1 Register data og databeskyttelsesloven

I grupper af 4-5 personer, diskuter for projekt (a), (b) og/eller (c):

1. Hvilke (hvis nogen) etiske overvejelser ser I i forbindelse med projektet?
2. Diskuterer definitionen af populationen og hvordan den skal udtrækkes.
3. Hvilke typer registre kunne være relevante for jer at søge om adgang til i forhold til projektet?
4. Hvilke typer af resultater kunne I forestille jer I gerne vil hjemtage i forbindelse med projektet?
5. Hvad skal man særligt være opmærksom på under dette projekt?

(a) I ønsker at undersøge effekten af ekstra tandeftersyn til folkeskoleelever fra socialt udsatte hjem, hvor det interessante outcome er den fremtidige trivsel for den enkelte elev.

(b) I har indsamlet et spørgeskema foretaget blandt praktiserende læger og deres patienter, som I har fået lagt ind på et projekt, hvor det skal kobles med data fra Danmarks Statistik og SDS, så I kan undersøge om patienter aktivt vælger læge efter lægers tilbøjeligheder til at godkende sygefravær.

(c) I har et projekt der ser på effekten af kontanthjælpsreformen, og vil specifikt undersøge hvilke effekter den har haft på helbredet af udsatte kontanthjælpsmodtagere.

2 Hjemsendelse

Opgave 0

1. Du er logget ind på dit projekt på forskningsserveren og sidder sammen med en kollega og arbejder på jeres fælles projekt. Du skal til møde og bliver derfor nødt til at gå en time. Kollegaen er autoriseret bruger på samme projekt, så I beslutter at han overtager tastaturet og arbejder videre. Er det en overskridelse af reglerne?
2. Du er ny bruger og har ikke erfaring i at arbejde med mikro data. Derfor har du bedt din erfarne kollega om hjælp, så I har aftalt at han kommer forbi og sidder ved siden af dig og ser på skærmen. Er det I overensstemmelse med sikkerhedsreglerne?
   (a) Hvis kollegaen er autoriseret bruger på samme projekt?
   (b) Hvis kollegaen har en autorisations aftale, men kollegaen behøver ikke at være autoriseret bruger på det specifikke projekt
   (c) Hvis kollegaen alene skal hjælpe med programmering, behøver han ikke at være autoriseret bruger?
3. Du sidder alene ved skærmen og arbejder, men skal lige hente kaffe. Du forlader derfor computeren og falder i snak ved kaffemaskinen og kommer derfor først tilbage 15 minutter senere. Er det i overensstemmelse med sikkerhedsreglerne?
   (a) Det er OK, hvis du låser skærmen
   (b) Nej, du skal logge helt af serveren, hvis du forlader computeren i mere end et par minutter.

4. Du sidder på en cafe og har din bærebare computeren med. Er det tilladt at logge ind og sidde og arbejde på dit projekt?

   (a) Ja - hvis jeg alene sidder og sikrer, at ingen andre kan se med

   (b) Nej - det er ikke tilladt at logge ind på projekter under FSE fra andre steder end din arbejdsplads eller din hjemmearbejdsplads.

Opgave 1

1. Er det ok at liste mikrodata på skærmen for at validere data?

2. Du har fået leveret mikrodata om uddannelse og ønsker at analysere frafald på ungdomsuddannelserne. For at validere data forsøger du at identificere dig selv i data og se, hvordan din egen uddannelseshistorik er registreret i registrene. Er det OK?

   (a) Ja- så længe du alene forsøger at identificere dig selv

   (b) Ja - det er ok, så længe du ikke har adgang til CPR numre, men alene forsøger at identificere ud fra andre oplysninger

   (c) Nej, det er ikke tilladt at forsøge at identificere enkeltpersoner. Heller ikke sig selv

3. Det viser sig, at der er fejl i data, så du beslutter at kontakte forskningsservice for at vise din kontaktperson det problem, du har fundet. Hvilke af nedenstående handlinger er tilladte:

   (a) Du lister de problematiske observationer og sender dem hjem for at videresende til din kontaktperson

   (b) Du laver et screendump, som du e-mailer til din kontaktperson

   (c) Du sender en mail til din kontaktperson med PNR nummeret på personen med de proqlematiske oplysninger og beder din kontaktperson om selv at slå personens oplysninger op

   (d) Du lister de/den problematiske observation og gemmer den i en fil på forskningsserveren og sender derefter en mail til din kontaktperson og fortæller, hvor filen ligger

Opgave 2

1. Må dette datasæt hjemsendes? Forklar hvorfor/hvorfor ikke.

pnr	Aar	Alder	Indkomst
123456789123	2010	30	300,000
123456789123	2011	31	310,000
123456789123	2012	32	305,000
123456789123	2013	33	250,000
.....	....	..	.....

2. Må dette datasæt hjemsendes? Forklar hvorfor/hvorfor ikke.

3. Er det tilladt at sende det sidste uddrag af data, hvis den ligger i en logfil eller i et program?

pnr	Aar	Alder	Indkomst
.	2010	30	300,000
.	2011	31	310,000
.	2012	32	305,000
.	2013	33	250,000
.	....	..	.....

4. Din institution har samlet data fra et spørgeskema og en kopi af de data bliver lagt på forskermaskinen i pseudononymiseret form. En kollega vil gerne lave analyser på datasættet udenfor forskermaskinen fordi hans ynglingsstatistiskprogram (muligvis Excel) ikke ligger på forskermaskinen. Da data ikke er ejet af Danmarks Statistik, overvejer han at hjemsende datasættet. Din kollega er lidt i tvivl og spørger dig til råds. Hvad vil du svare?

Opgave 3

For hver af nedenstående tabeller, diskturer hvorvidt tabellen må hjemsendes?

Table 1: Gennemsnitlig indkomst opdelt på stillingskategori i sygehus X

	Gns.	N
Ansatte	487,870.4	9,998
Leder	874,190.2	2
Total	487,947.7	10,000

Table 2: Gennemsnitlig indkomst opdelt på stillingskategori på sygehuse i region Z for sygeplejersker

	Gns.	N
Ansatte	443,594.1	923
Leder	701,458.4	77
Total	463,449.7	1,000

Table 3: Gennemsnitlig indkomst opdelt på stillingskategori i sygehus X, afdeling Y

	Gns.	N
Administrativ personel	429,079.7	6
Sygeplejersker	294,947.0	11
Sundhedsøkonomer	801,237.3	4
Kirurg	987,667.7	4
Øvrige ansatte	211,812.3	5
Total	467,785.9	30

Table 4: Gennemsnitlig indkomst opdelt på stillingskategori

	Gns.	N
Administrativ personel	429,079.7	6
Sygeplejersker	294,947.0	11
Sundhedsøkonomer	801,237.3	4
Kirurg	987,667.7	4
Øvrige ansatte	211,812.3	5
Total	467,785.9	30

Table 5: Min/max indkomst opdelt på stillingskategori i sygehuse i region Z

	min	max
Ansatte	208,329	727,475.0
Leder	520,906	1,227,474.4
Total	208,329	1,227,474.4
N		10,000

Table 6: Min/max indkomst opdelt på stillingskategori i sygehuse i region Z

	min	max
Ansatte	<250,000	>725,000
Leder	<525,000	>1,225,000
Total	<250,000	>1,225,000
N		10,000

Table 7: Gennemsnitlig indkomst opdelt på stillingskategori indenfor transportbranchen

	Gns.	N
Ansatte	478,725.0	5,500
Leder	498,953.2	4,500
Total	487,827.7	10,000

Opgave 4

1. Kan denne stump kode hjemsendes? Forklar hvorfor?

```
generate virksomhed = 0
replace  virksomhed = 1 if lbnr == "1239784"
replace  virksomhed = 2 if lbnr == "8493763"

/* Koder

1 = Mærsk
2 = Novo-Nordisk
0 = Øvrige virksomheder

*/
```

2. Hvad hvis vi sletter kommentaren omkring Koder?
3. Hvad hvis vi sletter koden, og kun beholder kommentaren?

Opgave 5

Man skal kende sine data før man hjemsender deskriptive stats. Diskuter under hvilke omstændigheder disse figurer er problematiske og under hvilke de er i orden at hjemsende. I det tilfælde, hvor figuren ikke kan hjemsendes, diskuter hvilke ændringer skal/kan laves for at figuren opfylder sikkerhedskravene.

Figure 1

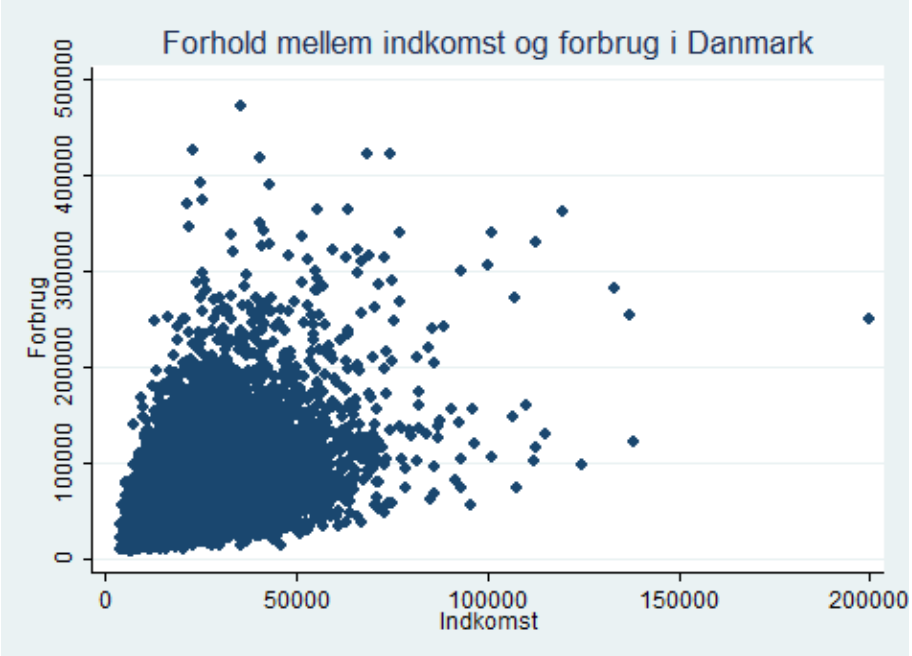


Figure 2

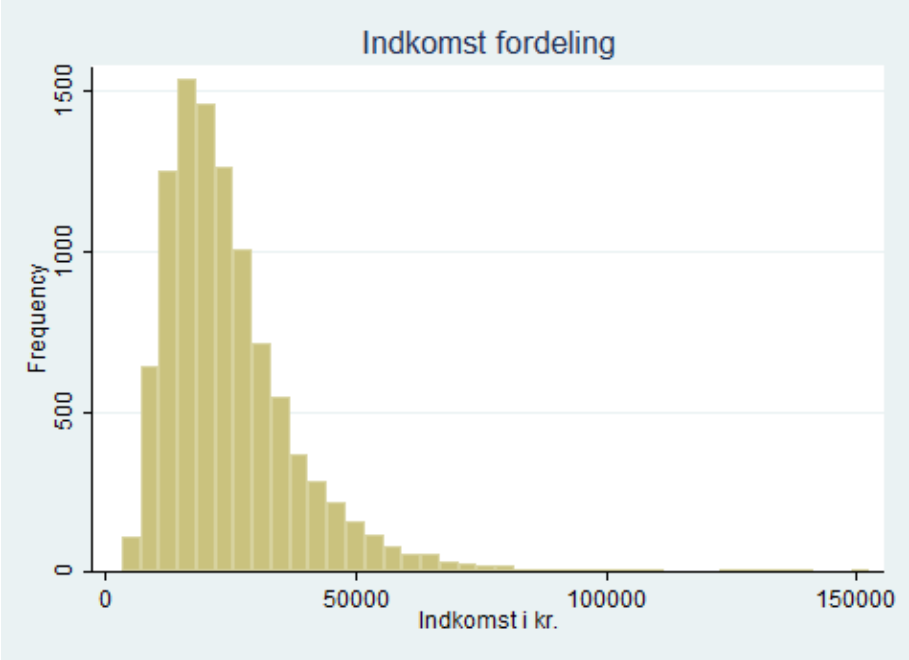
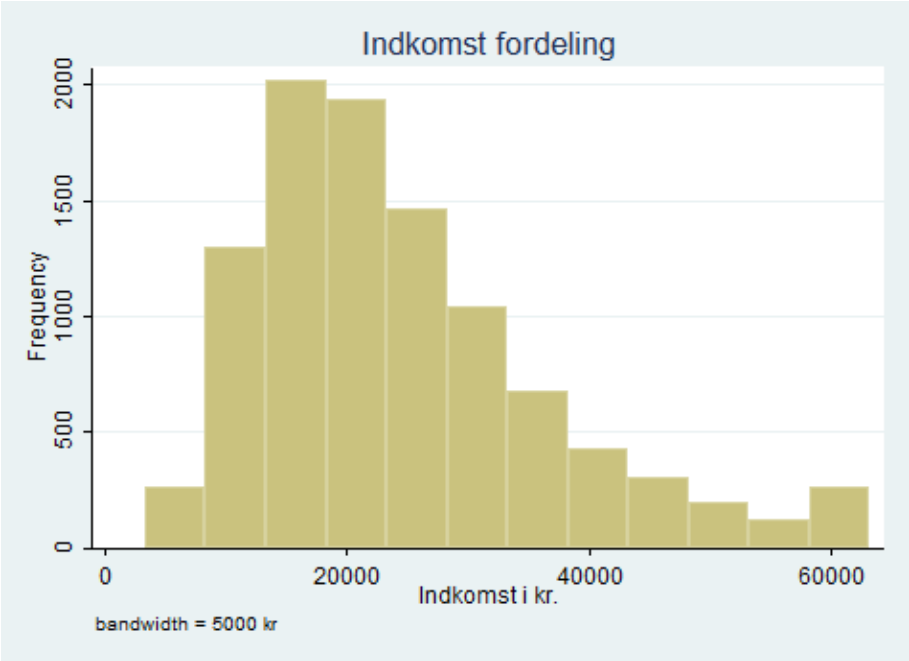


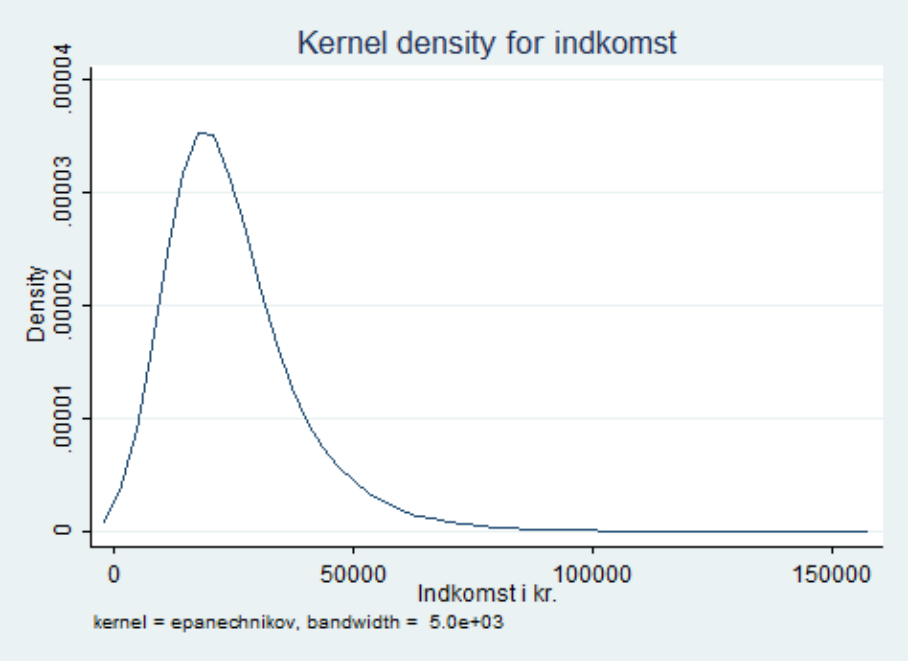
Figure 3



Opgave 6

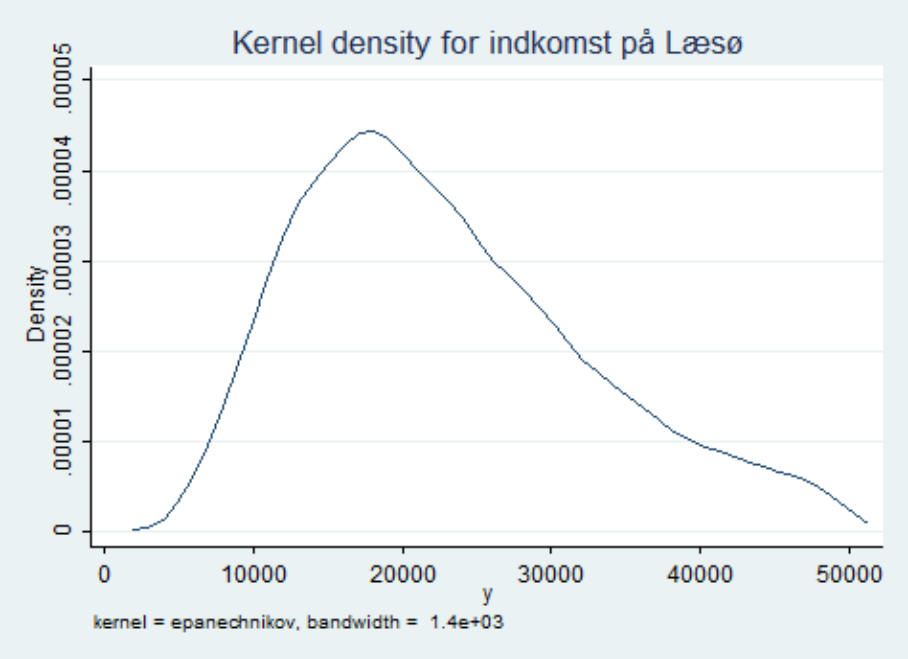
1. Diskuter under hvilke omstændigheder denne figur kunne være problematisk og hvornår den er ok.

Figure 4



2. Hvad med denne figur?

Figure 5



Opgave 7

En forsker vil implementere en algoritme for at kunne slette visse observationer, som opfylder en betingelse. Betingelse er meget specifik og kompliceret. Desuden kræver algoritmen en rekursiv tilgang. Derfor beslutter forskeren sig for at køre en løkke indtil hun ikke finder nogen observationer, der opfylder betingelsen. Indenfor løkken tester hun betingelsen ved at liste observationerne. Hun lister observationerne for at teste, om algoritmen tester betingelsen korrekt. Hun fortsætter indtil listingen er tom og skriver algoritmens output i en logfil som hun sender hjem.

1. Hvad er problemet med sådan en tilgang?
2. Diskuter hvornår logfilen bliver problematisk. Foreslå ændringer i algoritmen, hvor datasikkerhedsreglerne er overholdt.

Opgave 8

Det kan være farligt at liste observationer på individniveau og sætte outputet i en logfil.

1. Diskuter i hvilke situationer I vil liste mikrodata (enten på skærmen eller i en logfil).
2. Prøv at finde alternative løsninger, der er mere sikre med hensyn til datasikkerhedsreglerne.

Opgave 9

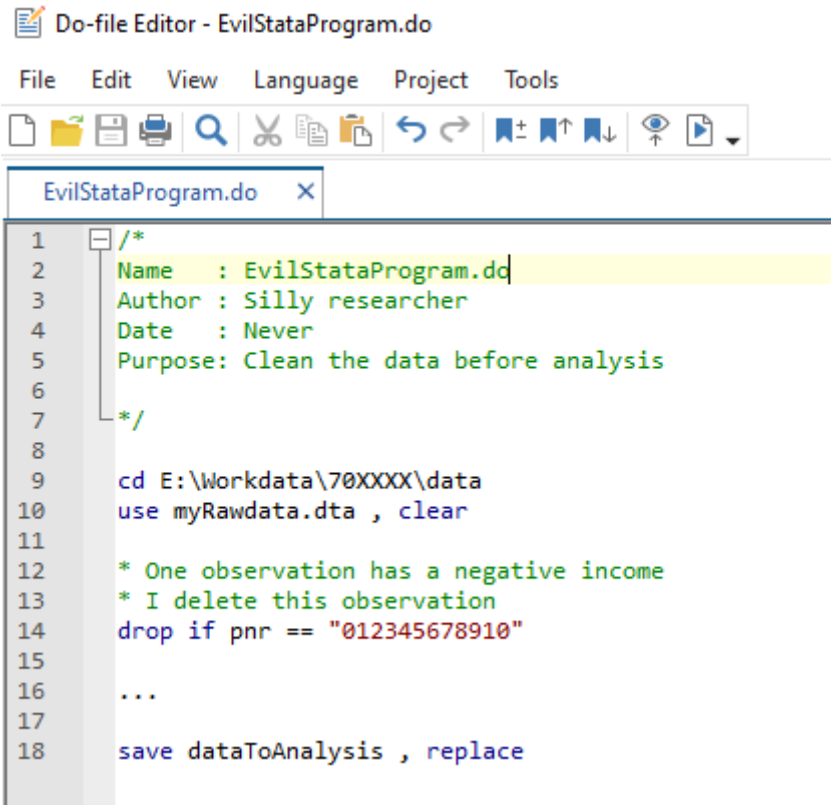
Diskuter hvorfor denne tabel er potentielt problematisk.

Table 8: Gennemsnitlig indkomst på Fanø opdelt på branche og alder

Branche/Alder	...	25-30 årige	N	...
...	...	...	...	...
Fisker	...	475,000	5	...
...	...	...	...	...
Total	...	...	...	...

Opgave 10

Du ønsker at hjemsende nedenstående Stataprogram. Er det OK? Hvad er problemet med sådan et program?



Opgave 11

Du finder personer med negativ indkomst og vil gerne vise det til en kollega. Må du hjemsende denne tabel?

Obs	ALDER	KON	CIVST	BRUTTO
1	18	1	3	355000
2	20	1	2	403456
3	52	2	2	102754
4	33	2	3	-7000
5	65	1	2	155000

Opgave 12

Jeg er i tvivl om sammenhængen mellem familienummer og oplysninger om antal hjemmeboende børn i familien i de data, jeg har fået leveret, så jeg bruger proc print i sas til at liste 20 observationer af familienummer og PNR. Både familienummer og PNR er anonymiseret så

1. Jeg hjemsender listen
2. Jeg tager et screendump
3. Jeg kalder på min kollega, så han kan se listen på min skærm
4. Andet?

Opgave 13

Jeg har indsamlet en survey, som jeg sidder og arbejder med lokalt. Jeg har søgt datatilsynet, så tilladelserne er i orden. Jeg har også sendt surveyen til Danmarks Statistik for at have muligheden for at koble surveyen med registerdata.

1. Jeg må gerne downloade mikrodata fra den survey, jeg selv har lagt op, og som jeg har tilladelse til at arbejde med lokalt.

2. Det er kun mikrodata, som Danmarks Statistik har leveret til mit projekt, jeg ikke må hjemsende leveret af Danmarks Statistik eller det er mine egne data.

3. Det er under ingen omstændigheder tilladt at hjemsende mikrodata, uanset om data er leveret af Danmarks Statistik eller det er mine egne data.

Opgave 14

1. Du har siddet og arbejdet i Stata hele dagen, og alle mine analyser ligger nu samlet i Stata's log fil. Hvordan får jeg bedst hjemsendt mine analyser?

2. Det er ikke tilladt at hjemsende tabeller, der indeholder oplysninger om medianer, maximum og minimum, da disse reelt er observationer for enkeltpersoner. Er det korrekt?

3. Du har siddet og arbejder på dit projekt i flere dage og har gemt 20-30 filer med output. Du er ikke helt sikker på indholdet i alle filer, og du har travlt. Hvad gør du?

   (a) Bruger Danmarks Statistiks nye scanningssystem til at filtrere data. De filer, der kommer advarsler på, tjekker du manuelt

   (b) Du tager den tid det tager at gennemse alle filerne, inden du downloader

   (c) Andet?

4. Det er sidst på dagen og du opdager, at du ved en fejl har hjemsendt mikrodata. Hvad gør du?

   (a) Du gør ingenting og satser på det ikke bliver opdaget

   (b) Du kontakter den autorisationsansvarlige på din institution og afventer, hvad han beslutter, at der skal gøres

   (c) Du kontakter straks Danmarks Statistik.

Opgave 15

Du har lavet nedenstående tabel. Hvad skal du være opmærksom på?

Table 9: Antal fødsler opgjort efter moderens alder og oprindelse i Gladsaxe kommune

Alder	Dansker	Indvandrer
18 år	0	1
19 år	1	1
20 år	1	0
21 år	8	2
22 år	6	1
23 år	6	2
24 år	12	3
25 år	7	3

1. Der er færre end 3 observationer i nogle af cellerne, så det er ikke tilladt at hjemsende tabellen.

2. Det nye hjemsendelsesværktøj har scannet tabellen og giver ikke nogen advarsel, så jeg kan trygt hjemsende tabellen.

Opgave 16

Jeg har en fil med en figur, hvor filttypen ikke er tilladt til hjemsendelse. Jeg overvejer at ændre det til pdf format, så den kan hjemsendes. Er det tilladt? Forklar hvorfor.

Opgave 17

Jeg er i tvivl om hvordan jeg skal reshape mine data. Jeg laver en post på Stack Overflow om det og laver en screenshot af data for at illustrere min problemstilling. Er det tilladt? Forklar hvorfor.