Jeg anvender data fra franske mænd under 40. Først importeres data:

```
proc import datafile="C:\Users\asbjp\Documents\KU\Videregående
statistik\hjemmeopgave\ESS8e01" out=ud replace DBMS=sav;
run;
```

Derefter dannes et nyt datasæt *minedata* hvor alle undtagen franske mænd under 40 er slettet fra datasættes fra:

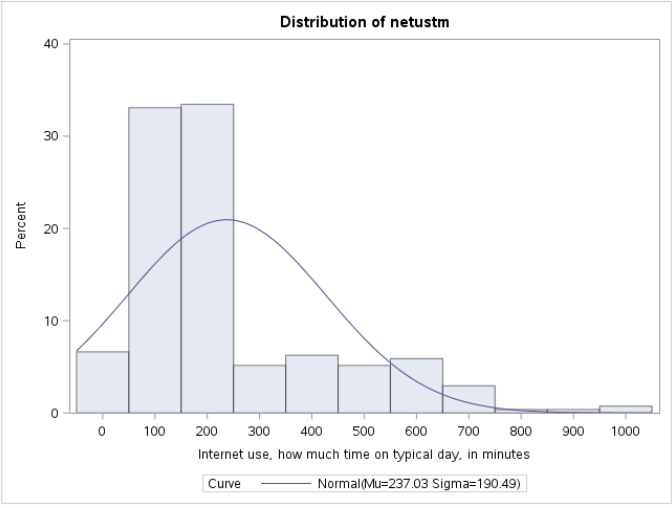
```
*Sortér data som beskrevet i opgaveformuleringen;
data minedata;
set ud;
if agea < 40 then age1=1;
if 40<= agea< 70 then age1=2;
if agea =>70 then delete;
*som en følge af de næste tre linier fjernes alle respondenter,
der aldrig bruger internet;
if netustm>5000 then delete;
log_netustm=log(netustm);
if log_netustm=. then delete;
*fjerner enkelte med manglende angivelse af køn;
if gndr=. then delete;
*de næste tre liner skal rettes til jf fordelingstabellen;
if cntry ne 'FR' then delete;
if age1 ne 1 then delete;
if gndr ne 1 then delete;
run;
```

a) Undersøg grafisk om variablene netustm og/eller log_netustm kan antages at være normalfordelte.

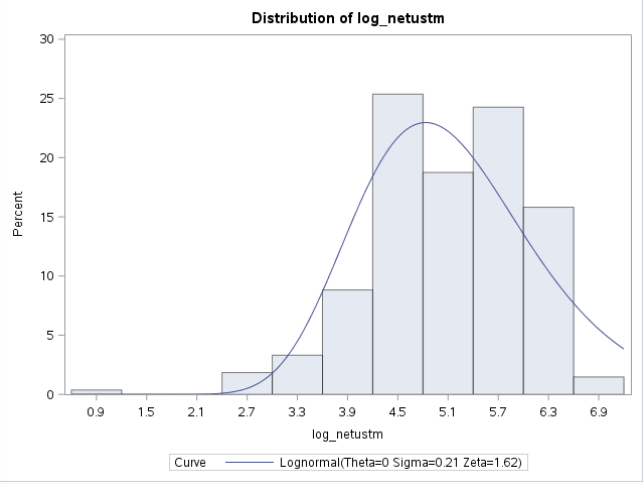
Den grafiske undersøgelse gennemføres ved at opstille histogrammer for variablene *netustm* (internetbrug) og logaritmen til denne. De holdes op mod en graf for henholdsvis normalfordelingen og logaritmen til normalfordelingen. Det gøres med koden:

```
proc univariate data=minedata;
*Definér de variable vi ønsker beskrevet;
var netustm;
var log_netustm;
*Definér histogram med relevant sammenligning;
histogram netustm/normal;
histogram log_netustm/lognormal;
run;
```

Histogrammet for *netustm* med hvor SAS har tilnærmet en normalfordeling med middelværdi 237,0294 og en standardafvigelse på 190,49163, samt en skewness på 1,476, og dermed er meget højreskæv og kurtosis på 1,96. Her ses det at denne fordeling afviger fra normalfordelingen idet en normalfordeling har kurtsis 3 og og skewness på 0.



Histogrammet for *log_netustm* med hvor SAS har tilnærmet en normalfordeling med middelværdi 5,144 og en standardafvigelse på 0,8742 samt en skewness på -0,767 og dermed er meget venstreskæv og kurtosis på 2,095. Her ses det at denne fordeling afviger fra normalfordelingen idet en normalfordeling har kurtsis 3 og og skewness på 0.



Dermed vil det ikke være præcist at antage at *neustm* og *log_neustm* er normalfordelte.

b) Test om middelværdien af netustm er lig med 197.9 i dit datasæt "minedata". Tallet 197.9 er gennemsnittet i det samlede ESS datamateriale. Benyt eventuelt den logaritmisk transformerede.
Det vil test om middelværdien af *netustm* er lig OECD gennemsnittet på 197,9. Dette gøres ved at udføre et t-test på mine data. Dette gøres via følgende stump kode:

```
proc ttest data=minedata h0=197.9;
var netustm;
run;
```

som giver følgende output

N	Mean	Std Dev	Std Err	Minimum	Maximum
272	237.0	190.5	11.5503	2.0000	1020.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
237.0	214.3	259.8	190.5	175.7	208.0

DF	t Value	Pr > |t|
271	3.39	0.0008

Det ses at den empirisk middelværdi bliver 237 med en standardafviglese på 190,5 og maks på 1020 og min på 2. Det ses at t-teststørrelsen bliver 3,39 og p-værdien på 0,8%. Testes der med et signifikansniveau på 5% fås den kritisk værdi til 1,96. Det ses heraf klart at t-teststørrelsen er større end den kritisk værdi og p-værdien er under signifikansniveauet hvormed vi kan afvise nulhypotesen og der er dermed må franske mænd under 40 internet forbrug være signfikant forskelligt fra 197,9.

***c)** Udfør en regressionsanalyse, hvor variablen trstep er responsvariabel og trstprl er den forklarende variabel. Kommenter resultatet.*
Trstep regresseres på *trstprl*. Dette gøres via følgende kode i SAS:

```
proc reg data=minedata;
model trstprl=trstep;
```

```
run;
```

Det giver os outputtet:

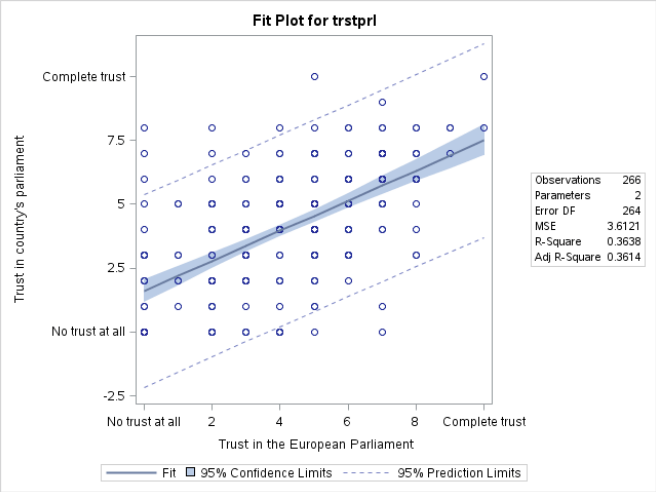
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1.60002	0.22930	6.98	<.0001
trstep	Trust in the European Parliament	1	0.58994	0.04802	12.29	<.0001

Vi estimerer en lineære regressionsmodel af følgende type

$$trstprl = \beta_0 + \beta_1 trstep + u$$

Dette giver et estimat $\hat{\beta}_1 = 0{,}59$. Dette skal fortolkes som at alt andet lige effekten af, at tilliden til EU-parlamentet øges med 1 point er, at tilliden til det nationale parlament øges med 0,59 point. Vi estimerer ligeledes $\hat{\beta}_0 = 1{,}6$ som siger at når tilliden til det EU-parlament er 0 er tilliden til det til det nationale parlament positivt, som siger at tilliden til den nationale parlament er højere end EU-parlamentet. Begge estimater er signifikant på 1% niveau.



Det fremgår af plottet af R^2=0,36, som betyder at variationen i tilliden til EU-parlamentet kan forklarer 36% af variation til det nationale parlament.

d) *Hvad gør følgende program?*

```
data zero_trust;
set minedata;
no_trust_ep=1;
if trstep>0 then no_trust_ep=0;
no_trust_prl=0;
if trstprl=0 then no_trust_prl=1;
run;
```

Linje 1 genererer et nyt datasæt, der hedder zero_trust. I linje 2 sættes datasættet til at tage udgangspunkt i datasættet *minedata*, dermed data fra ESS-undersøgelsen for franske mænd under 40. Linje 3 generer en ny variabel no_trust_ep, der som udgangspunkt er lig 1. Linje 4 fortæller, at variablen *no_trust_ep* skal antage værdien 0, hvis tilliden til Europaparlamentet, variablen *trstep* er større end 0. Linje 5 danner en variabel *no_trust_prl*, der som udgangspunkt er lig 0. Linje 6 fortæller, at variablen *no_trust_prl* skal antage værdien 1, hvis *trstprl* er lig 0.

Programmet danner altså med udgangspunkt i ESS-undersøgelsen to binære variable, nemlig en binær for ingen tillid til EU-parlamentet, der antager værdien 1, hvis man ikke har tillid til EU-parlamentet, og tilsvarende for det nationale parlament. Måden, de to variable dannes på, er lidt forskellig, da *no_trust_ep* som udgangspunkt sættes til 1, mens *no_trust_prl* som udgangspunkt sættes til 0.

e) *Undersøg i datasættet zero_trust, om der er samme andel med værdien 1 af de to variable no_trust_ep og no_trust_prl. Kan hypotesen testes?*

Første skridt er at beregne procentdelen som ikke har tillid hhv. EU- og det nationale parlament og dernæst teste om disse andele er statistisk forskellige fra hinanden.

The FREQ Procedure

no_trust_ep	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	231	84.93	231	84.93
1	41	15.07	272	100.00

no_trust_prl	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	235	86.40	235	86.40
1	37	13.60	272	100.00

Det ses her at 15% ikke har tillid til EU-parlament og 13,6 ikke har tillid til det national parlament. Vi tester om disse værdier er statistisk signifikant via en chi i anden test med følgende kode:

```
proc freq data=zero_trust;
tables no_trust_prl/testp=(0.864,0.136);
tables no_trust_ep/testp=(0.85,0.15);
run;
```

Der ligges ud med at teste, om den sande værdi for andelen af franske mænd under 40, der ikke har tillid til det nationale parlament, er lig 13,6%

no_trust_prl	Frequency	Percent	Test Percent	Cumulative Frequency	Cumulative Percent
0	235	86.40	86.40	235	86.40
1	37	13.60	13.60	272	100.00

Chi-Square Test for Specified Proportions	
Chi-Square	0.0000
DF	1
Pr > ChiSq	0.9989

Vi får en *p*-værdi på 0,99 og kan dermed ikke afvise nulhypotesen på et 5% signifikansniveau. Og dermed kan den sande værdi godt være 13,6%. Dernæst testes, om den sande andel, der ikke har tillid til EU-parlamentet, er 15%:

no_trust_ep	Frequency	Percent	Test Percent	Cumulative Frequency	Cumulative Percent
0	231	84.93	85.00	231	84.93
1	41	15.07	15.00	272	100.00

Chi-Square Test for Specified Proportions	
Chi-Square	0.0012
DF	1
Pr > ChiSq	0.9729

Vi får en *p*-værdi på 0,97 og kan dermed ikke afvise nulhypotesen på et 5% signifikansniveau. Og dermed kan den sande værdi godt være 15%. Ligeldes kan vi nu godt forventet at andelen der har tillid til det nationale og EU parlamentet er ens.

***f**) Test om de to variable no_trust_ep og no_trust_prl er uafhængige*
Vi tester om de to variable er uafhængige via følgende stykke kode som udfører en chi-i-anden test

```
proc freq data=zero_trust;
tables no_trust_prl*no_trust_ep/chisq;
run;
```

Vi får følgende output

Statistics for Table of no_trust_prl by no_trust_ep

Statistic	DF	Value	Prob
Chi-Square	1	65.9076	<.0001
Likelihood Ratio Chi-Square	1	48.6896	<.0001
Continuity Adj. Chi-Square	1	61.9555	<.0001
Mantel-Haenszel Chi-Square	1	65.6653	<.0001
Phi Coefficient		0.4922	
Contingency Coefficient		0.4416	
Cramer's V		0.4922	

Vi betragter chi-i-anden p-værdien på mindre end 0,1% så på et 5 procents signifikansniveau kan vi altså ikke afvise nulhypotesen om, at *no_trust_prl* og *no_trust_ep* skulle være uafhængige.

***g**) Test om den gennemsnitlige tid, respondenterne bruger internettet, er den samme for respondenter med no_trust_ep=0 som for respondenter med no_trust_ep=1,*

Vi udfører et t-test på nulhypotesen at den daglig forbrug af internet er ens for de to grupper. Dette gøres via følgende kode

```
proc ttest data=zero_trust;
class no_trust_ep;
var netustm;
run;
```

Dette giver følgende output

no_trust_ep	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		245.4	220.5	270.2	191.7	175.6	210.9
1		190.0	133.6	246.4	178.7	146.7	228.6
Diff (1-2)	Pooled	55.3766	-7.9511	118.7	189.8	175.1	207.3
Diff (1-2)	Satterthwaite	55.3766	-5.9291	116.7			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	270	1.72	0.0863
Satterthwaite	Unequal	57.591	1.81	0.0758

Givet et 5% signifikansniveau fås en kritisk t-værdi i t fordelingen på 1,96 og testes begge t-teststørrelse ses det at begge størrelser er under den kritisk værdi og dermed kan vi ikke afvise at det daglig internetforbrug ikke er det samme for de to grupper. Det skal dog bemærkes at ved f.eks. et 10% niveau er resultatet signifikant og dermed er vi et grænsetilfælde og det kunne indikere en forskel mellem de to grupper.

h) *Udfør en regressionsanalyse på datasættet zero_trust, hvor variablen trstep er responsvariabel og begge de variable trstprl og netustm er forklarende variable. Kommenter resultatet.*
Vi ønsker altså at estimere multiple lineære regressionsmodel:

$$trstep = \beta_0 + \beta_1 trstprl + \beta_2 netustm + u$$

Dette gøres ligsom i opgave c dog med en ekstra forklarende variable

```
proc reg data=zero_trust;
model trstep=trstprl netustm;
run;
```

vi får følgende resultat.

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1.38659	0.26428	5.25	<.0001
trstprl	Trust in country's parliament	1	0.60466	0.05029	12.02	<.0001
netustm	Internet use, how much time on typical day, in minutes	1	0.00123	0.00062833	1.96	0.0509

Dette giver et estimat $\hat{\beta}_1 = 0{,}605$. Dette skal fortolkes som at alt andet lige effekten af, at tilliden til det nationale parlament øges med 1 point er, at tilliden til EU-parlamentet øges med 0,605 point. Ligeledes estimeres $\hat{\beta}_2 = 0{,}00123$. Dette skal fortolkes som at alt andet lige effekten af, daglig internetforbrug øges med 1 point er, at tilliden til det nationale parlament øges med 0,605 point. Vi estimerer ligeledes $\hat{\beta}_0 = 1{,}38$ som ikke har en fortolkning i en multipel model. $\hat{\beta}_1$ og $\hat{\beta}_0$ er signifikant på et 1% niveau og $\hat{\beta}_2$ er næsten signifikant på et 5% niveau.