den er i første klasse, likevel er det å være mann i første klasse "dårligere" enn å være kvinne i tredjeklasse med tanke på overlevelse. Det viser seg at kjønn har større påvirkning på overlevelse fremfor hvilken klasse du er i, dette stemmer jo overens med at kvinner og barn går først når det evakueres. Det er kanskje da naturlig å se på hvordan alder spiller en rolle.

```
titanic |>
  mutate(
    Agegroup = cut(
      Age,
      breaks = c(-Inf,20, 30, 40, 50, 60, Inf),
      labels = c("0-20", "20-31", "31-40", "41-50", "51-60", "61+"))) |>
  group_by(Agegroup,Sex, Pclass) |>
  summarise(Andel = sum(Survived)/n())|>
  ggplot(aes(x = Agegroup, y= Andel, fill = Agegroup))+
  geom_col() +
  xlab("Aldersgrupper")+
  ylab("Andel overlevd")+
  facet_wrap(vars(Sex, Pclass))
```

```
`summarise()` has grouped output by 'Agegroup', 'Sex'. You can override using
the `.groups` argument.
```



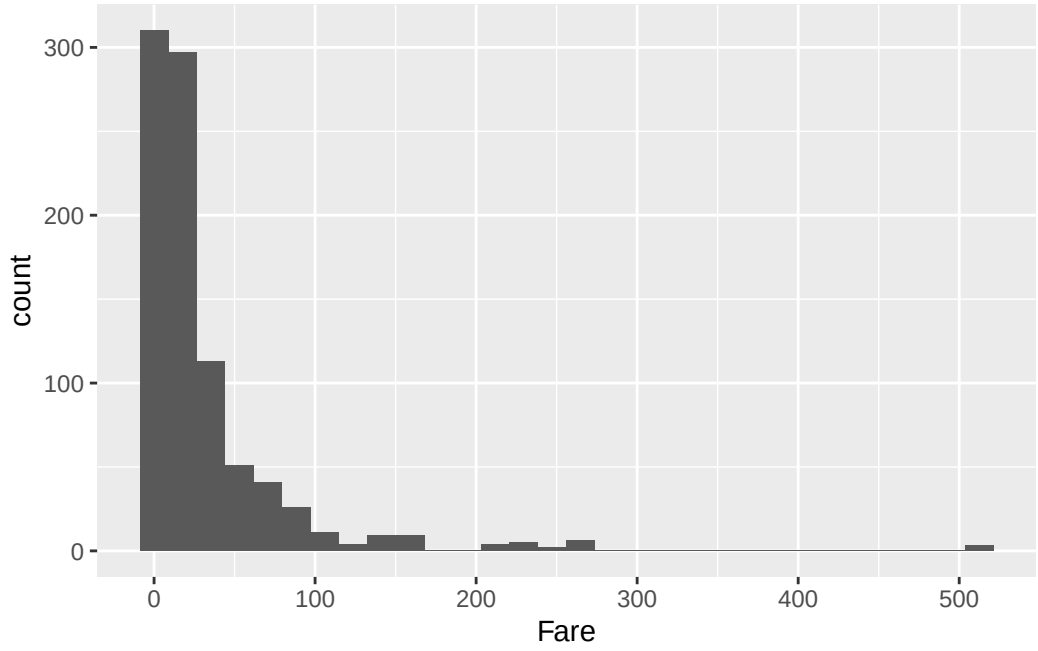
Deler inn i aldersgrupper for å letter kunne tolke dataen

Her ser vi mye av det samme som i plottet ovenfor, men med litt mer innsikt. Uansett aldersgruppe er det en høy andel av kvinner som overlever i første og andre klasse, derimot i tredje klasse var det en betraktlig mindre andel som overlevde uansett aldersgruppe, med untak av 61 år og eldre, noe som kanskje skyler et lite utvalg i denne gruppen. Menn ser vi at andelen overlevde er ganske lav uansett aldersgruppe. Likevel ser vi at de yngste mellom 0-20 har en liten fordel. Tendensen ser nærmest ut som at desto eldre du blir desto lavere blir andelen som overlever. Grafene gir oss ikke særlig tydlig svar på om alder egentlig spiller en rolle, men at det forsatt ser ut som at kjønn er den variabelen som har størst betydning.

Hva med prisen hvor mye de betalte, har det noe å si for overlevelsen.

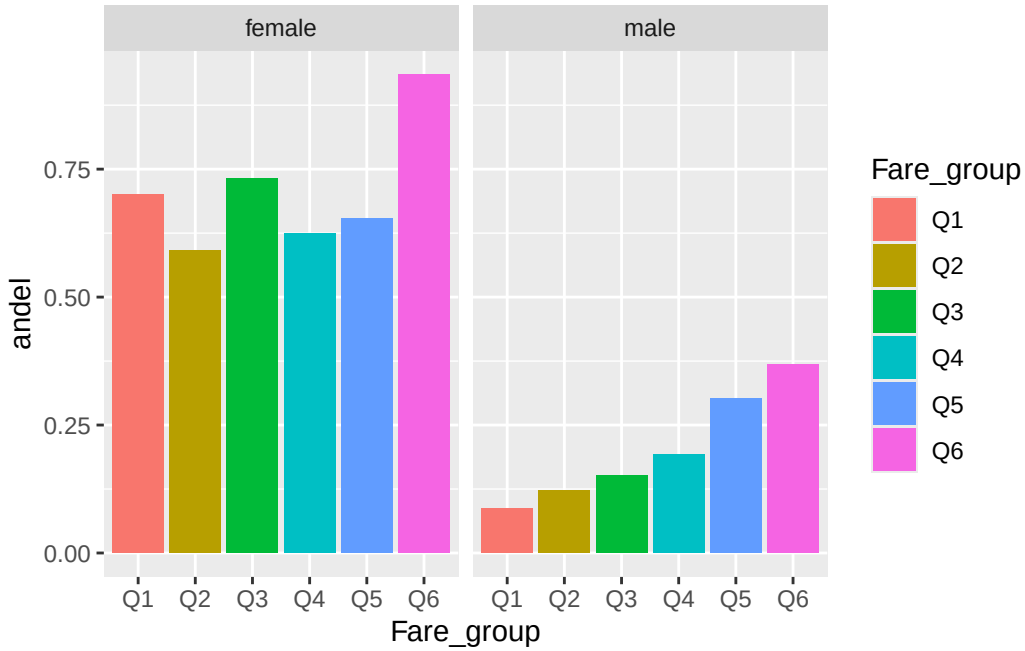
```
titanic |>
  ggplot(aes(x = Fare)) +
  geom_histogram()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
titanic |>
  filter(Fare < 200) |>
  mutate(Fare_group = ntile(Fare, 6)) |>
  mutate(Fare_group = factor(Fare_group, labels = c("Q1", "Q2", "Q3", "Q4", "Q5", "Q6"))) |>
  group_by(Fare_group, Sex) |>
  summarise(andel = sum(Survived)/n()) |>
  ggplot(aes(x = Fare_group, y = andel, fill = Fare_group)) +
  geom_col() +
  facet_wrap(vars(Sex))
```

```
`summarise()` has grouped output by 'Fare_group'. You can override using the
`.groups` argument.
```



Ettersom at fordlingen over hvor mye hver person betalte er høyre vridd har vi prøvd å dele dem inn i seks grupper, litt som for alder. Vi ser også at det er en del uteliggere, derfor har vi prøvd å filtrere disse bort. Det vi tolker fra grafene er at jo mer folk betalte for reisen desto større andel overlevde, spessielt hos menn. Funnet samsvarer også med grafen over at andelen menn i første klasse hadde en større andel overlevende. Tendensen er ikke like sterk hos kvinner. Noe som tyder på at det å være kvinner taler for at bilettprisen har mindre betydning for dem enn det den har hos menn.

```
titanic |>
  group_by(Embarked, Pclass) |>
  summarise(Antall = n())
```

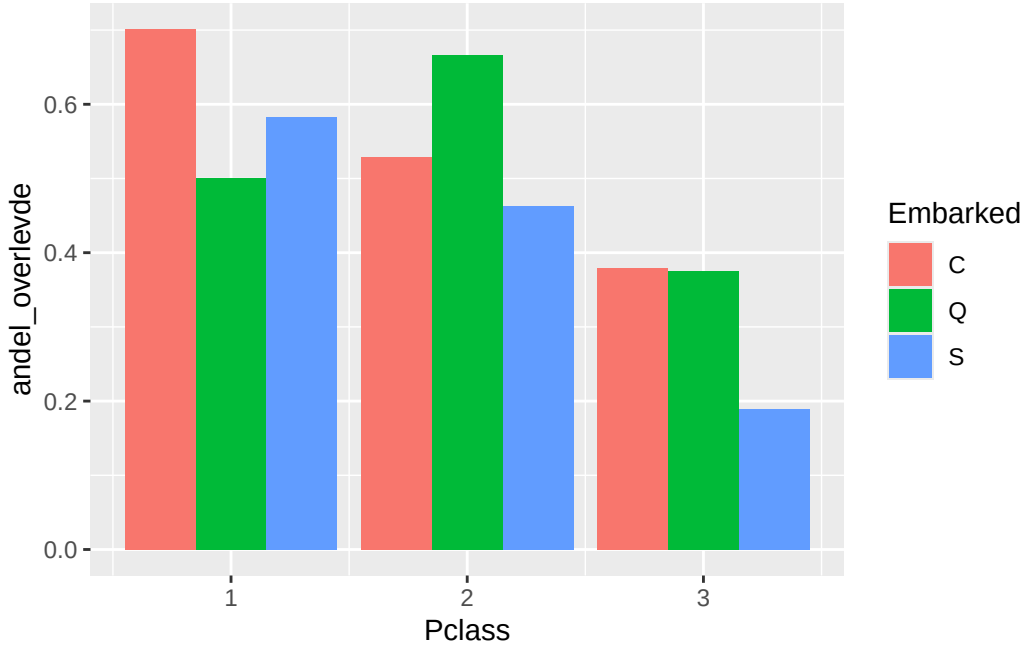
```
`summarise()` has grouped output by 'Embarked'. You can override using the
`.groups` argument.
```

```
# A tibble: 9 x 3
# Groups:   Embarked [3]
  Embarked Pclass Antall
  <chr>     <dbl>  <int>
1 C             1     87
2 C             2     17
3 C             3     66
4 Q             1      2
```

```
5 Q              2      3
6 Q              3     72
7 S              1    127
8 S              2    164
9 S              3    353
```

```
titanic |>
  group_by(Embarked, Pclass) |>
  summarise(andel_overlevde = sum(Survived)/ n()) |>
  ggplot(aes(x = Pclass, y = andel_overlevde, fill = Embarked )) +
  geom_col(position = position_dodge())
```

```
`summarise()` has grouped output by 'Embarked'. You can override using the
`.groups` argument.
```



Plotter for hver enkelt pclass, ettersom at en stor andel av de som kommer fra Queenstown går rett i tredje klasse noe som har en tydlig sammenheng med overlevelse. Grafen viser ikke noe særlig sammenheng mellom hvor du gikk ombord og overlevelse. Dette gir gi også mening ettersom at, det egentlig ikke burde ha noe å si,.

Neste:

Sjekke fair og embarked

Nye variabler: Familysize. Aldersgruppe?? fair_class

Sette opp modeller

Analyser