Del 1

Opgave 1

```
proc import datafile='/courses/d284cd65ba27fe300/Sommerskole/Data/ESS9e03_renset.sav' out=
proc format;
value  STRAT
1='unge m nd'
2='mid. m nd'
3=' ldre   m nd'
4='unge kvinder'
5='mid. kvinder'
6=' ldre   kvinder'
;
data DK;
set ESS9;
if cntry='DK';
*** forblive i EU ****;
blive_EU=0;
if vteurmmb=1 then blive_EU=1;
**********************;
if agea < 40 then age1=1;
if 40<= agea< 70 then age1=2;
if agea =>70 then age1=3;
if age1=1 and gndr=1 then  STRAT =1;
if age1=2 and gndr=1 then  STRAT =2;
if age1=3 and gndr=1 then  STRAT =3;
if age1=1 and gndr=2 then  STRAT =4;
if age1=2 and gndr=2 then  STRAT =5;
if age1=3 and gndr=2 then  STRAT =6;
*** der s ttes v gte   ***;
if strat=1 then vgt=184/177;
if strat=2 then vgt=234/267;
if strat=3 then vgt= 75/ 94;
if strat=4 then vgt=185/146;
if strat=5 then vgt=226/228;
if strat=6 then vgt= 96/ 87;

vgt1=(4823689/1572)*vgt;
*** politiske v gte   ***;
if prtvtddk=1 then pvgt=.2628/.2748;
if prtvtddk=2 then pvgt=.0458/.0585;
if prtvtddk=3 then pvgt=.0335/.0423;
if prtvtddk=4 then pvgt=.0419/.0593;
if prtvtddk=5 then pvgt=.2018/.1398;
if prtvtddk=6 then pvgt=.0083/.0122;
if prtvtddk=7 then pvgt=.1947/.2439;
if prtvtddk=8 then pvgt=.0753/.0390;
if prtvtddk=9 then pvgt=.0780/.0740;
if prtvtddk=10 then pvgt=.0480/.0431;
if prtvtddk=11 then pvgt=.0009/.0130;
Npvgt=(4823689/1230)*pvgt;
run;
```

Jeg introducerer politiske vægte i datasættet. Her betegner pvgt den politiske vægt, npvgt er den politisk optimale vægt og blive EU angiver en dummy for, hvorvidt man vil blive i EU eller forlade EU.

Opgave 2

```
proc contents data=dk;
run;
```

Ud fra ovenstående programstump ses det, at der er 1572 observationer, med kun 1230 respondenter tildelt vægte. Dette vil sige, der mangler nogen observationer ift. til det oprindelige datasæt. Er det oprindelige datasæt repræsentativt og der en systematik i værdierne, der er manglende, så er datasættet DK ikke længere repræsentativt.

Opgave 3

```
proc surveyfreq data=dk;
table prtclddk;
weight pvgt;
run;
```

Which party feel closer to, Denmark					
prtclddk	Frequency	Weighted Frequency	Std Err of Wgt Freq	Percent	Std Err of Percent
Socialdemokratiet - The Social democrats	265	255.78080	13.36521	28.5569	1.5227
Det Radikale Venstre - The Radical Liberal Party	58	46.76913	6.02822	5.2216	0.6821
Det Konservative Folkeparti - Conservative People's Party	40	34.30105	5.53401	3.8296	0.6205
SF Socialistisk Folkeparti - Socialist People's Party	67	54.46908	6.61287	6.0813	0.7477
Dansk Folkeparti - Danish People's Party	116	161.79458	14.18730	18.0637	1.4913
Kristendemokraterne - Christian Democrats	9	6.12295	2.03194	0.6836	0.2277
Venstre, Danmarks Liberale Parti - The Liberal Party	213	176.96506	10.92403	19.7575	1.2761
Liberal Alliance - Liberal Alliance	27	43.53167	8.64747	4.8601	0.9452
Enhedslisten - Unity List - The Red-Green Alliance	71	71.41650	8.29312	7.9734	0.9251
Alternativet - The Alternative	31	32.76228	5.88315	3.6578	0.6554
Other	9	11.77463	4.16426	1.3146	0.4630
Total	906	895.68772	8.66191	100.0000	
Frequency Missing = 324					

Prognosen er baseret på, hvad respondenter siger, de gerne vil stemme. Her benyttes de politiske vægte oprettet i opg 1. Dette stemmer relativt. Prognosen ligger dog en smule fra, hvad der reelt blev stemt til folketingsvalget i 2019. Dette kan dog afhænge af, hvor langt tid før datasættet blev indsamlet.

Hvis man ikke vægter, så kan ikke være sikker på at prognosen ikke rammer ordentligt. Folk vil fx ikke indrømme, hvad de stemmer eller fx ikke deltage i undersøgelsen. Hermed kan nogle grupperinger være underrepræsenteret.

Opgave 4

```
Proc surveymeans data=dk N=4200000;
where cntry='DK';
var prtclddk;
weight pvgt;
run;
```

The SURVEYMEANS Procedure

Data Summary	
Number of Observations	1572
Number of Observations Used	1230
Number of Obs with Nonpositive Weights	342
Sum of Weights	1219.06931

Statistics						
Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
prtclddk	Which party feel closer to, Denmark	906	4.692160	0.103466	4.48909865	4.89522138

Her ses det, at ved en standardafvigelse på 10% procent, så vil konfidensintervallet være mellem ovenstående.

Opgave 5

Vi benytter ikke længere strat-vægten, som er baseret på køn og alder. Vi tager nu kun højde for de politiske vægte, som vi har genereret.

Opgave 6

Vi kan genere en vægt, hvor demografien har den samme indflydelse på variablene som de politiske holdninger. Vi kan opskrive vægten, som:

$$DEMPOLVGT = 0.5 * vgt + 0.5 * pvgt$$

Denne vægt kombinerer demografi, køn og alder, med politiske holdninger. Vægten er ikke perfekt, da vi har vægtet parameterne 50/50. En bedre vægt, havde været en, der baserer sig på respondenternes svar på "hjælpespørgsmål". Vægten forsøger dog at mixe demografi og politiske holdninger.

Opgave 7

Senor Hans Bay har gentagende gange sagt, at den bedste vægt i datasættet er *anweight*. Hans Bay er en klog mand med mange stuer. Antallet af stuer i København=klygtighed. Hil Hans Bay den store.

Opgave 8

```
Proc surveymeans data=dk N=420000;
var TRSTEP TRSTPRL NWSPOL blive_EU;
weight anweight;
run;
```

Statistics						
Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
trstep	Trust in the European Parliament	1458	5.189676	0.079877	5.0329899	5.3463614
trstprl	Trust in country s parliament	1560	6.000021	0.076635	5.8497038	6.1503391
nwspol	News about politics and current affairs, watching, reading or listening, in minutes	1571	70.429864	2.192348	66.1296264	74.7301010
blive_EU		1572	0.775115	0.013280	0.7490673	0.8011634

Det kan ses, at danskerne er overmiddel positivt stemt for EU parlamentet. Samtidig ses det dog, at de er mere positivt stemt overfor deres eget parlament. Dette hænger sammen med afstanden til EU.

Respondenterene føler sig nok ikke på samme måde hørt i EU-parlamentet, som i det danske, og man hører sjældent noget om det i medierne, hvorfor tilliden til dem er lavere. Samtidig mener de adspurgte at de læser, ser el nyheder 70 min.

Ydermere mener 77,5% af respondenterne, at de gerne vil forblive i EU. Den lave konfidensintervalgrænse ligger på 75%, hvorfor det med sikkerhed kan siges, at majoriteten af den danske befolkning vil forblive i EU.

Opgave 9

```
data skaladata;
set dk;
skala1 = mean(of TRSTlgl TRSTplc TRSTplt TRSTprl TRSTprt TRSTep
TRSTun);
run;
proc means data=skaladata;
var  TRSTlgl TRSTplc TRSTplt TRSTprl TRSTprt TRSTep TRSTun skala1;
run;
data skaladata2;
set dk;
skala2 = mean(of stfdem stfeco stfedu stfgov stfhlth);
run;
proc means data=skaladata2;
var stfdem stfeco stfedu stfgov stfhlth skala2;
run;
```

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
trstlgl	Trust in the legal system	1555	7.6836013	1.9643430	0	10.0000000
trstplc	Trust in the police	1572	8.0031807	1.8001669	0	10.0000000
trstplt	Trust in politicians	1561	5.2011531	2.0426287	0	10.0000000
trstprl	Trust in country s parliament	1560	6.1666667	2.2181632	0	10.0000000
trstprt	Trust in political parties	1550	5.3045161	2.0472577	0	10.0000000
trstep	Trust in the European Parliament	1458	5.1927298	2.2610733	0	10.0000000
trstun	Trust in the United Nations	1474	6.4972863	2.0773386	0	10.0000000
skala1		1572	6.2965588	1.5791711	0	10.0000000

Figure 1: Trustvariable Skala

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
stfdem	How satisfied with the way democracy works in country	1550	7.3406452	2.0080759	0	10.0000000
stfeco	How satisfied with present state of economy in country	1531	7.1476159	1.9187140	0	10.0000000
stfedu	State of education in country nowadays	1547	7.7996122	1.7961940	0	10.0000000
stfgov	How satisfied with the national government	1527	5.3817944	2.1412160	0	10.0000000
stfhlth	State of health services in country nowadays	1559	6.6703015	2.0507888	0	10.0000000
skala2		1572	6.8665394	1.4258849	0	10.0000000

Figure 2: Satisfaction Skala

Derved er det muligt at konkludere, man kan konsturere skalaer for de to variabel katagorier.

Del 2

Opgave 1

```
proc import datafile ='/courses/d284cd65ba27fe300/Sommerskole/Data/pisa_dnk_18_renset.sav'
dbms=sav replace out=DK18;
run;

proc reg data=DK18;
model score_m=ESCS;
run;
```

Root MSE	72.15139	R-Square	0.1411
Dependent Mean	498.38473	Adj R-Sq	0.1410
Coeff Var	14.47705		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	484.80648	0.92280	525.36	<.0001
ESCS	Index of economic, social and cultural status	1	34.79815	0.99589	34.94	<.0001

Det kan udledes, at R^2-forklaringsgraden er 14,1%. Estimatet er givet ved $34,8$. Hver gang ESCS stiger med en enhed, så ændres score m med 34,8. Dette afhænger dog af, hvilken form for variabel ESCS og score m er. Det kan også være procent el.

Opgave 2

```
proc mi data=dk18 seed=1 Nimpute=1 /*noprint*/ out=outmi;
em itprint  outem=outem;
var score_m escs;
run;
```

Missing Data Patterns

Group	score_m	ESCS	Freq	Percent	Group Means	
					score_m	ESCS
1	X	X	7431	97.05	498.384732	0.390200
2	X	.	226	2.95	460.584865	.

Opgave 3

```
proc reg data=outmi;
model score_m=escs;
run;
```

Root MSE	72.15365	R-Square	0.1406
Dependent Mean	497.26905	Adj R-Sq	0.1405
Coeff Var	14.50998		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	483.96567	0.90619	534.06	<.0001
ESCS	Index of economic, social and cultural status	1	34.62817	0.97834	35.39	<.0001

Det kan ses, at forklaringsgraden og estimatet er en smule formindsket. Men alt i alt hænger dette godt sammen med metoden brugt. Denne danner de mangelende observationer ud fra sammenhængen med andre variable, der stadig er der.

Opgave 4

```
proc surveyimpute seed=1 data=dk18 method=hotdeck(selection=srswr);
var score_m escs;
output out=dk18mimpute;
run;
```

Imputation Summary

Observation Status	Number of Observations	Sum of Weights
Nonmissing	7431	7431
Missing	226	226
Missing, Imputed	226	226
Missing, Not Imputed	0	0

Fra ovenstående tabel kan det udledes, at 226 manglende observationer er dannet. Derved er hullerne i datasættet lukket

Opgave 5

```
proc reg data=dk18mimpute;
model score_m=escs;
run;
```

Root MSE	72.50889	R-Square	0.1322
Dependent Mean	497.26905	Adj R-Sq	0.1320
Coeff Var	14.58142		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	484.07906	0.91426	529.48	<.0001
ESCS	Index of economic, social and cultural status	1	33.67135	0.98620	34.14	<.0001

Det kan udledes, at R^2-forklaringsgraden er faldet og beta-estimatet er mindre end tidligere spørgsmål. Dette hænger sammen med, at disse observationer er random generated, hvorfor sammenhængen bliver mindre i datasættet.

Opgave 6

Analysen giver de bedste resultater, hvis man bruger proc mi-metoden, da disse resultater fastholder varians og beta-estimatet, pga observationer er dannet efter sammenhængen mellem andre variable i observationen selv. Dette kan dog give et forkert billede af den virkelige sammenhæng, da man bearbejder resultaterne til sin fordel. Nogle gange vil det være en fordel at få random generated variable, da dette vil styrke analysen, men vil formentligt sænke sammenhængen mellem observationerne.

Opgave 7

```
data pisany;
set dk18;
if STRATUM='DNK0101' then vgt=0.066/0.269;
if STRATUM='DNK0202' then vgt=0.249/0.267;
if STRATUM='DNK0303' then vgt=0.505/0.350;
if STRATUM='DNK0404' then vgt=0.180/0.115;
Nvgt=(59968/7657)*vgt;
run;
```

Ved at aflæse hhv. univers- og stikprøvevægte i den givne tabel og indsætte både universstørrelse og stikprøvestørrelsen. Koden er ovenstående er benyttet.

Opgave 8

```
proc means data=pisany;
var score_m score_n score_r;
weight vgt;
run;
```

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
score_m	matematik scoren	7657	508.5304818	75.6400008	236.0975000	738.6043000
score_n	naturfag scoren	7657	491.5053314	86.1457680	163.8209000	757.4121000
score_r	laese scoren	7657	499.6332838	88.0033467	188.6991000	745.9385000

Opgave 9

```
proc means data=pisany;
var score_m score_n score_r;
weight W_fstuwt;
run;
```

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
score_m	matematik scoren	7657	509.3983745	213.1718847	236.0975000	738.6043000
score_n	naturfag scoren	7657	492.6370334	241.5883212	163.8209000	757.4121000
score_r	laese scoren	7657	501.1299338	248.2160909	188.6991000	745.9385000

Det kan ses, at scoren bliver højere. Altså vægtes det anderledes eller også bliver børnene bare bedre i skolen.

Opgave 10

```
proc means data=dk18;
class skoleid;
var score_r escs;
output out=b mean=;
proc reg data=b;
model score_r=escs;
run;
```

Programstumpen giver os først gennemsnittet på de forskellige skoler. Dette er baseret dens skoleid. Her findes gennemsnittene for hver skole. Efterfølgende laver man en regression på disse gennemsnit. Det udledes, at jo højere socioeconmisk status skolen samlede set har, det bedre score børnene. Forklaringsgraden er nu på 39,57%, mens beta-estimatet er $71,1$.

Root MSE	37.61352	**R-Square**	0.3957
Dependent Mean	488.23633	**Adj R-Sq**	0.3939
Coeff Var	7.70396		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	463.00278	2.61843	176.82	<.0001
ESCS	Index of economic, social and cultural status	1	71.06529	4.71459	15.07	<.0001