1 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?
   Svar: Det er en overskridelse af reglerne. Adgangen til fjernforbindelsen er personlig, jf. punkt 4 af brugeraftalen.

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?

   Svar: c. Man skal være autoriseret på samme projekt for at kunne se indholdet på et projekt. Man skal dog afvikle kørlser fra sin egen adgang (jf. spørgsmål 1)

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.

   Svar b. Det står i brugeraftalen under punkt 3. Lukning af fjernforbindelsen betyder ikke, at man skal lukke sin session, blot afkoble fjernforbindelsen. At låse sin skærm er til gengæld ikke tilstrækkelig. På Windows serveren kan man enten vælge "disconnect", hvilket afbryder forbindelsen, eller man kan også vælge "sign out", som vil afslutte sesssionen.

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.

   Svar: b. Ifølge punkt 2 i brugeraftalen skal arbejdet på forskermaskinen ske ved opkobling fra egen arbejdsplads eller hjemmefra. Hvis man skal arbejde på forskermaskinen fra udlandet, så skal man opkoble fra den autoriserede institution, for eksempel igennem en VPN-forbindelse. Se forskningsservices vejledning vedr. opkobling til forskermaskinen fra udlandet.

Opgave 1

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

   Svar: Det står ikke nogen steder, men det må man gerne, ellers bliver registerforskning betydeligt vanskeligere.

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

   Svar: c jf. punkt 7 i brugeraftalen om forsøg på identifikation af enkeltpersoner eller enkeltvirksomheder.

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

   Svar: Man må aldrig hjemsende data på individ- eller virkskomhedsniveau, jf. brugeraftalen pkt. 7, selvom det er til ansatte i Forskningsservice

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
.....	....	..	.....

   Svar: Nej. Tabellen indeholder mikrodata med både en nøglevariabel og personoplysninger. Selvom pnr er en pseudonymiseret udgave af cpr.nr., betragtes det som mikrodata.

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

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

   Svar: Nej. Indeholder stadig mikrodata, selvom nøglevariablen (pnr) er fjernet.

3. Er det tilladt at sende det sidste uddrag af data, hvis den ligger i en logfil eller i et program? Svar: Nej. Det gør ikke nogen forskel. Pas derfor på med at liste mikrodata i logfiler eller programkoder.

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?

   Svar: Selvom data ikke kommer fra DST og man ejer data, når det bliver lagt på serveren, bliver det betragtet som mikrodata. Jeg vil svare, at det ikke tilladt, og man skal derfor undlade det. Der findes i øvrigt et Excel-lignende program (LibreOffice) på Danmarks Statistiks server (EDIT server på FSE-windows).

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

Svar 1: Må ikke hjemsendes. Gennemsnittet for "Ledere" er baseret på mindre en fem observationer. Fjern linjerne "Leder" og "Total" (ellers kan man regne sig frem til den oprindelige tabel).

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

Svar 2: Kan hjemsendes.

Svar 3: Tabel har celler med mindre end fem observationer. Kan ikke hjemsendes. Man skal samle de to kategorier Sundhedsøkonomer og Kirurg.

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

Svar 4: Tabellen har celler med mindre end fem observationer. Kan ikke hjemsendes. Man skal samle de to kategorier Sundhedsøkonomer og Kirurg. Det skal man muligvis også med de øvrige ansatte, da vi er på kanten af kriteriet for fem observationer. Overskriften på tabellen er desuden meget detaljeret. Det er et eksempel på, at man skal overveje at gå over grænsen på fem observationer. Tabellen er generelt for detaljeret og er på kanten at være personfølsom. Jeg ville undlade at sende denne tabel selv i en mere aggregeret udgave.

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

Svar 5: Afhængig af konteksten kan min. og maks. værdier afsløre personfølsomme oplysninger. Det er muligvis problemet i det her tilfælde, især vedr. "Leder", hvor man nemt kan identificere, hvem lederen med den højeste indkomst er.

Svar 6: Man har prøvet at disketionere tabellen. Kriteriet må være det beløb, der indeholder mindst tre personer, men man kan godt overveje at gå over denne grænse afhængig af konteksten.

Svar 7: Intet problem her; kan hjemsendes.

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

*/
```

Svar: Nej, der er mikrodata i programkoden. LBNR er en nøglevariablen hos DST og betragtes derfor som mikrodata. Endnu værre: det er lykkedes brugeren at identificere enkelte virksomheder bag lbnr. En sådan identifikation må ikke foretages og sanktioneres med en udelukkelese fra forskerordningen for den enkelte forsker og en lukning af institutionens adgang til deres projekter i minimum en måned. Svaret for punkt 2 og 3 er derfor nej i begge tilfælde.

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

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.

Svar Figur 1: Må ikke hjemsendes. Der er outliers (ekstreme værdier), der kan afsløre enkeltpersoner. Selvom figurer med scatter plot af observationer med individdata kan hjemsendes, så længe de ikke afsløre enkelte personer, vil jeg finde en anden måde at formidle disse data. Man kan overveje at bruge rankordningen i fordelingen af forbrug og indkomst i stedet for beløb. Generelt falder figur 1-5 under kategorien grå zone. Hvis I er tvivl, så undlad at hjemsende materialet.

Svar Figur 2: Afhængig af konteksten kan de ekstreme værdier være problematiske. Denne fil kan for eksempel afsløre maks. værdien for en enkelt person (150.000 i dette tilfælde). Hvis man har lidt mere baggrund om populationen bag denne figur, kan den være problematisk at hjemsende. Jeg vil derfor undlade at hjemmesende denne figur og prøve at disketionnere ydeligere som i figur 3 ved at brede "bins" ud og trunkere fordelingen på toppen.

Svar Figur 3: Se kommentaren vedr. figur 2

Figure 1

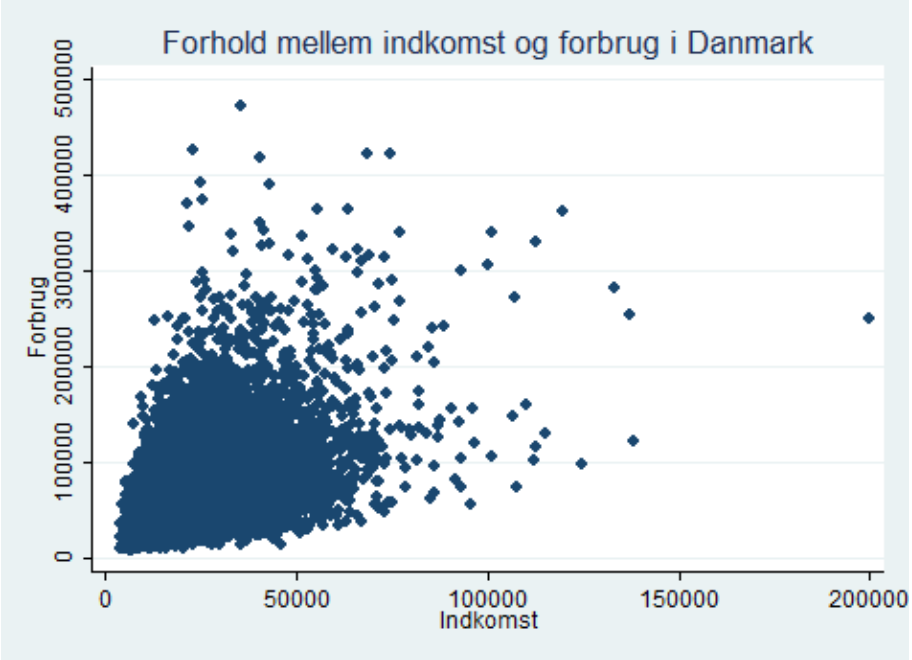


Figure 2

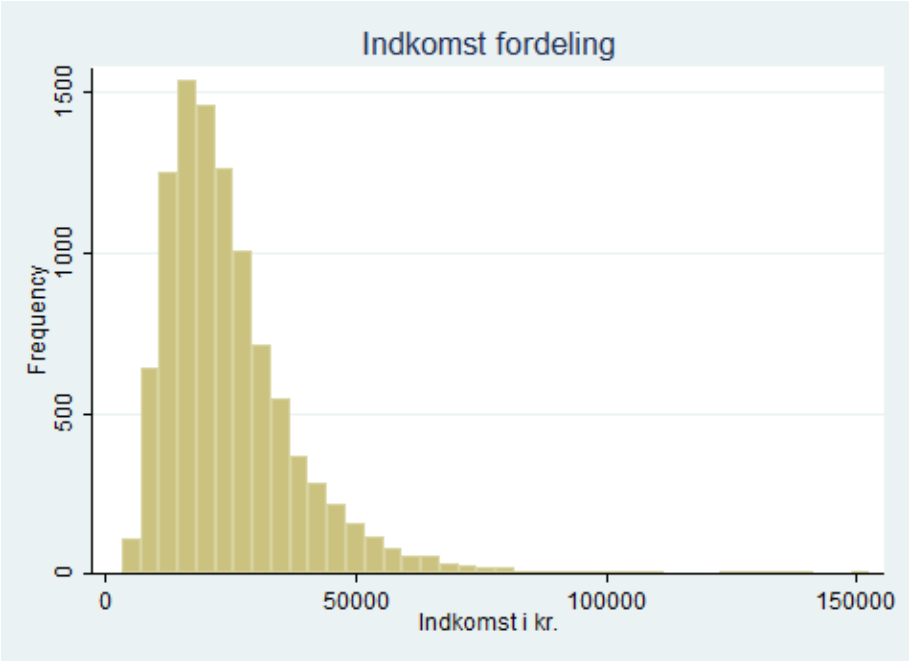
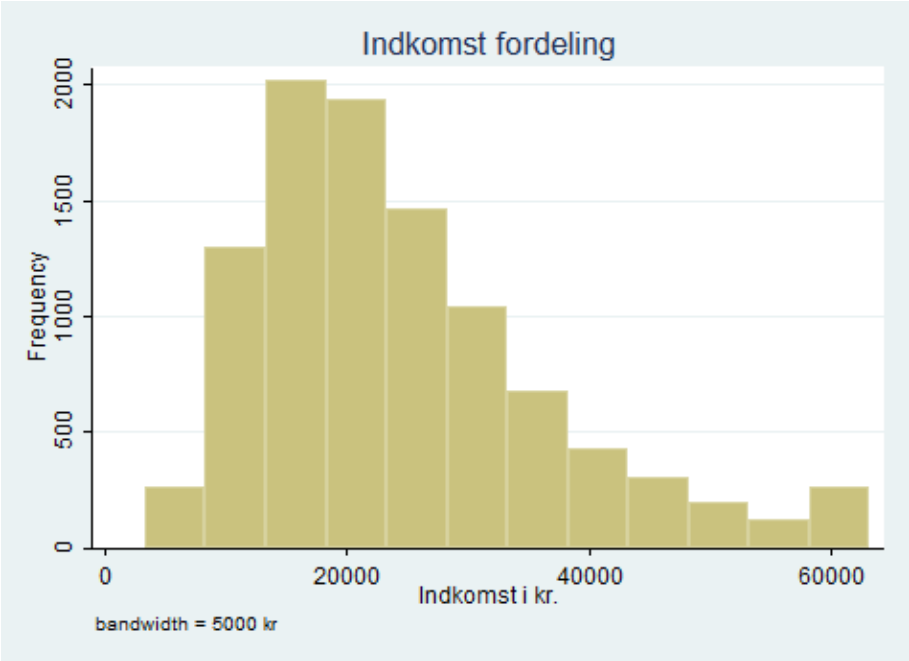


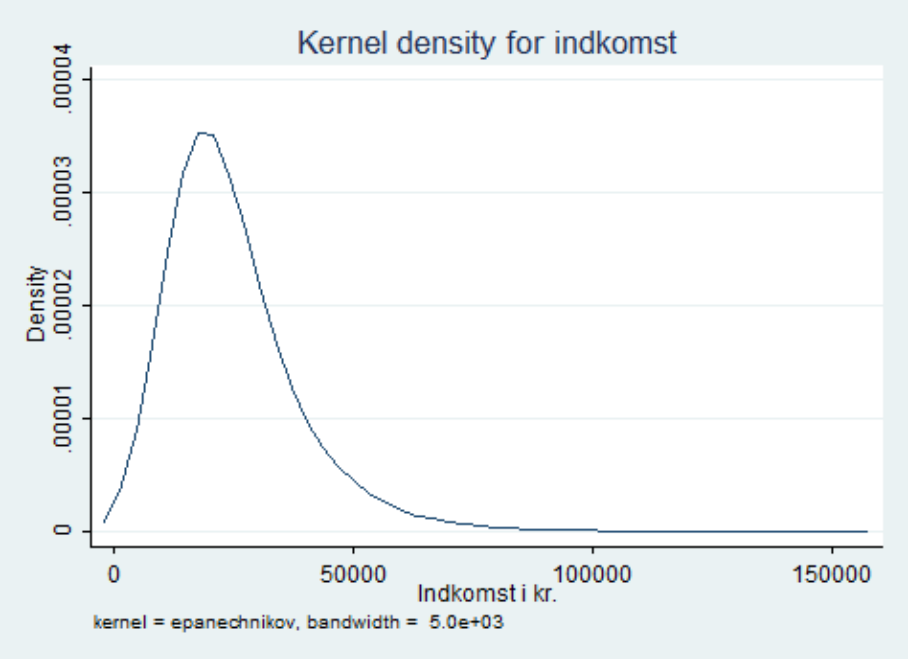
FIGURE 3


Indkomst fordeling
Frequency
0
500
1000
1500
2000
0
20000
40000
60000
Indkomst i kr.
bandwidth = 5000 kr

Opgave 6

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

Figure 4

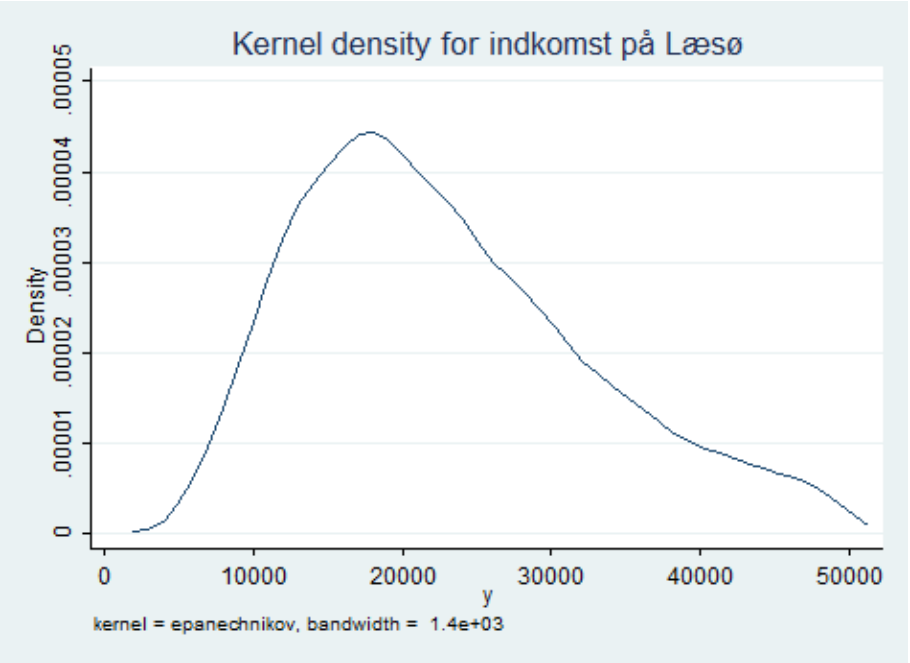


Svar Figur 4: Det er figur 2 i en kernel density udgave. Det er svært at identificere enkelte personer ud fra denne graf, og det er lidt sværere at afgøre, hvad maks. værdien er, men vi har et interval, der kan give oplysninger om den rigeste person. Prøv generelt at bruge en skala, hvor der er tilstrækkelig data for hvert punkt, og hvor hælen ikke er for tynd.

2. Hvad med denne figur?

Svar Figur 5: Vi får lidt mere baggrund om populationen (Læsø), som er lille. Hvis populationen rent faktisk er alle på Læsø, kan man gennemskue, hvem den rigeste person er, og man har derfor fået ny viden. Jeg vil brug en anden type figur for at formidle indkomstfordelingen på Læsø. En tabel med indkomstsintervaller for eksempel.

FIGURE 5


Kernel density for indkomst på Læsø
Density
0
.00001
.00002
.00003
.00004
.00005
0
10000
20000
30000
40000
50000
y
kernel = epanechnikov, bandwidth = 1.4e+03

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.

Svar: Det er farligt! Det kan være problematisk at liste personoplysninger i en logfil, da den vil indholde mikrodata, og man kan ved en fejl risikere at hjemsende logfilen. Det vil være en bedre tilgang at bruge assert-funktionen. Funktionen tester om et kriterie er opfyldt eller ej og stopper programmet, hvis det ikke tilfældet. Assert findes både i R og Stata og stopper programmet med en fejlmelding, hvis kriteriet ikke er opfyldt.

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).

   Svar: Nogle gange er man nødt til at undersøge årsagen til ekstreme eller absurde værdier ved at se, hvordan de hænger sammen med de øvrige data (i forhold til de andre rækker eller på tværs af andre variabler for den samme række).

2. Prøv at finde alternative løsninger, der er mere sikre med hensyn til datasikkerhedsreglerne.

   Svar: Assert er en løsning. Begræns jer til at liste observationer på skærmen og ikke i en logfil. Trimming, Winsorizing eller winsorization https://da.abcdef.wiki/wiki/Winsorizing af data

Opgave 9

Diskuter hvorfor denne tabel er potentielt problematisk.

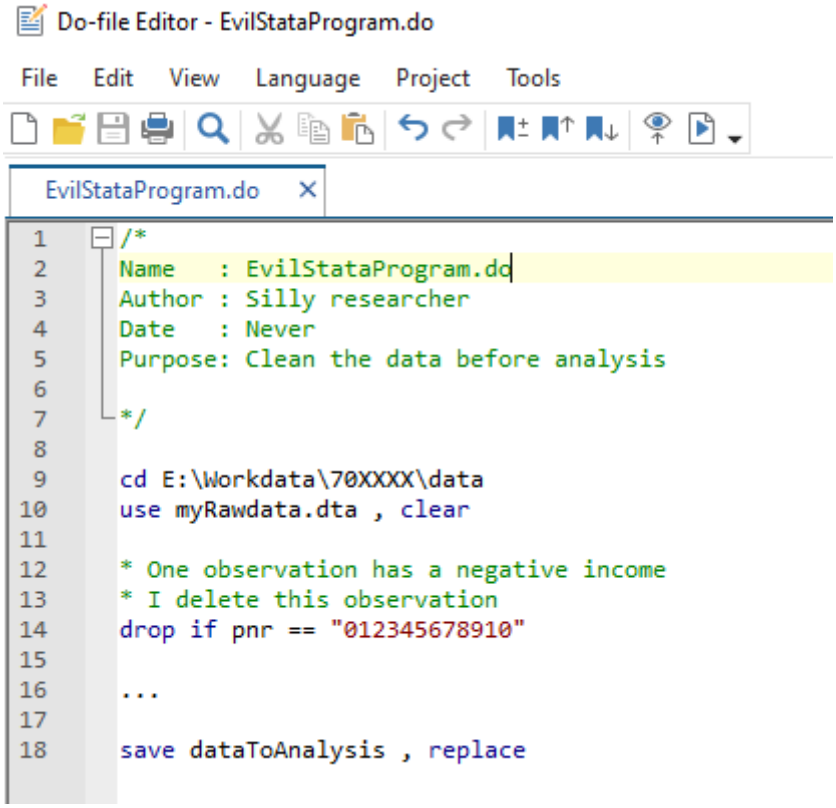
Table 8: Gennemsnitlig indkomst på Fanø opdelt på branche og alder

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

Svar: Man kan overveje at gå over grænsen på fem observationer. Der er fare for identifikation af personer i dette tilfælde.

Opgave 10

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



Svar: Nej, den indedolder mikrodata (pnr = ...)

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

Svar: Det er mirkoro data. Der er tale om individdata, da indkomst er listet med enkelte værdier, som indikerer, at data er på individniveau.

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?

Svar : d. Jeg kalder på en kollega, som er AUTORISERET PÅ PROJEKTET. Jeg kan også henvise til filen til kontaktpersonen i Forskerservice for at undersøge, om der er tale om en fejl i dataleverancen.

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.

Svar: 3.

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?

   Svar: Opret en fil for hver tabel. Jeg vil eventuelt samle flere resultater i en tabel. Undgå at hjemsende en lang logfil, som I bagefter bearbejder udenfor serveren. Lav generelt et script for hver tabel I skal hjemsende.

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?

   Svar:Det er faktisk ikke helt korrekt. De bliver problematiske og skal ikke hjemsendes, hvis de afslører enkeltepersoner. Problemet er ikke så meget, at der kunne være en person bag tallet, men mere, at den kan afsløre, hvem personen er.

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?

   Svar: b og c. Jeg reducerer antallet af filer, som jeg hjemsender for at minimere risikoen for at begå en fejl.

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.

Svar: a, hvis man godt kan lide at gamble. Personligt vælger jeg løsning c og kontakter den autorisationsansvarlige på min institution.

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.

Svar: Tabellen må ikke hjemsendes som den er, fordi den er for detaljeret. Man bliver nødt til at lave bredre aldersgrupper. Scanningsprogrammet vil ikke give en advarsel, da det ikke kan fange sådanne tilfælde. Husk at scanningsprogrammet leder efter mønstre, der ligner nøglevariabler. Programmet kan ikke fortælle, om en tabel er for detaljeret eller ej.

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.

Svar: Man må ikke ændre filtypen ved at ændre på filextension (fx Ændre .gph til .pdf i windows exlporer er ikke tilladt og vil gøre filen ulæselig). Man må dog gerne gemme en fil i et andet filformat ved at eksportere figuren til for eksempel pdf-format.

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.

Svar: Må må tage ikke screenshots af mikrodata på skærmen, og screenshots af serveren i det hele taget.