ag_comments

```
library(flextable)
```

1 Generelt

Fin litteraturgjennomgang (se mine forslag for bruk av inline-sitering for å gjøre dette lettere). Flott gjennomgang av hva dere finner i data. Det avsluttende avsnittet *Samsvarer det som fremkommer fra den deskriptive statistikken med funnene fra literaturen?* er virkelig bra.

Dere har utført et grundig og flott arbeid så jeg har prøvd å gjøre tilsvarende i min tilbakemelding. Jeg har prøvd å vise hvordan dere bedre kan utnytte mulighetene i Quarto både når det gjelder kryssreferanser og siteringer. I tillegg bør dere prøve å legge av dere uvaner fra Word med å formatere overskrifter. Bruk de 6 overskriftsnivåene ut fra logisk organisering og stol på de typografiske valgene som er gjort. Hvis dere virkelig mener å ha gode argumenter for å endre på dette kan det gjøres vha. css og scss for html output, templates og styles for Word og lasting av sty-filer i LaTeX. Generelt vil jeg ikke anbefale dette med mindre dere har akutt behov for å finne noe å fylle dagene med ;-)

I punkt-listingen nedenfor har jeg prøvd å dokumentere endringene jeg forslår. Noe er sikkert glemt. Det finner dere ved å sjekk diff på Github eller History under Git-fanen.

Jeg kan godkjenne assignment 2 uten problem.

2 Andre kommentarer

- Litt usikker på hvorfor dere har valgt å bruke `output: bookdown::pdf_document2: number_sections: true` Hvis det er for å få nummererte avsnitt vil jeg heller anbefale Quarto løsningen. bookdown::pdf_document2 benytter rmarkdown som fremdeles er støttet, men utviklingen går mot det nye systemet Quarto. Jeg har lagt inn en Quarto løsning som gir nummererte avsnitt.
- Jeg vil anbefale dere å gjøre mer bruk av såkalt «inline citation» som APA støtter. Se mitt forslag til endring mht. sitering i avsnittene *Lønn og høyde* og *Lønn og bmi*. Gjør at siteringene blir mer naturlig integrert med den flytende teksten.

- Det er generelt lurt å bruke informative «citation keys», f.eks vil britt2009 istedenfor published2009 gjøre det lettere å se hvilket paper man viser til. BetterBibTeX utvidelsen til Cotero kan lage slike «cite keys» automatisk.
- Kan være greit å sitere R og modelr pakken. Bruker man funksjonen `citation()` får man ferdig bibitem for R som bare kan kopieres inn i bib filen. Jeg har bare lagt til «cite key-en» R2024.
- Tilsvarende for *modelr* er `citation("modelr")`. Her har jeg lagt til `modelr2023` som «cite-key».
- Første setning i avsnittet *Lønn og kjønn* er: «Studier fremhever kjønnsforskjeller i lønn som et vedvarende problem.» Her burde dere hatt noen kildehenvisninger til disse studiene.
- Bruk overskriftnivåene i Quarto, header 1-6. Tenk på logisk struktur, ikke utseende. Hvis dere mener at standard formatering ikke holder mål (det ligger typografisk kunnskap bak utformingen av disse så dere bør ha en god grunn) er det mulig å endre disse. Dette er hovedtanken bak strukturerte tekstsystem (som LaTeX, html, quarto etc.). Forfatteren skal bekymre seg om struktur. Utseende skal ha fornuftige defaults som tar utgangspunkt i typografisk tradisjon, men kan overstyres hvis man virkelig ønsker dette.
- Ikke legg inn formatering (bold i overskriften Innleding) i f.eks overskrifter. Jobb med defaults og så kan dere eventuelt overstyre default formatering til slutt hvis dere virkelig mener det er gode argumenter for dette.
- Jeg har forsøkt å endre overskriftene etter logisk struktur.
- Libraries bør lastes helt i starten av dokumentet slik man lett ser hvilke pakker som må være innstallert for å kjøre dokumentet. Gir man chunken navnet setup (vha. `#| label: setup`) vil denne bli kjørt hver gang man kjører dokumentet. Bruker man da en ny funksjon fra en pakke man har lagt til i setup vil denne «automatisk» virke.
- Jeg har lagt inn litt kode som bruker vtable::st() for å generere tabeller. Funksjonen gir kompakte tabeller med mye informasjon og krever lite arbeid.
- `st()` har som default at den setter tittel på tabellene. Siden vi bruker Quarto til å gjøre dette får vi dobbelt opp. For å ungå dette bruker vi argumentet out = "retrun" i `st()`.
- Kommenterte ut `print(total_missing)` i en chunk og viser hvor man kan få dette fint inn i teksten.
- print(height_stats), lagt inn litt kode for å skrive denne ut som flextable.
- Endret caption: "Deskriptiv statistikk for høyde." til `#| tbl-cap: Deskriptiv statistikk for høyde.` som er pandoc måten å gjøre dette på.
- `@ref(tbl-normInc)`, bruk heller `@tbl-normInc` som er pandoc kryssreferanse og vil virke i «alle» format.
- `label: men-low-income` må være `label: tbl-men-low-income` for at kryssreferanser og caption skal virke. Tilsvarende for `#| label: female-low-income`, men-norm-income etc. Label for table MÅ begynne med tbl-, for figure med fig-, for kodelisting lst-. Finnes en del til, men dette er de viktigste
- `\@ref(fig-income-married)` endret til `@fig-income-married`.
- Endret `#| tbl-cap: "Deskriptiv statistikk for delsettet med ingen inntekt."` til `#| tbl-cap: Deskriptiv statistikk for delsettet med ingen inntekt.` som tydeligvis er pandoc måten. Har en anførselstegn rundt teksten

forsvinner mellomrom mellom tallet (tabellnummeret) og teksten. Dette lærte seg selv nå ;-)

- Har lagt inn argumentet `show_coltype = FALSE` i `as_flextable()` for å slippe rekken i header som angir hvilken type kolonnen er.
- `\@ref(fig-norm-inc)` endret til `@rfig-norm-inc`
- Endret `#| fig-cap: "Histogram som viser fordeling av inntekt i populasjonen til hele datasettet" til#|` fig-cap: Histogram som viser fordeling av inntekt i populasjonen til hele datasettet.‘ for å få mellomrom mellom tabellnummer og tittel.
- `\@ref(fig-height-edufac)` endret til `@fig-height-edufac`
- Avsnittet *Splittet på kjønn.* Her lager dere flere data subsets slik som height_zero_inc_male, heights_zero_inc_female etc. Så vidt jeg kan se bruker dere ikke disse senere i oppgaven. Er dette tilfelle vil jeg anbefale å lage tabellene ad-hoc slik som jeg har forslått for height_zero_inc_male. Da unngår en å få så mange objekter definert i Environment. Dette sparer minne, men gjør det også lettere å holde oversikten. Hvis dere derimot brukte height_zero_inc_male mange steder i dokumentet, f.eks til figurer, statistisk analyse etc., ville det å lage subsettet height_zero_inc_male være en god ide.
- Har endret litt på width i tabellene. Dere hadde 16 mm for alle kolonnne. Dette ga bryting av tekst i første kolonne og std. dev. heading. Endret til å la første kolonne være 30 mm (s.a not married ikke ble brudt). Lot resten være default og da fikk også Std. Dev. tilstrekkelig bredde. Vi er litt sårbare når vi kjører med kompakte tabeller (line_spacing 0.3) siden linjebrudd vil gi overskriving av tekst.
- Har lagt til koden:

```
as_flextable(max_row = 30, show_coltype = FALSE) |>
line_spacing(space = 0.3, part = "body") |>
width(j = 2:8, width = 1.2, unit = "cm") |>
fit_to_width(max_width = 14, unit = "cm") |>
delete_part("footer")
```

for å få tabellene riktig formatert. Lar `line_spacing()` bare omfatte «kroppen» til tabellen s.a. vi har rom for bryting av lange titler på kolonner. Antyder bredde på kolonnene untatt den lengst til venstre som trenger større bredde, før jeg bruker `fit_to_width()` for å tilpasse til bredden på teksten. Merk også argumentene brukt i `as_flextable()` for å få med aller rekker (default for `as_flextable.data.frame()` er bare 10) og droppe kolonne type som ikke er relevant her. Et tips for å finne tekstvidden brukt i et Quarto pdf dokument (egentlig LaTeX dokument) er å legge inn tex kommandoen

`\the\textwidth`: 434.55125pt

Et LaTeX typografisk punkt har bredden: $\frac{1}{72{,}27}$ inch $\approx 0,013837$ inch $\approx 0,3515$ mm

Dette gir at tekstvidden er $434,55 * 0,3515 = 152,74$ mm som jeg har avrundet til 15 cm.

- Av tekniske årsaker har `line_spacing()` ingen virkning i pdf (dvs. LaTeX) dokumenter. Her må vi benytte hash-pipoen `#| ft.arraystretch: 1.2` for å få mer kompakte tabeller.

2.1 Andre kommentarer visualisering

- i `fig-count-income`: ggplot har skiftet fra `size`til `linewidth` som argument for å angi linjebredde. Enn så lenge virker begge, men det kan være lurt å skifte til ny standard. `size` brukes fremdeles for å angi punkt-størrelser etc..
- i `fig-BMI-income`: Her er det så mye data at det kan være lurt å skifte til et mindre punktsymbol. Har vist bruk av de to parametrene pch (print character) og cex (størrelse). Disse finnes dokumentert graphics::par.
- i fig-sex-income: Her vil jeg faktisk foreslå å droppe `facet_wrap()`. Det er enklere å sammenlikne tetthetene når de er i samme figur.
- i fig-height-inc-edu: Jeg tror jeg ville droppet dem som har NA for edu. De stjeler for mye plass. Har også foreslått å endre størrelse og alpha for punktene i geom_point. Legg merke til hvordan disse blir satt *utenfor* `aes()`. Vi øsnker bare å endre størrelse og alpha ikke at dette skal være egenskaper som illustrerer en variabel, da må de stå utenfor aes() i en mapping. Foreslår også å bruke heltrukket linje istedenfor 'dashed' siden vi bare har en linje.
- i fig-height-inc-sex: Viser bare et annet triks for å bli kvitt irriterende melding fra `geom_smooth()`.
- i fig-income-married: Her tror jeg dere har funnet mye av kjernen i lønnsforskjellene. I disse dataene ser det ut til at giftemål har svært forskjellig effekt for menn og kvinner (kanskje knyttet til tradisjonelle forestillinger om forsørgeransvar og å bli forsørget).
- i fig-height-norminc-sex (som er en variant av fig-height-inc-sex) har jeg foreslått å ta bort de svært høye inntektene (siden de kan ha svært stor innvrikning på regresjonslinjene). Jeg foreslår også å droppe facet_wrap() her. Dette fordi det er lettere å sammenlikne linjer i samme figur. Se bruken av argumentet group i geom_smooth() for å få en regresjonslinje for hver av kjønnene. Jeg lar også egenskapene colour og linestyle bli bestemt av kjønn slik at det er lett å se forskjell på linjene. For punktene har jeg forsøkt med geom_jitter() og å la også shape avhenge av sex. Å vise så mye data i en figur blir aldri perfekt og geom_point() kunne også vært brukt.