Rettevejledning "Programmering og Anvendt Statistik med SAS" juni 2014

1.1

Program:

```
ods trace on;
proc contents data=sasprog.Programmering_stdhuse;
run;
ods trace off;
proc contents data=sasprog.Programmering_stdhuse;
ods output Variables=asdf;
run;
Proc print; run;
```

Ods trace viser i loggen navnene på de forskellige output elementer som proc contents danner.
Herefter bruges ”ods output”, hvoraf output-elemntet ”variables” gemmes i et ny datasæt kaldet ”asdf” i workmappen.

1.2

Den første bid danner et nyt tomt datasæt kaldet ”var_votes” i mappen work.

```
data var_votes;
run;
```

Der bliver nu dannet et nyt datasæt kaldet ”var_list” også i mappen work,

```
data var_list;
```

Denne linje siger at vores nye datasæt dannes ud fra datasættet ”programmering_vars”

```
set sasprog.Programmering_vars;
```

Der dannes nu en ny variabel kaldet ”random”, der for hver observation får en tilfældig værdi af en uniform fordeling fra 0-1. Tallet 12345 er et seed, der sikrer, at der dannes de samme tilfældige tal, hver gang programmet afvikles.

```
random = ranuni(12345);
run;
```

Næste bid siger at vi nu vil sortere i datasættet ”var_list” efter variablen "random"

```
proc sort data=var_list;
by random;
run;
```

På denne måde dannes et datasæt af variabelnavnene i datasættet Programmeinrng_vars med variabelnavnene sorteret i tilfældig rækkefølge.

Der laves igen et nyt datasæt i work mappen, denne gang kaldet “test_list”.

```
data test_list;
```

Det nye datasæt bliver defineret som det gamle datasæt “var_list”, men kun med de to første observationer, og kun variablen “name” bliver beholdt.

```
set var_list (obs=2 keep=name);/*I makroen benyttes obs=&varnum og ikke obs=2*/
run;
```

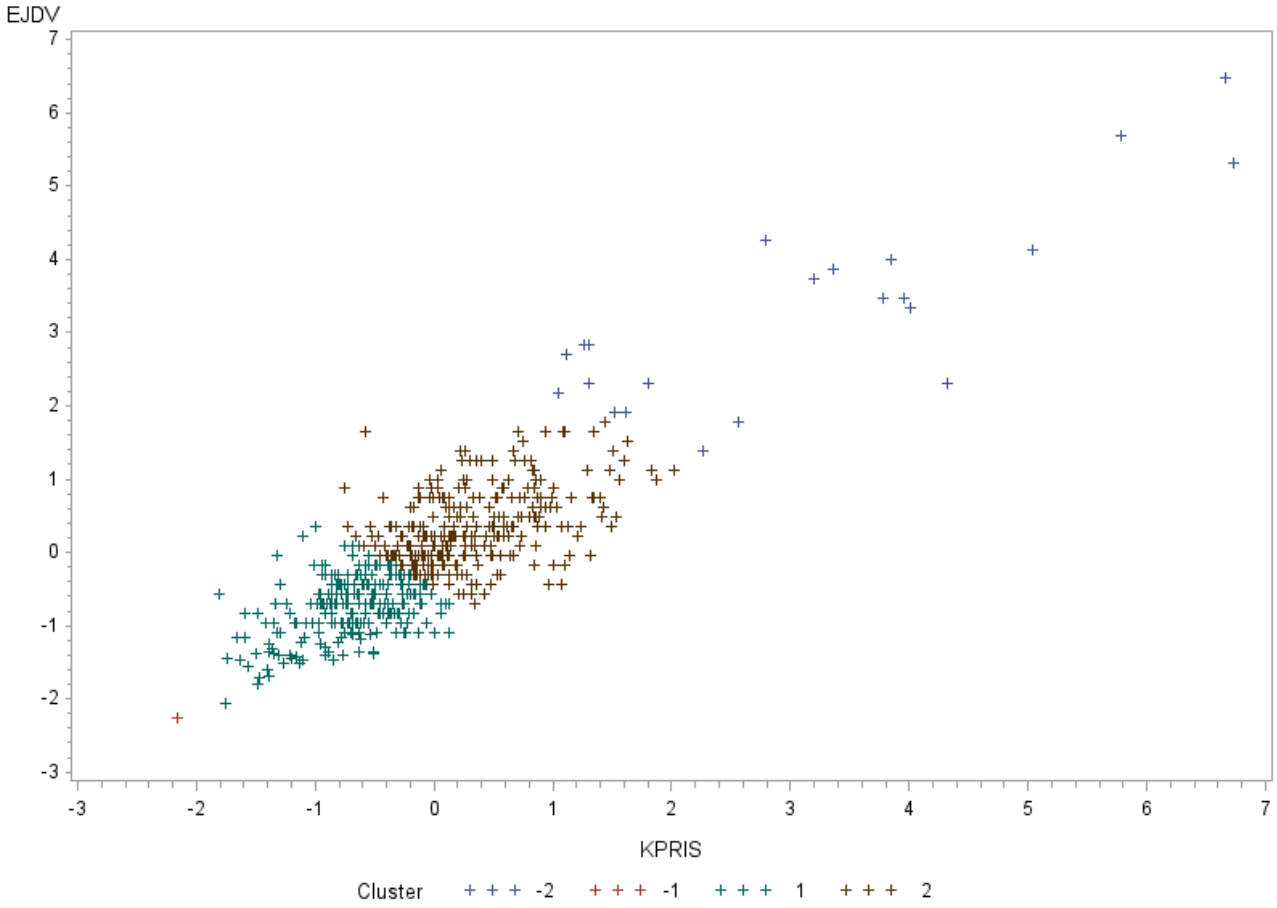
Datasættet test_list vil altså indeholde 2 tilfældigt udvalgte variabelnavne.

2

```
proc fastclus data=sasprog.Programmering_stdhuse maxclusters=100 maxiter=0
outstat=cluster_stat outseed=mean1;
var ejdv kpris;
run;
Data seed;
set mean1;
if _freq_>15; *Her sorteres alle de klustre, der har kun har 15 eller færre
observationer, fra;
run;
*Her sorteres alle de klustre, der har kun har 15 eller færre observationer, fra;
Proc fastclus data=sasprog.Programmering_stdhuse seed=seed strict=2 maxc=2
out=ny;
var ejdv kpris;
run;
Proc sort data=ny;
by cluster;
run;
Proc means;
var ejdv kpris;
by cluster;
run;
Proc gplot;
plot ejdv*kpris=cluster;
run;
```

Resultatet kan formidles ved den dannede graf, der viser to klustre, der skærer cigaren midt over og outliere blandt de dyreste og det allerbilligste hus. Outlierne er klustrene med negativt fortegn, hvor den numeriske værdi angiver det nærmeste egentlige kluster. Kommentarer af denne type bør medtages.

Hvis det første kald af Proc Fastclus udelades, så der ikke findes initalseeds hvor initialklustre med under fx 15 observationer udelades, risikeres, at der dannes et meget lille kluster af de dyreste huse.



3.1

```
data rsq;
set cluster_stat;
if _type_ = 'RSQ';
drop _type_ cluster over_all;
run;
proc transpose data=rsq out=rsq2;
run;
```

Koden danner et nyt datasæt kaldet ”rsq” i ”work” mappen. Det tager udgangspunkt i datasættet ”cluster_stat”. ”If” funktionen siger, at kun observationer, hvor variablen _type_ er lig RSQ, skal beholdes til det nye datasæt. ”Drop” funktionen dropper herefter de tre variable ”_type_, cluster og over_all” så kun variablene EJDS og KPRIS er tilbage

Bagefter transponeres datasættet, og det nye transponerede datasæt kaldes ”rsq2”. Dette datasæt vil indeholde RSQ-værdien for de to observationer, dvs variablene EJDV og KPRIS i en variabel kaldet COL1.

3.2

```
data rsq2;
set rsq2;
length variable_name $32.;
name = _name_;
run;
proc sort data=rsq2;
by descending col1;
run;
data var_votes;
set var_votes rsq2(obs=1);
run;
```

Denne kode danner datasættet ”rsq2” ud fra datasættet "rsq". Først tilføjes en ny variabel ”variable_name”, hvor ”length” funktionen siger, at den skal fylde 32 tegn i outputvinduet. Herefter dannes en ny variabel kaldet ”name”, der er lig den gamle variabel ”_name_”, og tager de samme værdier.

Programmet sorteres nu efter variablen col1, men som descending og ikke ascending, der er standard. Derved kommer den største værdi af col1, dvs den største værdi af RSQ øverst. Til sidst overskrives det gamle dataset” var_votes”, med sig selv, og datasættet rsq2, hvor kun den første observation medtages. Da ”var_votes” oprindeligt er et tomt dokument bliver der først oprettet en tom observation og herefter kommer observationen fra rsq2 ind. Der suppleres altså op med den variabel, der har den højeste RSQ-værdi.

4.1

Kaldet af makrovariblen &lib afsluttes som oftest med en blank. Men i denne forbindelse skal det være et punktum lige efter det dannede biblioteksnavn, så datasætnavnet kan komme lige efter. For at løse dette problem kan kaldet af en makrovariable udover at afsluttes med en blank også afsluttes med et punktum. Det ene punktum er altså punktummet mellem libname og datasætnavn - det andet er afslutning af et kald af en makrovariabel.

4.2

%put skriver noget til logvinduet. Her skrives indholdet af en makrovariabel, der kaldes med & efterfulgt af makronavnet.

4.3

Her skal redegøres for betydningen af %, &. Desuden skal betydningen af do-løkken forklares:

I løkken udvælges først et antal (&varnum) variabelnavne tilfældigt. Dernæst udføres clusteranalysen og det registreres, hvilken af de udtrukne variable, der har den største RSQ. Dette variabelnavn tilføjes så datasættet var_votes.

5

Makroen udvælger stikprøver af potentielle variable til en klusteranalyse. Derefter vælges den variable der fra hver stikprøve forklarer bedst. Til sidst skrives i frekvenstabellen ud hvor mange gange de enkelte variable havner som den bedste adskiller. Den bedste adskiller vil ofte være den bedste forklarende variable i en fx en regressionsmodel.

En fuld beskrivelse af makroen funktion kan ikke forventes i en besvarelse til topkarakter. Men det giver plus ved bedømmelsen, når der vises en delvis forståelse for at RSQ værdien overføres til variablen COL1, så der til sidst gives en linie til den pågældende variable i datasættet var_votes. Disse delelementer er en del af pensum

Makroen er sakset fra følgende kilde, hvori dens funktion yderligere beskrives:

http://support.sas.com/resources/papers/proceedings14/1300-2014.pdf