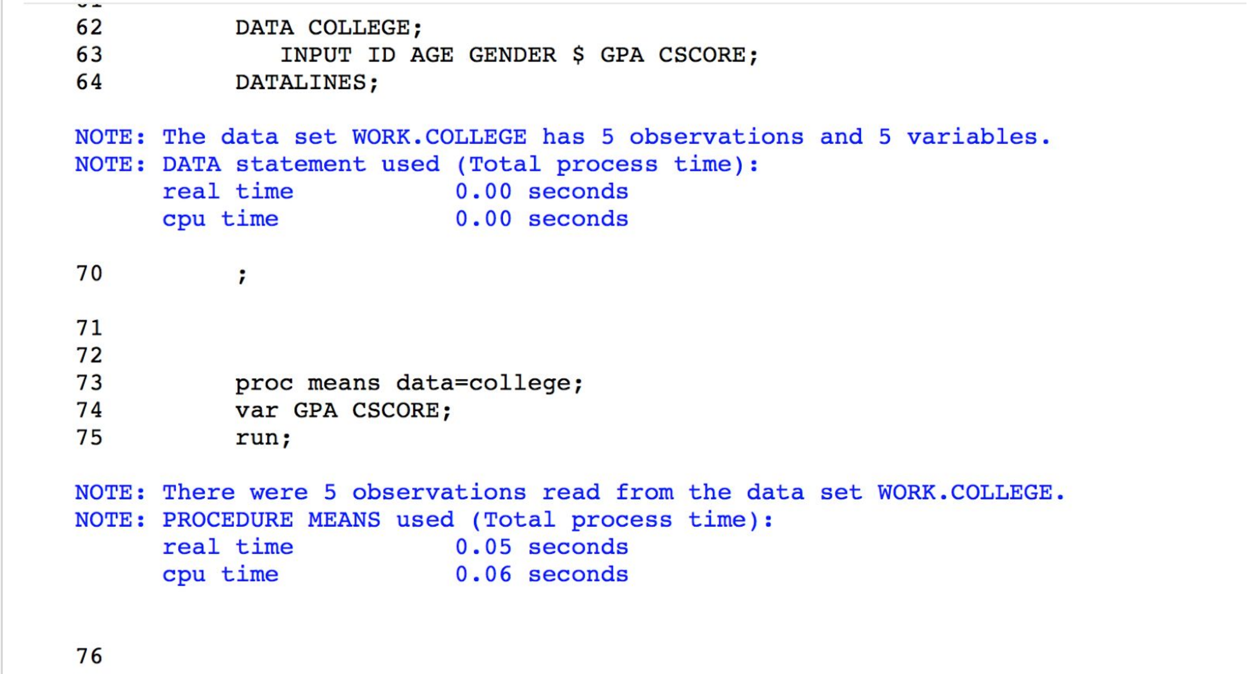
Homework 9

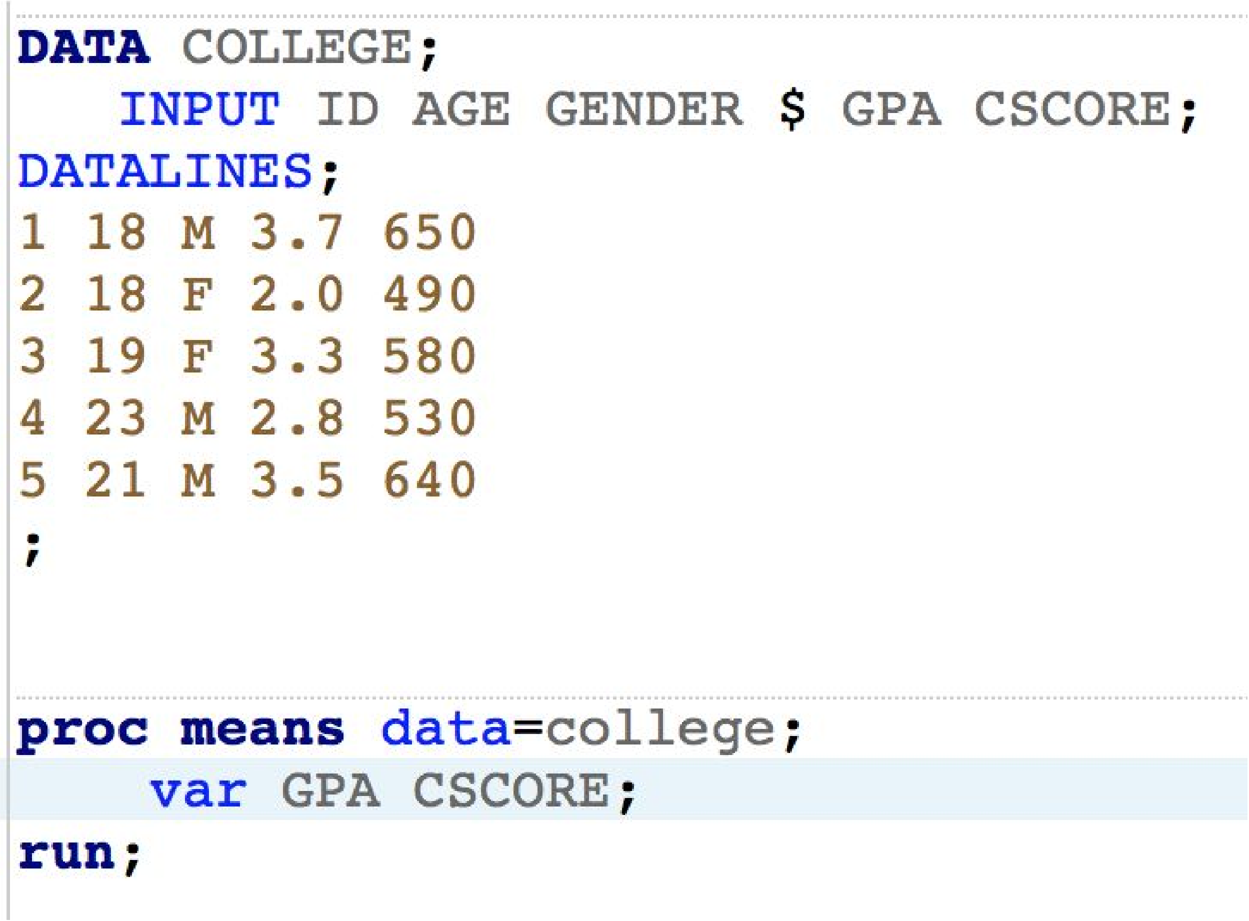
1. Create the following SAS dataset on 5 college students:

```
DATA COLLEGE;
   INPUT ID AGE GENDER $ GPA CSCORE;
DATALINES;
1 18 M 3.7 650
2 18 F 2.0 490
3 19 F 3.3 580
4 23 M 2.8 530
5 21 M 3.5 640
;
```

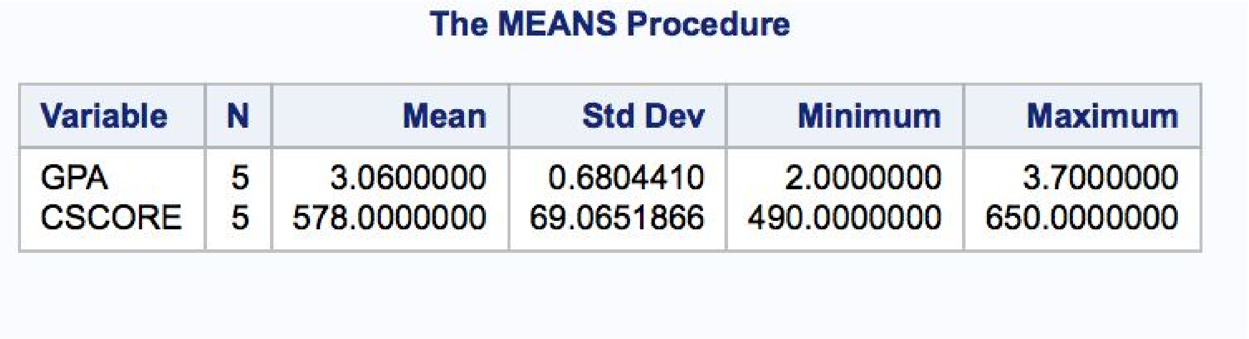
(a) Add statements necessary to compute the mean grade point average and mean college entrance exam score.



```
          DATA COLLEGE;
             INPUT ID AGE GENDER $ GPA CSCORE;
          DATALINES;

NOTE: The data set WORK.COLLEGE has 5 observations and 5 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds

          ;



          proc means data=college;
          var GPA CSCORE;
          run;

NOTE: There were 5 observations read from the data set WORK.COLLEGE.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.05 seconds
      cpu time            0.06 seconds
```



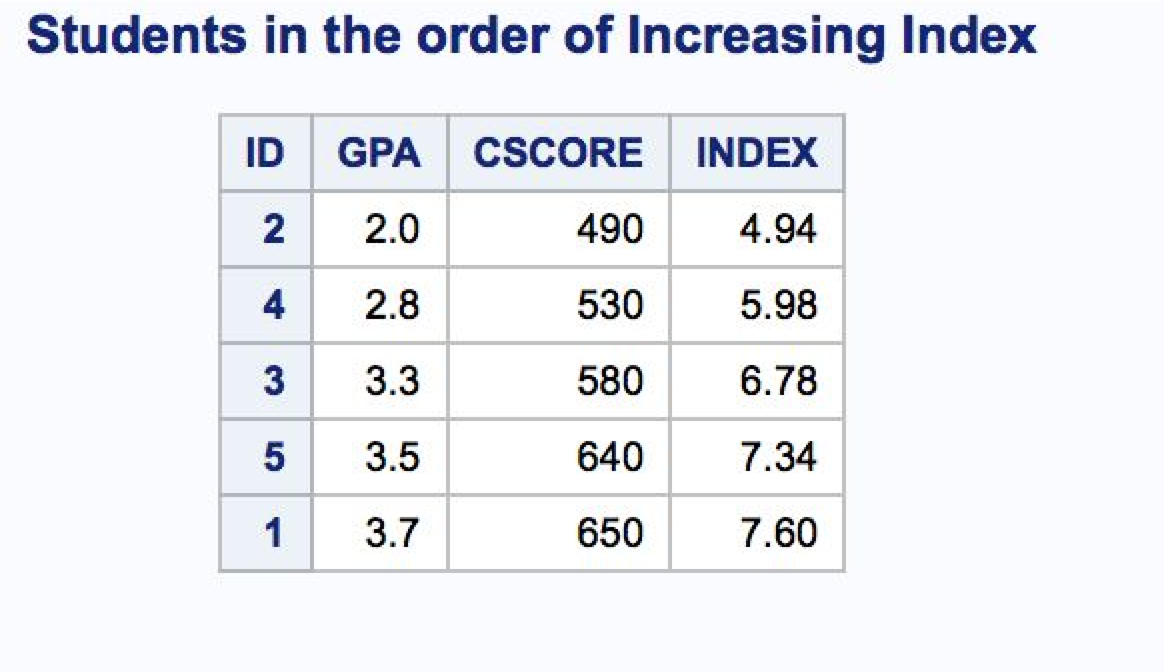
```
DATA COLLEGE;
   INPUT ID AGE GENDER $ GPA CSCORE;
DATALINES;
1 18 M 3.7 650
2 18 F 2.0 490
3 19 F 3.3 580
4 23 M 2.8 530
5 21 M 3.5 640
;


proc means data=college;
    var GPA CSCORE;
run;
```



The MEANS Procedure

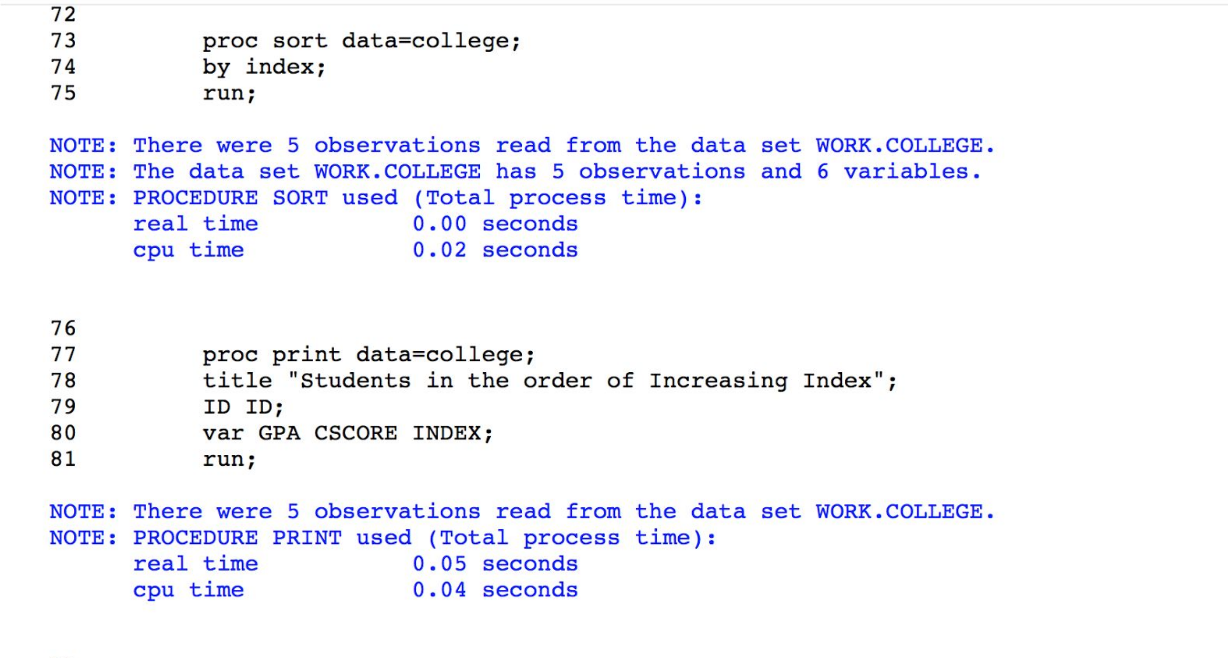
Variable	N	Mean	Std Dev	Minimum	Maximum
GPA CSCORE	5 5	3.0600000 578.0000000	0.6804410 69.0651866	2.0000000 490.0000000	3.7000000 650.0000000

(b) We want to compute an index for each subject, as follows:
INDEX=GPA + 3 x CSCORE/500
Modify your program to compute the INDEX for each student and to print a list of students in order of increasing INDEX. Include in your listing the student ID, GPA, CSCORE and INDEX.

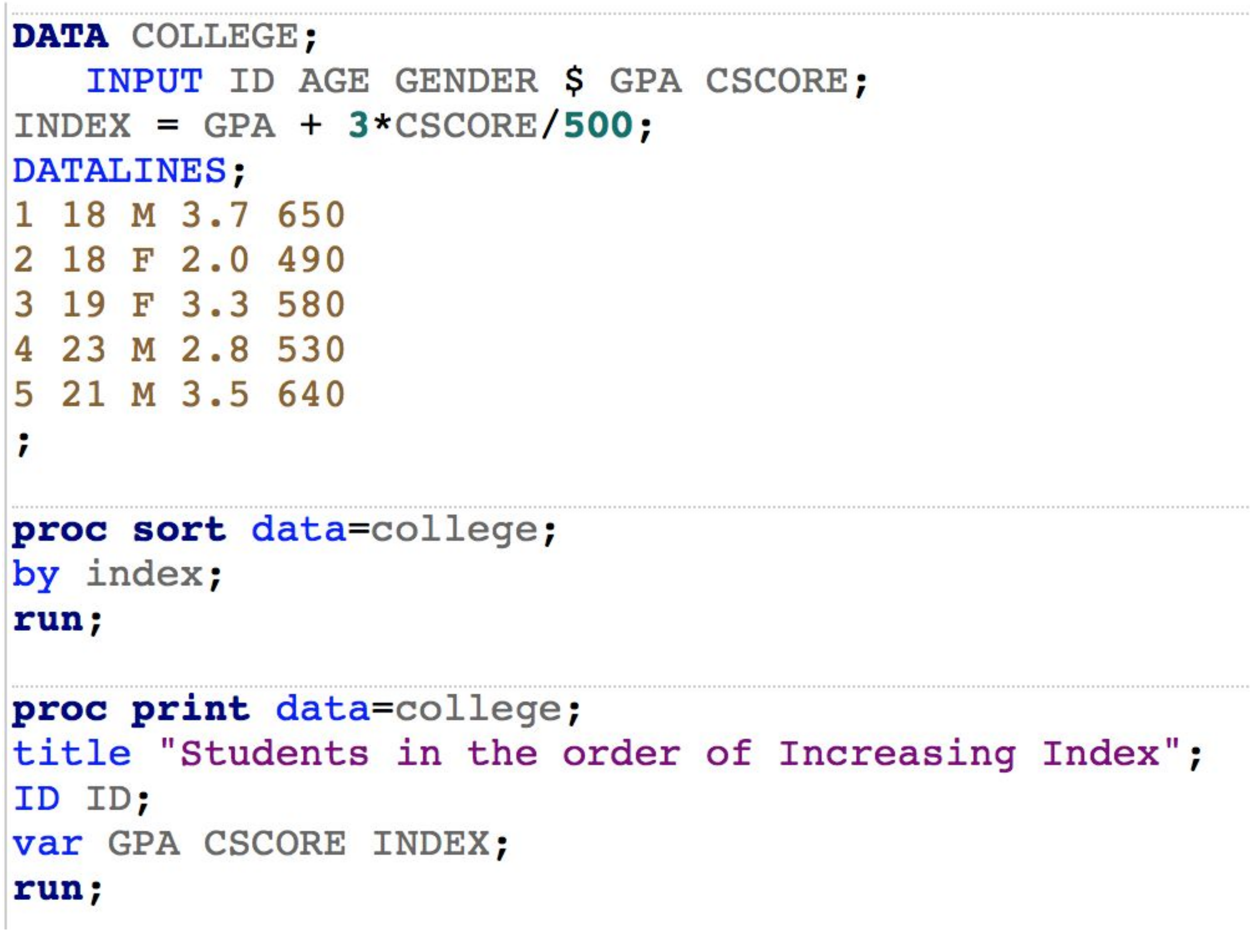


Students in the order of Increasing Index

ID	GPA	CSCORE	INDEX
2	2.0	490	4.94
4	2.8	530	5.98
3	3.3	580	6.78
5	3.5	640	7.34
1	3.7	650	7.60

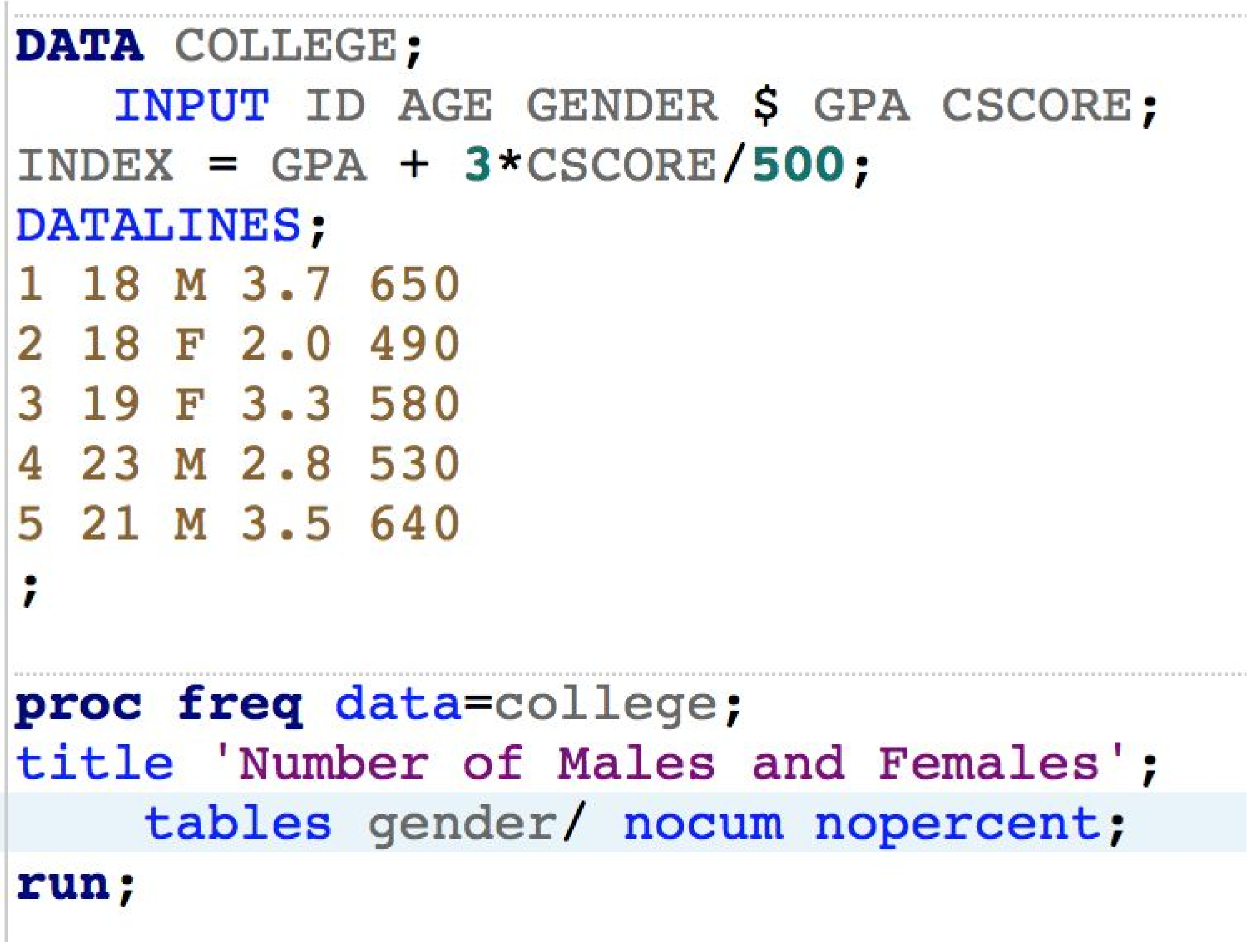


```
72
73          proc sort data=college;
74          by index;
75          run;

NOTE: There were 5 observations read from the data set WORK.COLLEGE.
NOTE: The data set WORK.COLLEGE has 5 observations and 6 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time           0.00 seconds
      cpu time            0.02 seconds


76
77          proc print data=college;
78          title "Students in the order of Increasing Index";
79          ID ID;
80          var GPA CSCORE INDEX;
81          run;

NOTE: There were 5 observations read from the data set WORK.COLLEGE.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.05 seconds
      cpu time            0.04 seconds
```

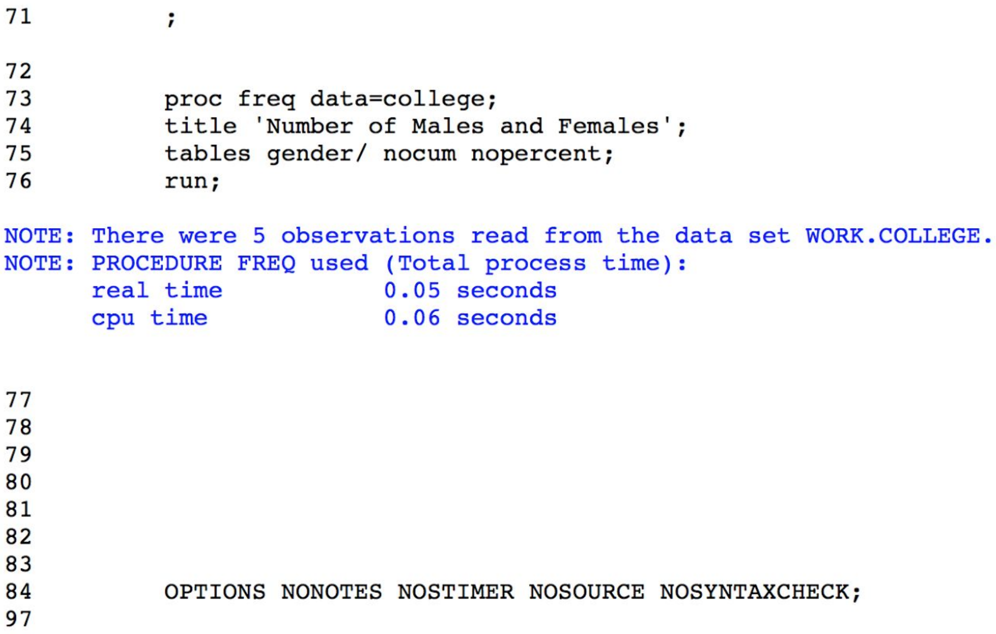


```
DATA COLLEGE;
   INPUT ID AGE GENDER $ GPA CSCORE;
INDEX = GPA + 3*CSCORE/500;
DATALINES;
1 18 M 3.7 650
2 18 F 2.0 490
3 19 F 3.3 580
4 23 M 2.8 530
5 21 M 3.5 640
;

proc sort data=college;
by index;
run;

proc print data=college;
title "Students in the order of Increasing Index";
ID ID;
var GPA CSCORE INDEX;
run;
```

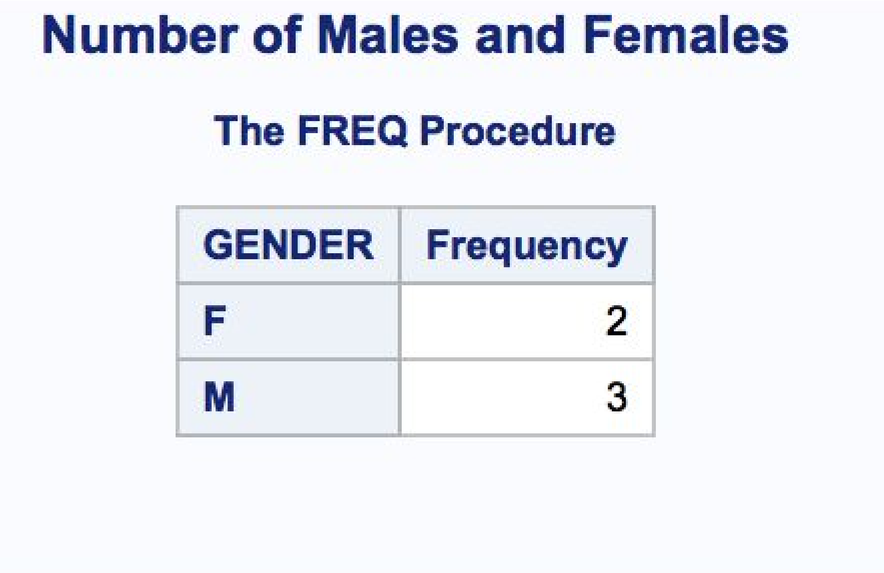
2. Add the necessary statements to compute the number of males and females in the previous problem.



```
DATA COLLEGE;
   INPUT ID AGE GENDER $ GPA CSCORE;
INDEX = GPA + 3*CSCORE/500;
DATALINES;
1 18 M 3.7 650
2 18 F 2.0 490
3 19 F 3.3 580
4 23 M 2.8 530
5 21 M 3.5 640
;

proc freq data=college;
title 'Number of Males and Females';
    tables gender/ nocum nopercent;
run;
```



```
          ;


          proc freq data=college;
          title 'Number of Males and Females';
          tables gender/ nocum nopercent;
          run;

NOTE: There were 5 observations read from the data set WORK.COLLEGE.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.05 seconds
      cpu time            0.06 seconds









          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;

```



Number of Males and Females

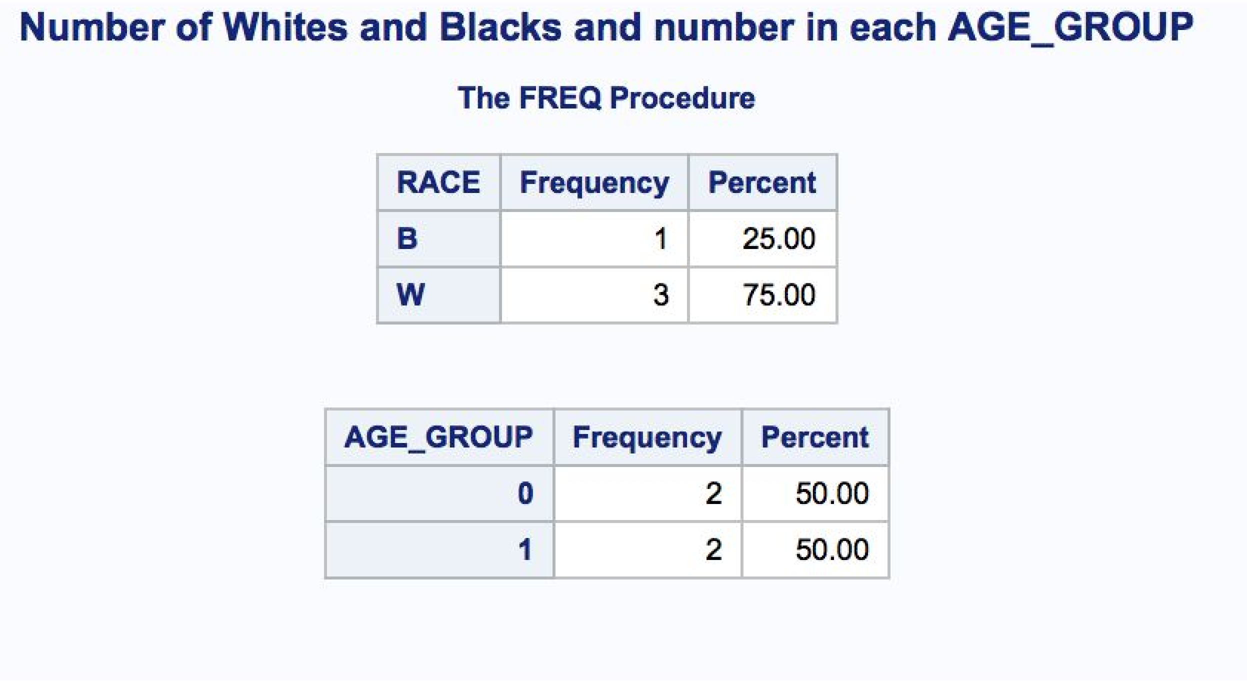
The FREQ Procedure

GENDER	Frequency
F	2
M	3

3. Use the data below and create a new variable (AGE_GROUP) that has a value of 1 for ages between 0 and 35 and 2 for ages greater than 35.

```
DATA TAXPROB;
      INPUT SS SALARY AGE RACE $;
      FORMAT SS SSN11.;
   DATALINES;
   123874414 28000 35 W
   646239182 29500 37 B
   012437652 35100 40 W
   018451357 26500 31 W
   ;
```

Compute the number of whites (W) and blacks (B) and the number in each age group. Use the appropriate option to omit cumulative statistics from the output.

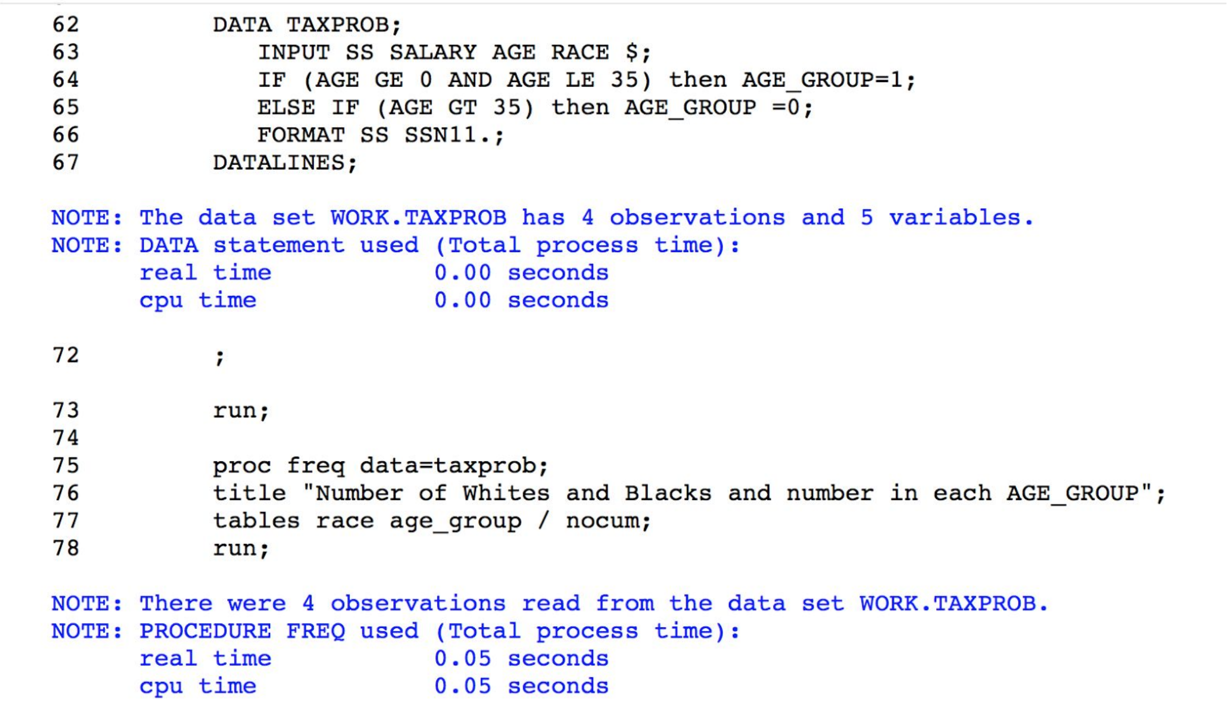


Number of Whites and Blacks and number in each AGE_GROUP

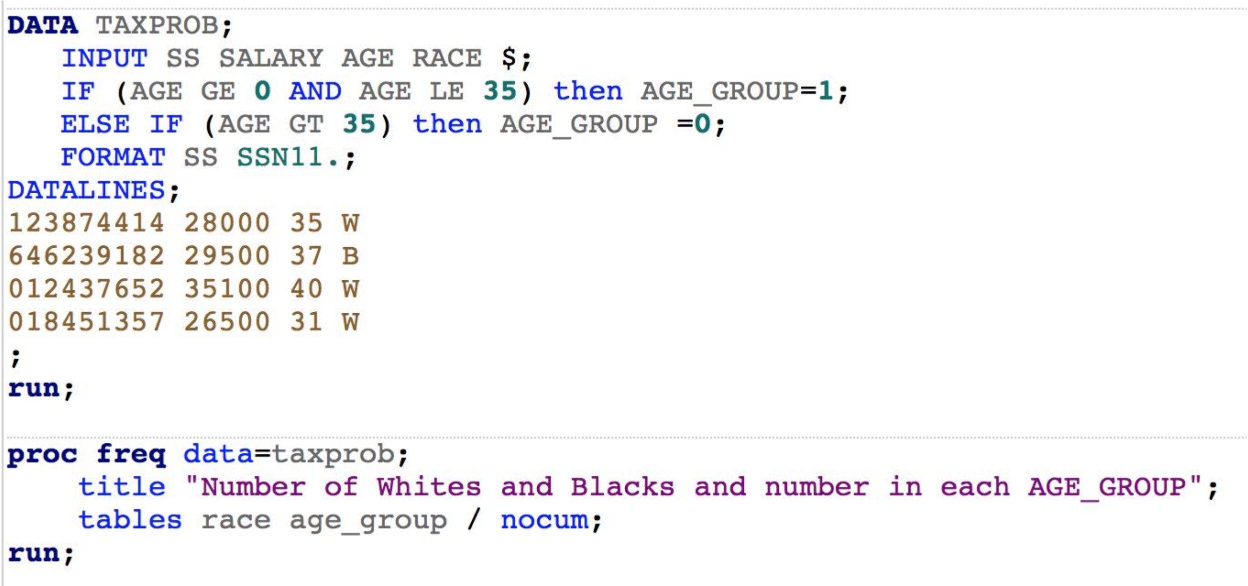
The FREQ Procedure

RACE	Frequency	Percent
B	1	25.00
W	3	75.00

AGE_GROUP	Frequency	Percent
0	2	50.00
1	2	50.00



```
62          DATA TAXPROB;
63             INPUT SS SALARY AGE RACE $;
64             IF (AGE GE 0 AND AGE LE 35) then AGE_GROUP=1;
65             ELSE IF (AGE GT 35) then AGE_GROUP =0;
66             FORMAT SS SSN11.;
67          DATALINES;

NOTE: The data set WORK.TAXPROB has 4 observations and 5 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds

72          ;

73          run;
74
75          proc freq data=taxprob;
76          title "Number of Whites and Blacks and number in each AGE_GROUP";
77          tables race age_group / nocum;
78          run;

NOTE: There were 4 observations read from the data set WORK.TAXPROB.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.05 seconds
      cpu time            0.05 seconds
```

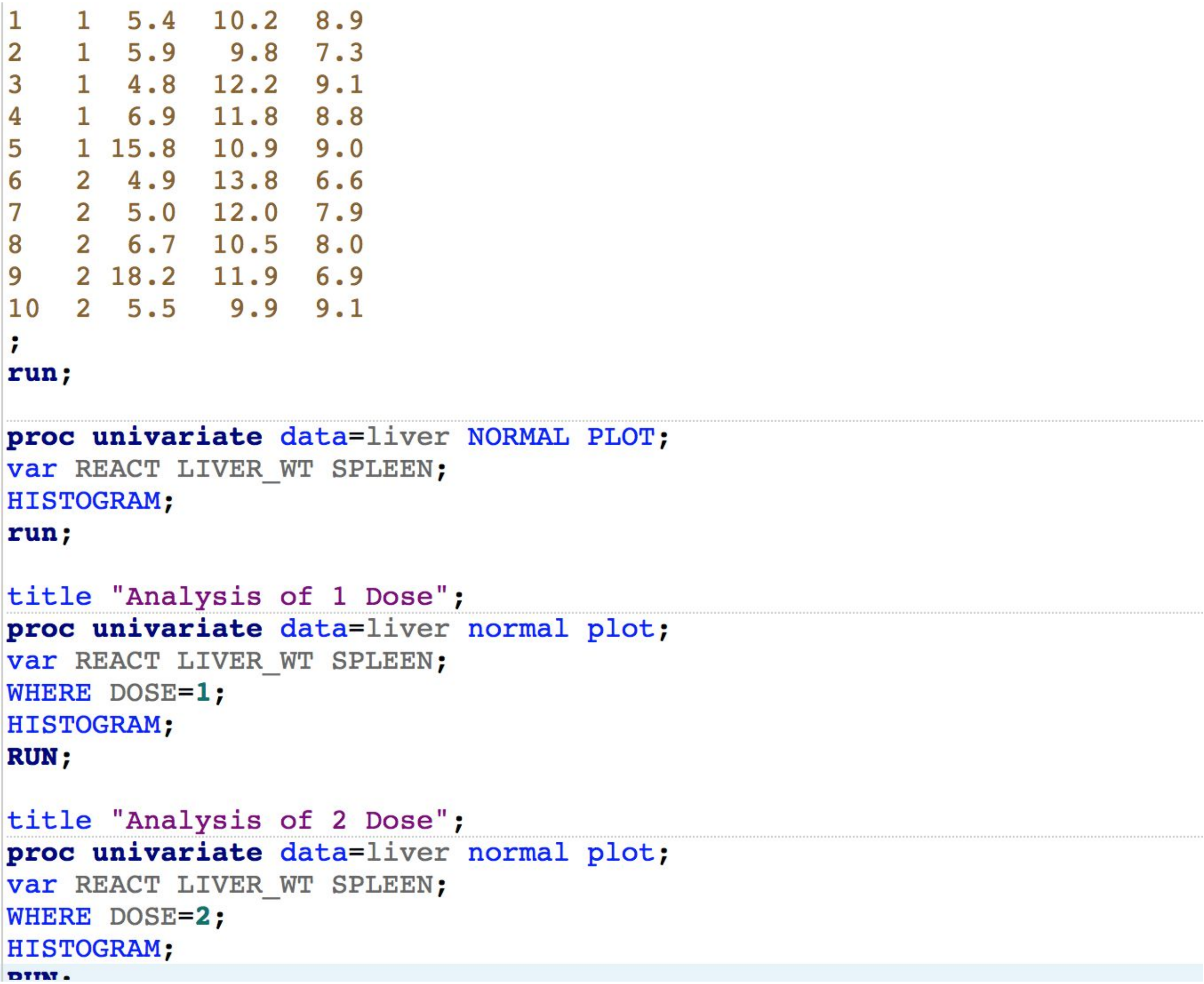


```
DATA TAXPROB;
   INPUT SS SALARY AGE RACE $;
   IF (AGE GE 0 AND AGE LE 35) then AGE_GROUP=1;
   ELSE IF (AGE GT 35) then AGE_GROUP =0;
   FORMAT SS SSN11.;
DATALINES;
123874414 28000 35 W
646239182 29500 37 B
012437652 35100 40 W
018451357 26500 31 W
;
run;

proc freq data=taxprob;
    title "Number of Whites and Blacks and number in each AGE_GROUP";
    tables race age_group / nocum;
run;
```

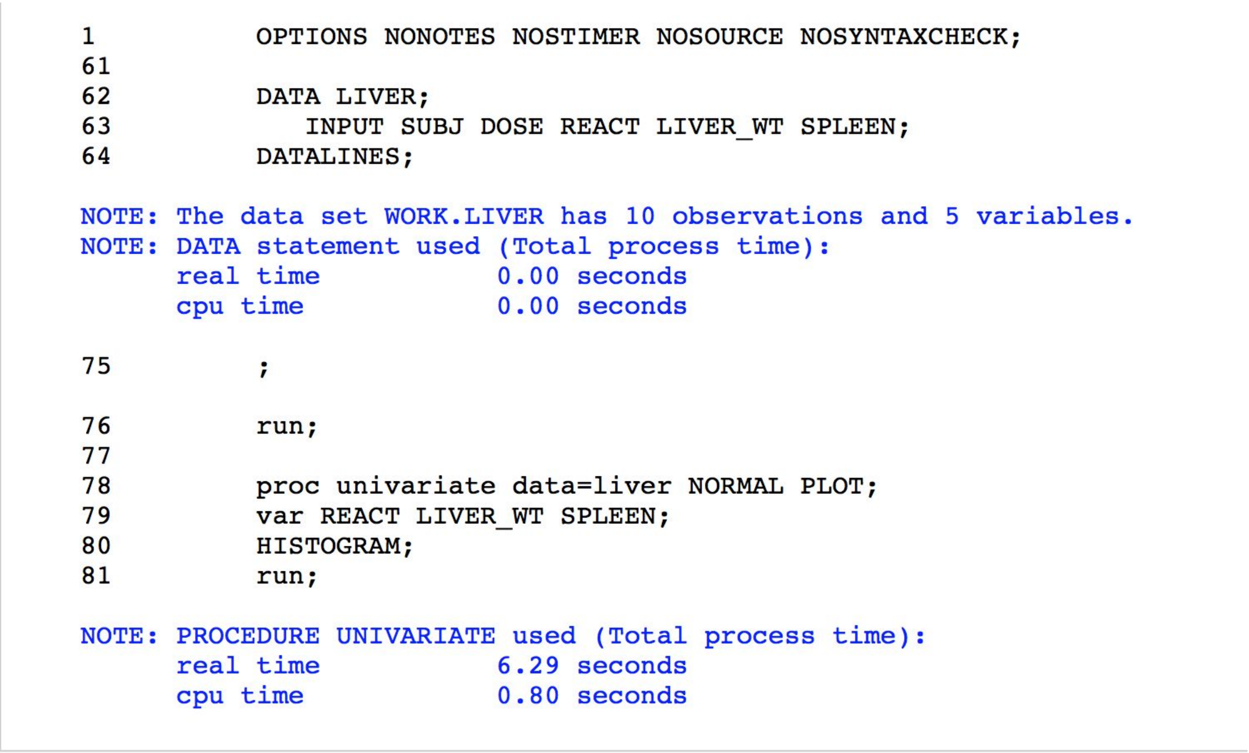
4. Use this data and PROC UNIVARIATE to produce histograms, normal probability plots, and boxplots and test the distributions for normality. Do this for variables like REACT, LIVER_WT, and SPLEEN, first for all subjects and then separately for each of the two DOSES.

```
DATA LIVER;
   INPUT SUBJ DOSE REACT LIVER_WT SPLEEN;
```

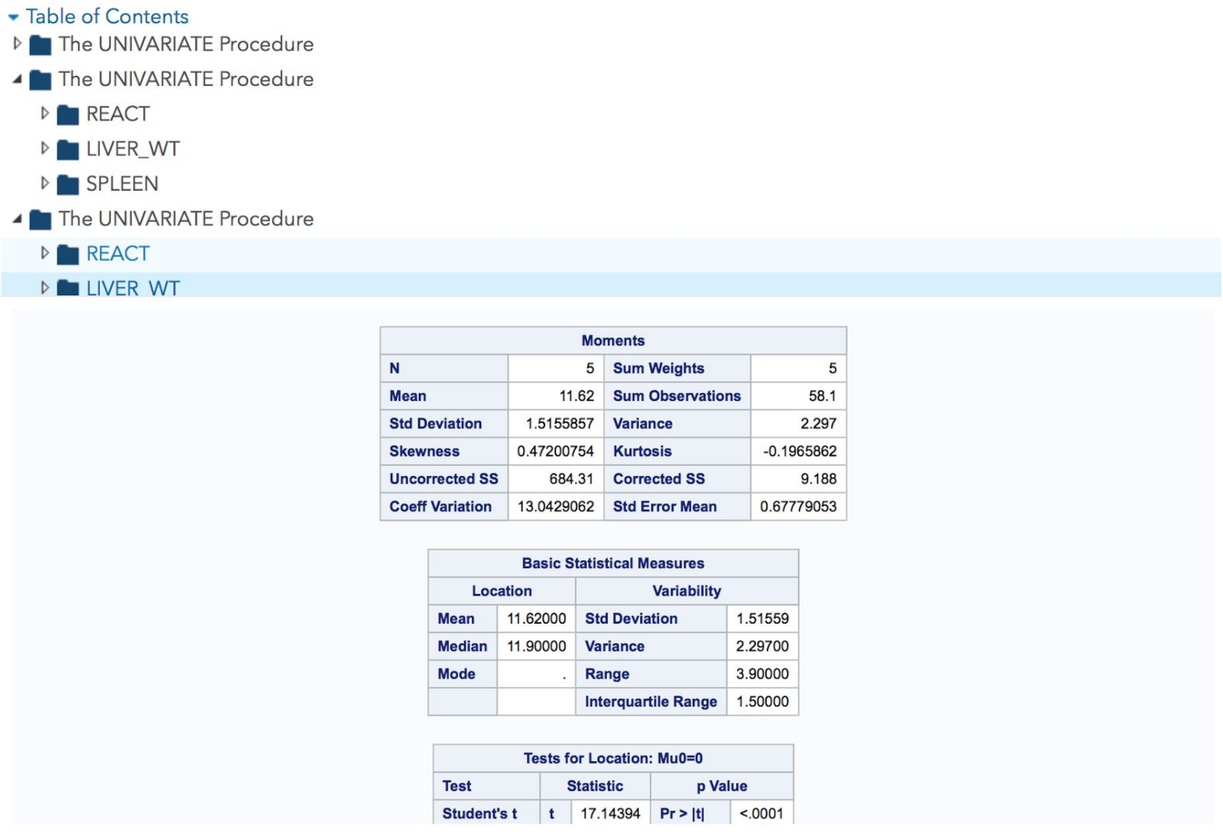
```
DATALINES;
1    1   5.4   10.2   8.9
2    1   5.9    9.8   7.3
3    1   4.8   12.2   9.1
4    1   6.9   11.8   8.8
5    1  15.8   10.9   9.0
6    2   4.9   13.8   6.6
7    2   5.0   12.0   7.9
8    2   6.7   10.5   8.0
9    2  18.2   11.9   6.9
10   2   5.5    9.9   9.1
;
```



```
1    1   5.4   10.2   8.9
2    1   5.9    9.8   7.3
3    1   4.8   12.2   9.1
4    1   6.9   11.8   8.8
5    1  15.8   10.9   9.0
6    2   4.9   13.8   6.6
7    2   5.0   12.0   7.9
8    2   6.7   10.5   8.0
9    2  18.2   11.9   6.9
10   2   5.5    9.9   9.1
;
run;

proc univariate data=liver NORMAL PLOT;
var REACT LIVER_WT SPLEEN;
HISTOGRAM;
run;

title "Analysis of 1 Dose";
proc univariate data=liver normal plot;
var REACT LIVER_WT SPLEEN;
WHERE DOSE=1;
HISTOGRAM;
RUN;

title "Analysis of 2 Dose";
proc univariate data=liver normal plot;
var REACT LIVER_WT SPLEEN;
WHERE DOSE=2;
HISTOGRAM;
RUN;
```



```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
61
62         DATA LIVER;
63            INPUT SUBJ DOSE REACT LIVER_WT SPLEEN;
64         DATALINES;

NOTE: The data set WORK.LIVER has 10 observations and 5 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds

75         ;

76         run;
77
78         proc univariate data=liver NORMAL PLOT;
79         var REACT LIVER_WT SPLEEN;
80         HISTOGRAM;
81         run;

NOTE: PROCEDURE UNIVARIATE used (Total process time):
      real time           6.29 seconds
      cpu time            0.80 seconds
```



Moments			
N	5	Sum Weights	5
Mean	11.62	Sum Observations	58.1
Std Deviation	1.5155857	Variance	2.297
Skewness	0.47200754	Kurtosis	-0.1965862
Uncorrected SS	684.31	Corrected SS	9.188
Coeff Variation	13.0429062	Std Error Mean	0.67779053

Basic Statistical Measures			
Location		Variability	
Mean	11.62000	Std Deviation	1.51559
Median	11.90000	Variance	2.29700
Mode	.	Range	3.90000
		Interquartile Range	1.50000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	17.14394	Pr > |t|	<.0001

5. What’s wrong with this program?

```
1    DATA 123;
2       INPUT AGE STATUS PROGNOSIS DOCTOR GENDER STATUS2
3             STATUS3;
4    (data lines)
       ;
5    PROC CHART DATA=123 BY GENDER;
6       VBAR STATUS
7       VBAR PROGNOSIS;
8    RUN;
9    PROC PLOT DATA=123;
10       DOCTOR BY PROGNOSIS;
11    RUN;
```

- THE DATA SET NAME IS INVALID, SHOULD NOT START WITH THE AGE VARIABLE OR IN MORE SIMPLY WITH THE NUMBER
- YOU DID NOT INCLUDED THE STATUS 1 BUT HAVE STATUS 2 AND 3
- I WOULD SUGGEST MAYBE ADDING SOME ID VARIABLE TO KEEP TRACK OF THINGS EASILY
- THE PROGRAM DO NOT HAVE A DATALINES AND CARDS STATEMENTS
- NO SEMICOLON AFTER PROC CHART
- ALSO IF YOU USE THE BY GENDER STATEMENT YOU SHOULD FIRST SORT IT BY GENDER BEFORE USING BY ON THE VARIABLE
- NO SEMICOLON AFTER VBAR STATUS
- IN THE LINE 10 I THINK THERE SHOULD BE A PLOT STATEMENT BEFORE ANYTHING IS WRITTEN;
- LASTLY WHEN YOU CONSTRUCT A PLOT YOU SHOULD HAVE ONE VARIABLE AGAINST ANOTHER NOT LIKE A DOCTOR BY … ; DOCTOR * PROGRANOSIS IS MORE APPROPRIATE