SAS Programming (BIOL-4V190)

Chapter 17

Counting Frequencies

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

17.1 Introduction

PROC FREQ provides frequency counts of variable values.
Like PROC MEANS, PROC FREQ can be used to create data sets containing these frequency counts and some statistics.

17.2 Counting Frequencies

By default, PROC FREQ generates frequency counts and percentages for all variables in the data set.

Syntax:

```
proc freq;
run;
```

17.3 Selecting Variables for PROC FREQ

Adding the TABLE (or TABLES) statement controls the variables which will appear in the output.
The NOCUM option on the TABLES statement suppresses the printing of the cumulative statistics.
The NOPERCENT option on the TABLES statement suppresses the printing of the percentages.

Syntax:

```
proc freq;
  tables variablename1 variablename2...variablename’n’ / < nocum nopercent >;
run;
```

This is illustrated in Program 17-2.

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

17.4 Using Formats to Label the Output

Using formatted variables or adding a format statement in the PROC FREQ procedure will generate output which displays the formatted values.

Syntax:

```
proc freq;
  tables variablename1 variablename2...variablename’n’;
  format variablename1 format. variablename2 $format.;
run;
```

This is illustrated in Program 17-3.

17.5 Using Formats to Group Values

Formats can also be used to recategorise or regroup the data based on the formatted values.

This is the same technique as was seen in PROC MEANS.

This is illustrated in Program 17-4.

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

17.6 Problems Grouping Values with PROC FREQ

When missing values are part of an ‘OTHER’ category created by PROC FORMAT, by default, all values falling into this category will be excluded from the PROC FREQ output.

The data set LEARN.GROUPING has one variable, X, and is shown on page 349.

A format is created to group the values of X in PROC FREQ output.

```
*Program 17-5 Demonstrating a problem in how PROC FREQ groups values - page 349;
proc format;
   value two
      low-3 = 'Group 1'
      4-5   = 'Group 2'
      other = 'Other values';
run;

title "Grouping Values (First Try)";
proc freq data=learn.grouping;
   tables X / nocum nopercent;
   format X two.;
run;
```

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

The output from the PROC FREQ procedure is shown on page 350.

The counts of 4 and 6 for Groups 1 and 2 are correct, but the count of 2 for the Missing Values is not correct. There is only 1 missing value in the data. The additional “missing” value is actually the value of X=6.

When the keyword ‘other’ is used in a format used by PROC FREQ, the procedure assigns all values in the category to the lowest value found in the data.

In this case, the smallest value is a missing value, so SAS groups all values in this category into the missing category.

The problem is solved by creating a specific category for missing values in PROC FORMAT.

```
*Program 17-6 Fixing the grouping problem - page 350;
proc format;
   value two
      low-3 = 'Group 1'
      4-5   = 'Group 2'
      .     = 'Missing'
      other = 'Other values';
run;
```

Rerunning the PROC FREQ code, the correct output is shown on the bottom of page 350.

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

17.7 Displaying Missing Values in the Frequency Table

Sometimes it is desirable to include missing values in the PROC FREQ output (and counts).

This can be accomplished by adding the MISSING option on the TABLES statement.

Syntax:

```
proc freq;
  tables variablename1 variablename2...variablename’n’ | missing;
run;
```

PROC FREQ results with and without the missing option are shown on page 352.

Notice the impact on the frequency counts and percentages when including/excluding the missing values.

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

17.8 Changing the Order of Values in PROC FREQ

By default, PROC FREQ orders the output based on the internal (underlying or raw) data values.

To change the order, use the ORDER= option on the PROC FREQ statement.

The values of ORDER are:

INTERNAL: internal or raw data values (default), smallest to largest

FORMATTED: formatted values

FREQ: frequency of values, largest to smallest

DATA: observation order in the data set – useful when working with data that may be sorted a certain way

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

In this example, a format is created to format the variable that will be used in PROC FREQ.

```
*Program 17-8 Demonstrating the ORDER= option of PROC FREQ - page 353;
proc format;
   value darwin
      1 = 'Yellow'
      2 = 'Blue'
      3 = 'Red'
      4 = 'Green'
      . = 'Missing';
run;

title "Default Order (Internal)";

proc freq data=test;
   tables Color / nocum nopercent missing;
   format Color darwin.;
run;
```

The output is shown on page 354.

Without the ORDER= option, the values are listed in ascending order of the unformatted or underlying data values.

Yellow is listed first since its data value is 1, Blue is next since its data value is 2, etc.

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

The display can be changed by adding the ORDER= option.

```
*Program 17-9 Demonstrating the ORDER= formatted, data, and freq options - page 354;
title "ORDER = formatted";
proc freq data=test order=formatted;
   tables Color / nocum nopercent;
   format Color darwin.;
run;

title "ORDER = data";
proc freq data=test order=data;
   tables Color / nocum nopercent;
   format Color darwin.;
run;

title "ORDER = freq";
proc freq data=test order=freq;
   tables Color / nocum nopercent;
   format Color darwin.;
run;
```

The output are shown on page 355.

When ORDER=FORMATTED, the values are listed in ascending order of the formatted values or Blue, Green, etc.

When ORDER=DATA, the data are listed in the order in which the values are found in the data.
The first four values in the data are 3 4 1 2, so the data are displayed as Red (3), Green (4), Yellow (1), Blue (2).

When ORDER=FREQ, the data are listed in descending order of frequency.
Red is the most frequently occurring value so it is listed first, followed by the next most frequently occurring value, Blue, then Yellow and Green.

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

17.9 Producing Two-Way Tables

To generate two-way tables, an asterisk is placed between two variable names on the TABLES statement.

Syntax:

```
proc freq;
  tables variablename1*variablename2;
run;
```

The output from this example is shown on the next slide.

```
*Program 17-10 Requesting a two-way table - page 356;
title "A Two-way Table of Gender by Blood Type";
proc freq data=learn.blood;
   tables Gender * BloodType;
run;
```

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

The values within each cell have been color coded to make it easier to understand.

```
A Two-way Table of Gender by Blood Type
                  The FREQ Procedure
             Table of Gender by BloodType
Gender(Gender)     BloodType(Blood Type)
Frequency|
Percent  |
Row Pct  |
Col Pct  |A       |AB      |B       |O       |  Total
---------+--------+--------+--------+--------+
Female   |    178 |     20 |     34 |    208 |    440
         |  17.80 |   2.00 |   3.40 |  20.80 |  44.00
         |  40.45 |   4.55 |   7.73 |  47.27 |
         |  43.20 |  45.45 |  35.42 |  46.43 |
---------+--------+--------+--------+--------+
Male     |    234 |     24 |     62 |    240 |    560
         |  23.40 |   2.40 |   6.20 |  24.00 |  56.00
         |  41.79 |   4.29 |  11.07 |  42.86 |
         |  56.80 |  54.55 |  64.58 |  53.57 |
---------+--------+--------+--------+--------+
Total         412       44       96      448     1000
            41.20     4.40     9.60    44.80   100.00
```

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

17.10 Requesting Multiple Two-Way Tables

Multiple two-way tables can be requested by using parentheses to group variables on the TABLES statement.

Multiple TABLES statements can also be used.

```
proc freq data=learn.blood;
   title 'generating multiple tables';
   tables Gender * (agegroup BloodType);
   tables agegroup*bloodtype;
run;
```

In this example, 3 tables will be generated:
Gender by Agegroup
Gender by BloodType
agegroup by bloodtype

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

17.11 Producing Three-Way Tables

To generate three-way tables, an asterisk is placed between the three variable names on the TABLES statement.

Multi-way table requests can generate ALOT of output.

It is sometimes helpful to use the LIST option on the TABLES statement to compress this output into a more compact table.

Syntax:

```
proc freq;
  tables variablename1*variablename2*variablename3 / list;
run;
```

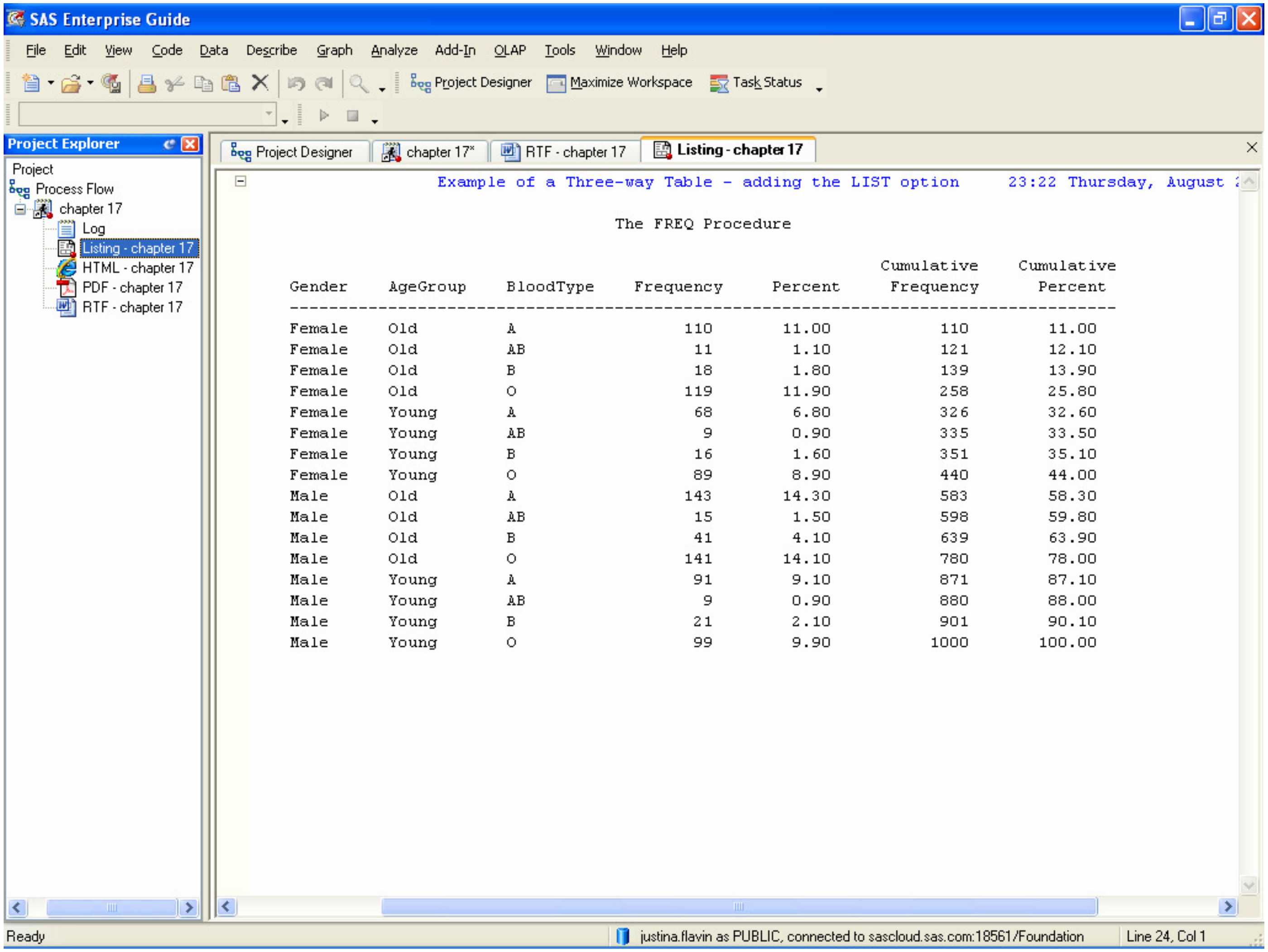
Note that not all possible combinations of all variables will necessarily be displayed in the output.

Only the combinations that actually occur in the data are shown.

```
title "Example of a Three-way Table - adding the LIST option";
proc freq data=learn.blood;
   tables Gender * AgeGroup * BloodType / list;
run;
```

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

Here is the output from the example.



```
                     Example of a Three-way Table - adding the LIST option     23:22 Thursday, August

                                          The FREQ Procedure

                                                                      Cumulative    Cumulative
   Gender     AgeGroup    BloodType    Frequency     Percent     Frequency      Percent
   ------------------------------------------------------------------------------------------
   Female     Old         A                  110       11.00           110        11.00
   Female     Old         AB                  11        1.10           121        12.10
   Female     Old         B                   18        1.80           139        13.90
   Female     Old         O                  119       11.90           258        25.80
   Female     Young       A                   68        6.80           326        32.60
   Female     Young       AB                   9        0.90           335        33.50
   Female     Young       B                   16        1.60           351        35.10
   Female     Young       O                   89        8.90           440        44.00
   Male       Old         A                  143       14.30           583        58.30
   Male       Old         AB                  15        1.50           598        59.80
   Male       Old         B                   41        4.10           639        63.90
   Male       Old         O                  141       14.10           780        78.00
   Male       Young       A                   91        9.10           871        87.10
   Male       Young       AB                   9        0.90           880        88.00
   Male       Young       B                   21        2.10           901        90.10
   Male       Young       O                   99        9.90          1000       100.00
```

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

Creating Frequency Data Sets Using PROC FREQ

Counts and frequencies generated by PROC FREQ can be routed to data sets by using an OUT= statement.

The data set contains the variables on the tables statement plus the variables COUNT and PERCENT.

Variables with cumulative counts and cumulative frequencies are not available.

Adding the NOPRINT option to the PROC FREQ statement will cause the output to be suppressed and only a data set will be generated.

Syntax:

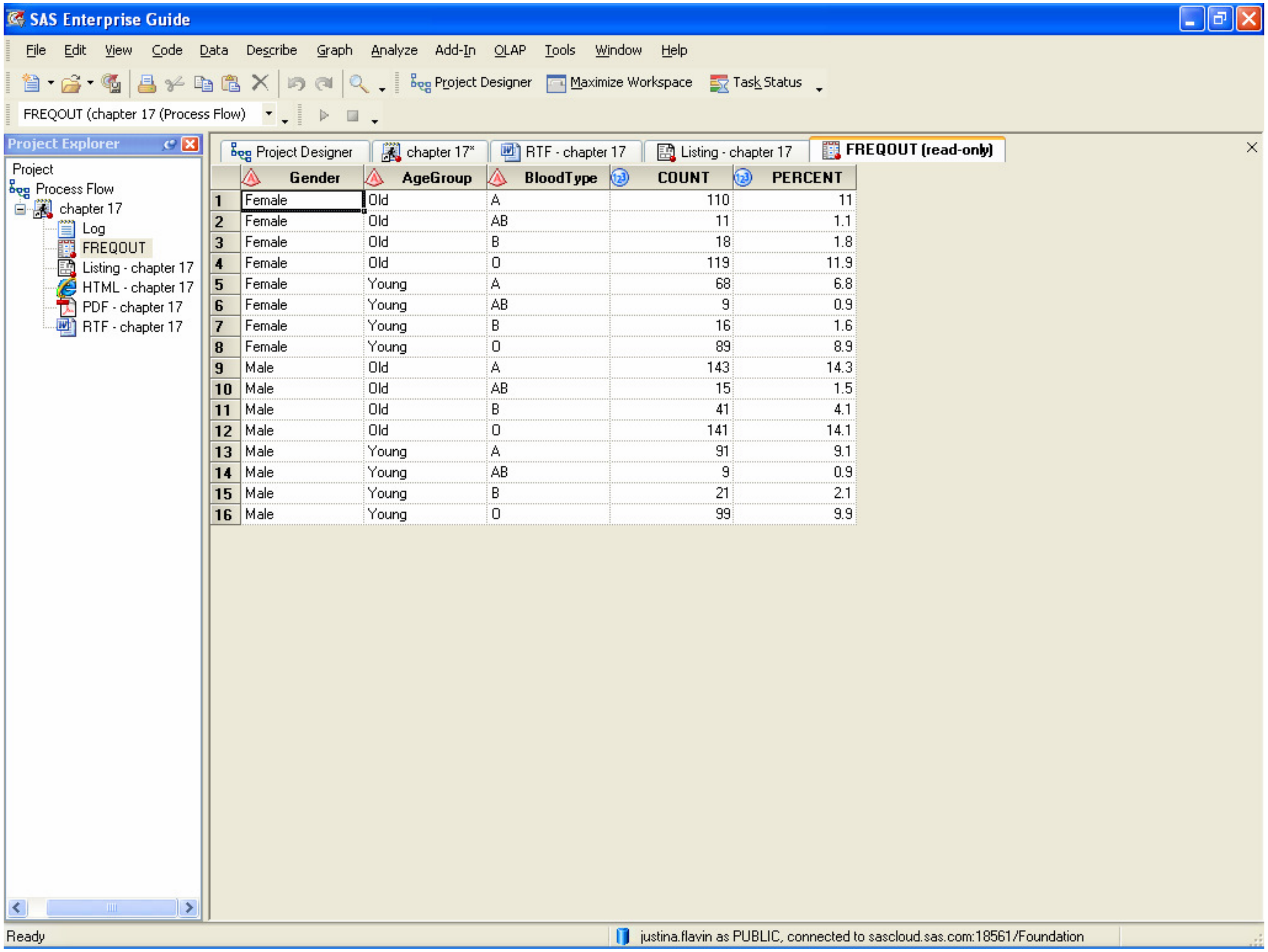
```
proc freq noprint;
  tables variablename1*variablename2 / out=datasetname;
run;
```

Here is an example:

```
proc freq data=learn.blood;
   tables Gender*AgeGroup*BloodType / list out=freqout;
run;
```

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

Here is the FREQOUT data set.



	Gender	AgeGroup	BloodType	COUNT	PERCENT
1	Female	Old	A	110	11
2	Female	Old	AB	11	1.1
3	Female	Old	B	18	1.8
4	Female	Old	O	119	11.9
5	Female	Young	A	68	6.8
6	Female	Young	AB	9	0.9
7	Female	Young	B	16	1.6
8	Female	Young	O	89	8.9
9	Male	Old	A	143	14.3
10	Male	Old	AB	15	1.5
11	Male	Old	B	41	4.1
12	Male	Old	O	141	14.1
13	Male	Young	A	91	9.1
14	Male	Young	AB	9	0.9
15	Male	Young	B	21	2.1
16	Male	Young	O	99	9.9

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

*****Performing a Chi-Square Analysis Using PROC FREQ*****

Chi-Square statistics can be generated by PROC FREQ by specifying the appropriate statistics keywords on the TABLES statement.

CELLCHI2 – individual cell chi-square values

CHISQ – overall chi-square value

EXPECTED – expected cell counts

Syntax:

```
proc freq;
  tables variablename1*variablename2 / chisq cellchi2 expected;
run;
```

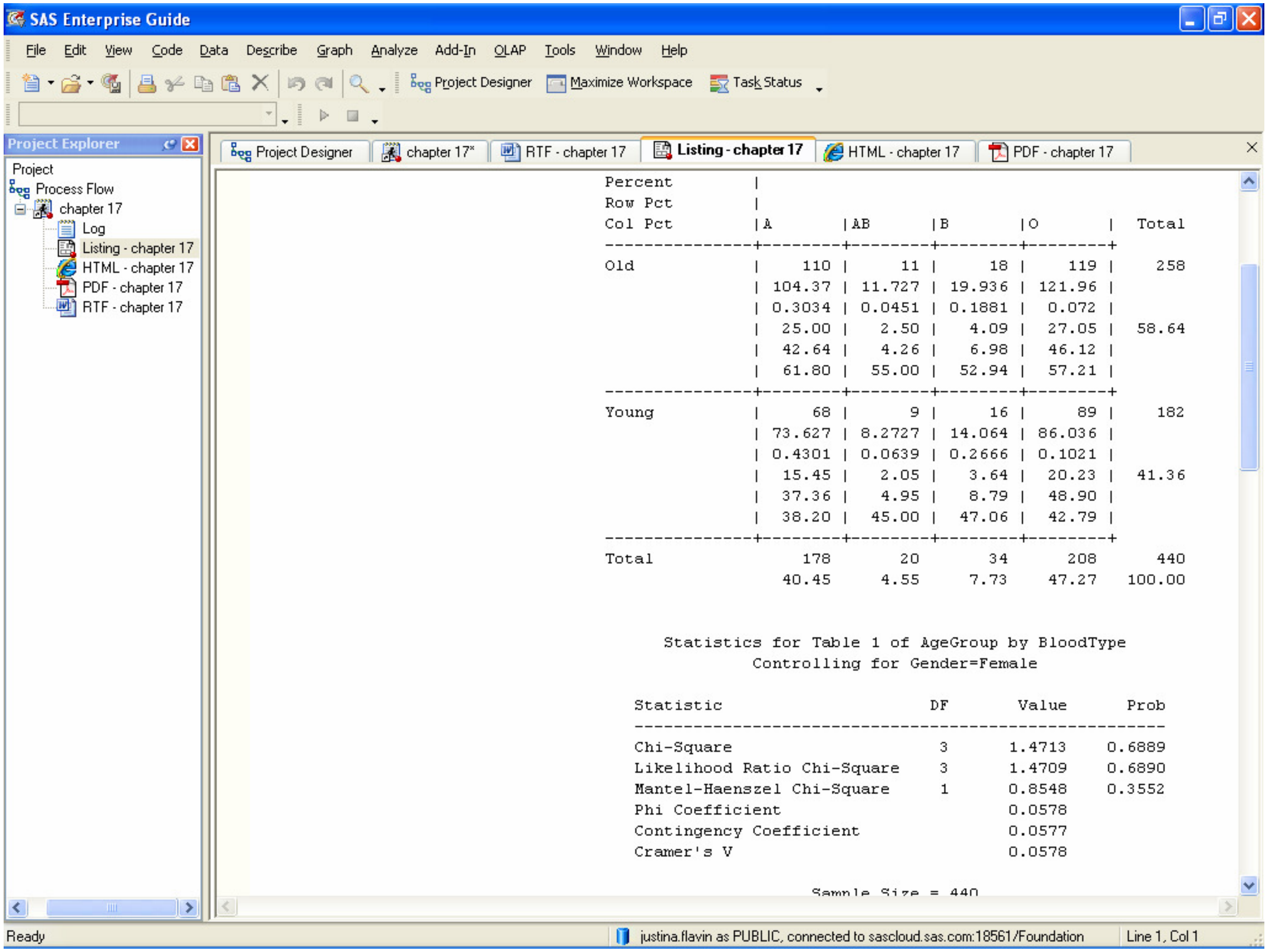
Additional information on the statistics options available in PROC FREQ may be found in the online documentation.

Here is an example:

```
proc freq data=learn.blood;
  tables Gender*AgeGroup*BloodType / chisq cellchi2 expected cmh;
run;
```

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

When statistics options are added to the PROC FREQ code, an additional table with the statistical results is displayed underneath the table of frequency counts.



```
Percent         |
Row Pct         |
Col Pct         |A       |AB      |B       |O       |  Total
----------------+--------+--------+--------+--------+
Old             |    110 |     11 |     18 |    119 |    258
                | 104.37 | 11.727 | 19.936 | 121.96 |
                | 0.3034 | 0.0451 | 0.1881 |  0.072 |
                |  25.00 |   2.50 |   4.09 |  27.05 |  58.64
                |  42.64 |   4.26 |   6.98 |  46.12 |
                |  61.80 |  55.00 |  52.94 |  57.21 |
----------------+--------+--------+--------+--------+
Young           |     68 |      9 |     16 |     89 |    182
                | 73.627 | 8.2727 | 14.064 | 86.036 |
                | 0.4301 | 0.0639 | 0.2666 | 0.1021 |
                |  15.45 |   2.05 |   3.64 |  20.23 |  41.36
                |  37.36 |   4.95 |   8.79 |  48.90 |
                |  38.20 |  45.00 |  47.06 |  42.79 |
----------------+--------+--------+--------+--------+
Total                178       20       34      208      440
                   40.45     4.55     7.73    47.27   100.00


      Statistics for Table 1 of AgeGroup by BloodType
              Controlling for Gender=Female

   Statistic                     DF       Value      Prob
   ------------------------------------------------------
   Chi-Square                     3      1.4713    0.6889
   Likelihood Ratio Chi-Square    3      1.4709    0.6890
   Mantel-Haenszel Chi-Square     1      0.8548    0.3552
   Phi Coefficient                       0.0578
   Contingency Coefficient               0.0577
   Cramer's V                            0.0578

                     Sample Size = 440
```

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

*****Creating Data Sets Containing Statistics*****

To create data sets containing statistics generated by PROC FREQ, an OUTPUT statement must be added.

The data set will contain the statistics specified on the OUTPUT line.

Frequency count data from the TABLES statement is not included, but as previously shown, this data may be output to a data set by using the OUT=option on the TABLES statement.

Syntax:

```
proc freq;
  tables variablename1*variablename2 / statistics-keywords;
  output out=datasetname statistics-keywords;
run;
```

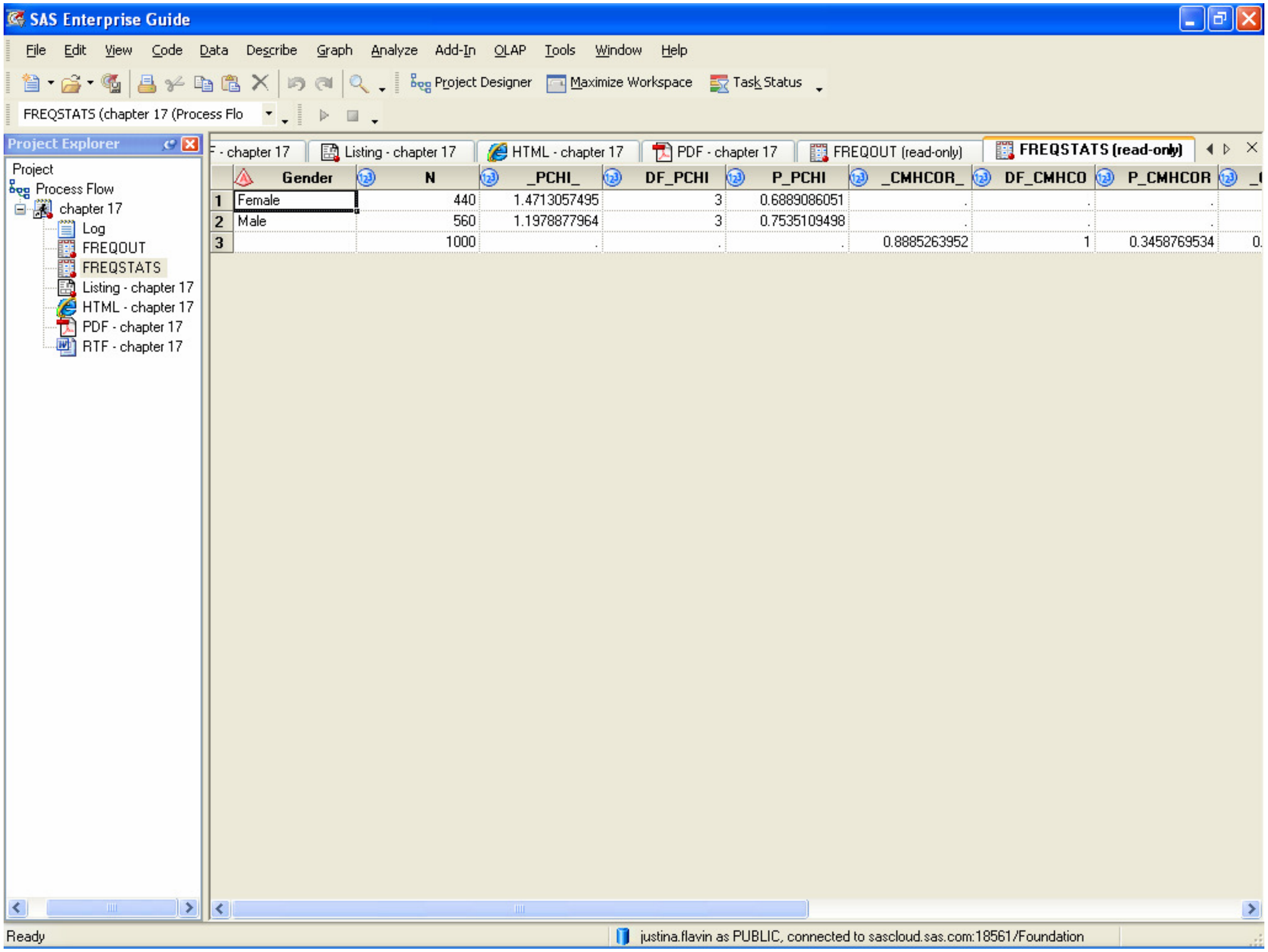
Additional information about the contents, structure, and variable naming conventions of the output data set are available in the online documentation under the OUTPUT statement and OUTPUT Data Sets.

This example illustrates outputting two data sets.
FREQOUT contains the information from the TABLES statements.
FREQSTATS contains the statistical results.

```
proc freq data=learn.blood;
   tables Gender*AgeGroup*BloodType / chisq cellchi2 expected cmh out=freqout;
   output out=freqstats cmh pchi;
run;
```

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

The FREQSTATS data set is shown below. The statistics written out to the data set must be specified on the OUTPUT statement. Different statistics can be displayed on the output generated by the TABLES statement.



	Gender	N	_PCHI_	DF_PCHI	P_PCHI	_CMHCOR_	DF_CMHCO	P_CMHCOR	_(
1	Female	440	1.4713057495	3	0.6889086051	.	.	.	
2	Male	560	1.1978877964	3	0.7535109498	.	.	.	
3		1000	.	.	.	0.8885263952	1	0.3458769534	0.

Ready   justina.flavin as PUBLIC, connected to sascloud.sas.com:18561/Foundation

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

Additional Topic: PROC UNIVARIATE

PROC UNIVARIATE provides descriptive univariate statistics on numeric variables.

The syntax is very similar to that of PROC FREQ and PROC MEANS and like those procedures, output data sets can also be produced.

Crude data plots can also be obtained from PROC UNIVARIATE.

Basic PROC UNIVARIATE

By default, PROC UNIVARIATE generates univariate statistics output for all numeric variables in the data set.

Syntax:

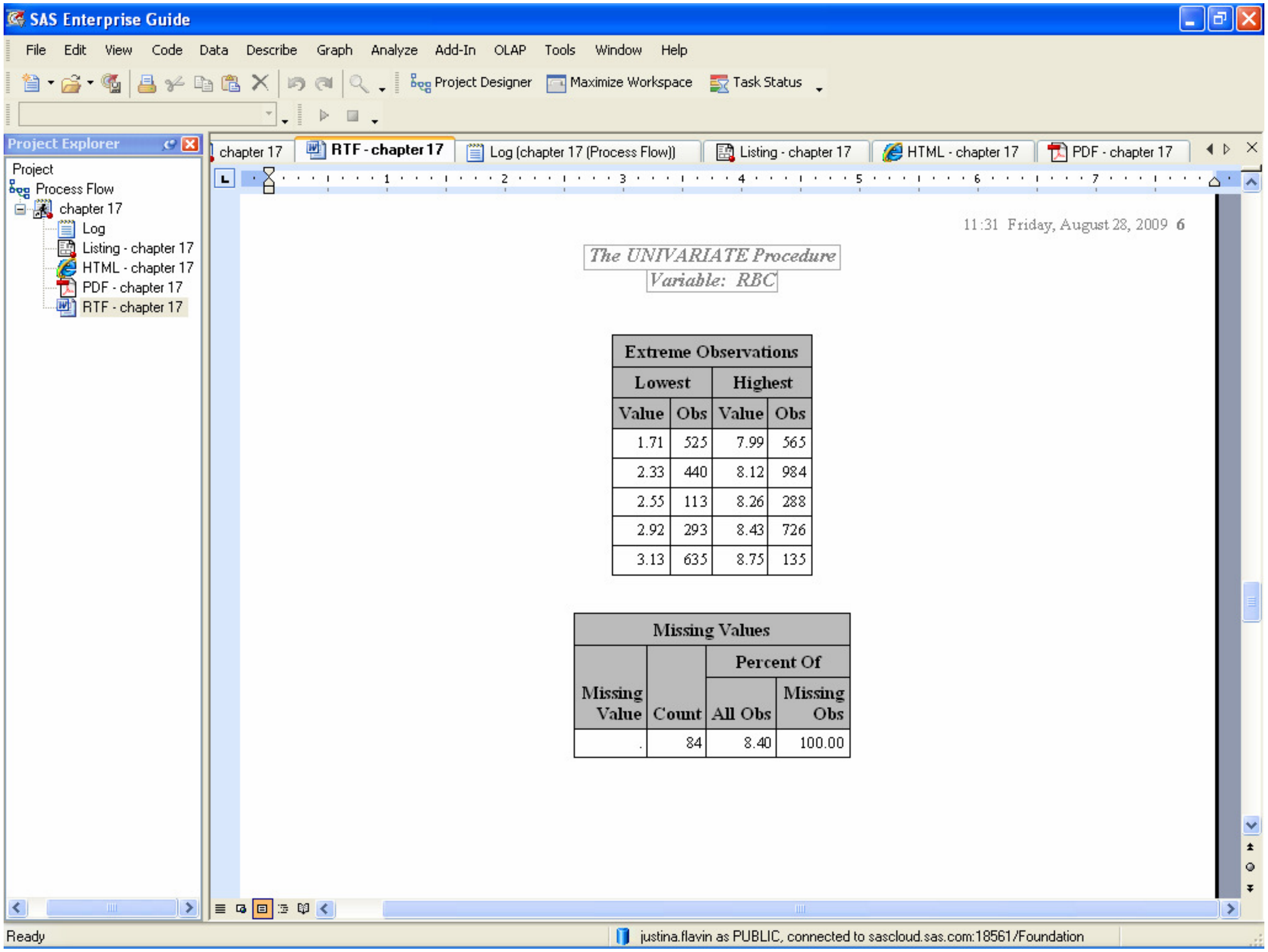
```
proc univariate;
run;
```

Example:

```
proc univariate data=learn.blood;
run;
```

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

PROC UNIVARIATE generates quite a bit of output. The list of Extreme Observations for RBC is shown below. These are the 5 lowest and 5 highest values of RBC, identified by the Observation Number.



11:31 Friday, August 28, 2009 6

The UNIVARIATE Procedure
Variable: RBC

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1.71	525	7.99	565
2.33	440	8.12	984
2.55	113	8.26	288
2.92	293	8.43	726
3.13	635	8.75	135

Missing Values			
		Percent Of	
Missing Value	Count	All Obs	Missing Obs
.	84	8.40	100.00

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

Use the VAR Statement to Select the Variables

Adding an ID statement causes the value of the ID variable(s) to be added to the list of Extreme Observations to better identify these values.

Syntax:

```
proc univariate;
  id variablename1 variablename2...variablename’n’;
  var variablename1 variablename2...variablename’n’;
run;
```

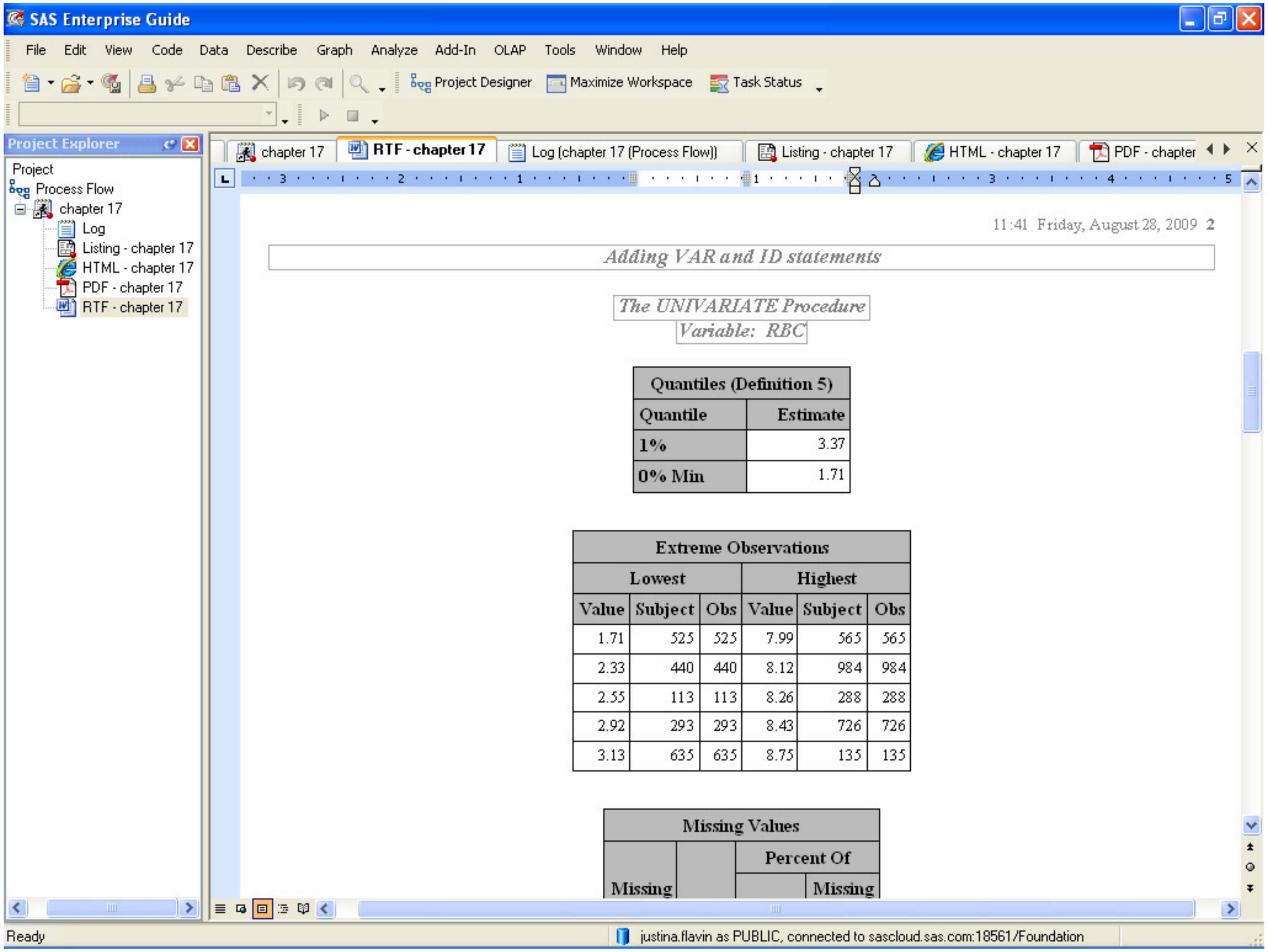
Example:

```
proc univariate data=learn.blood;
  title Adding VAR and ID statements;
  var rbc wbc;
  id subject;
run;
```

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

In the example code, the statement ID subject was added.

Now the Extreme Observations list also includes the value of Subject for each observation listed.



11:41 Friday, August 28, 2009 2

Adding VAR and ID statements

The UNIVARIATE Procedure
Variable: RBC

Quantiles (Definition 5)	
Quantile	**Estimate**
1%	3.37
0% Min	1.71

Extreme Observations					
Lowest			**Highest**		
Value	**Subject**	**Obs**	**Value**	**Subject**	**Obs**
1.71	525	525	7.99	565	565
2.33	440	440	8.12	984	984
2.55	113	113	8.26	288	288
2.92	293	293	8.43	726	726
3.13	635	635	8.75	135	135

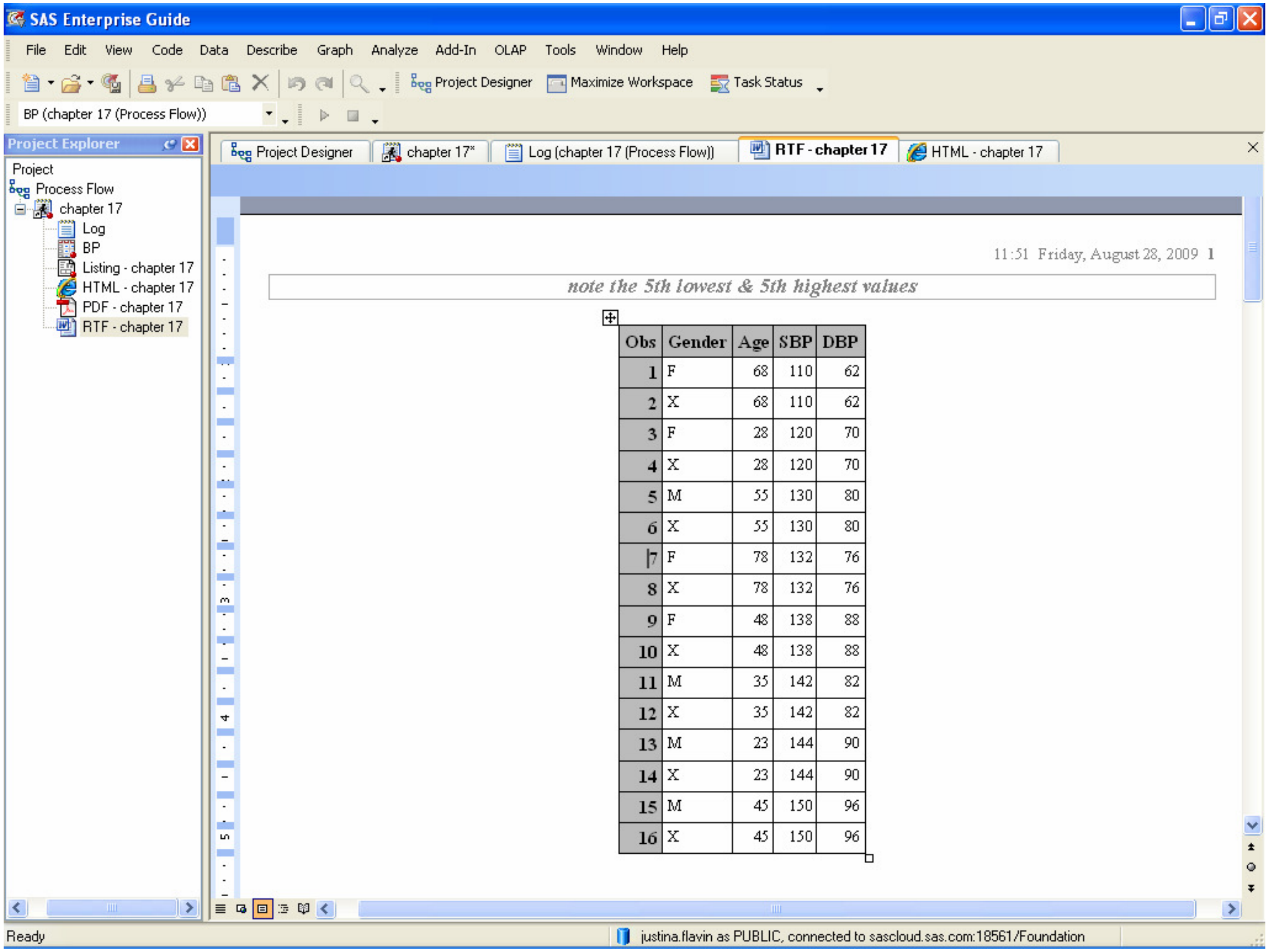
Missing Values			
		Percent Of	
Missing			**Missing**

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

One caveat to using the list of Extreme Observations is that there may be more observations in the data set with these highest & lowest values, but SAS only prints 5 of them.

In the example code, we create a data set that will have duplicate values of the analysis variable.

Here is a list of the data sorted by ascending order of SBP. Note the 5 lowest and 5 highest values.



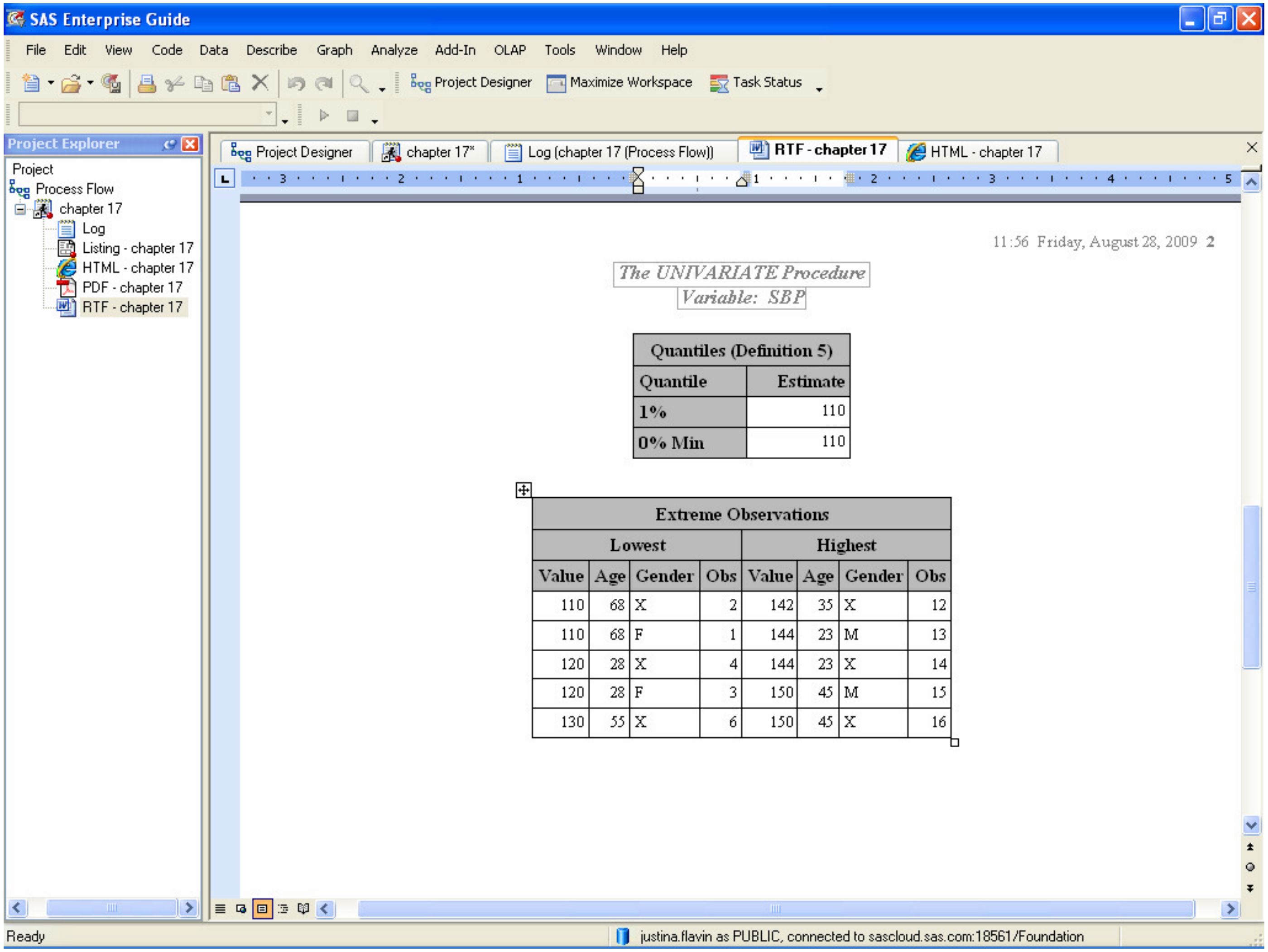
11:51 Friday, August 28, 2009 1

note the 5th lowest & 5th highest values

Obs	Gender	Age	SBP	DBP
1	F	68	110	62
2	X	68	110	62
3	F	28	120	70
4	X	28	120	70
5	M	55	130	80
6	X	55	130	80
7	F	78	132	76
8	X	78	132	76
9	F	48	138	88
10	X	48	138	88
11	M	35	142	82
12	X	35	142	82
13	M	23	144	90
14	X	23	144	90
15	M	45	150	96
16	X	45	150	96

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

Notice that the 5th lowest value of 130 is for AGE=55, GENDER=X. From the listing we know that there is also an observation of 130 for AGE=55, GENDER=M. The same is true for the 5th highest value. There is another observation for the value of 142 which doesn't get listed. Once SAS has identified 5 values, it does not expand the list to accommodate duplicates.



11:56 Friday, August 28, 2009 2

The UNIVARIATE Procedure

Variable: SBP

Quantiles (Definition 5)	
Quantile	**Estimate**
1%	110
0% Min	110

Extreme Observations							
Lowest				Highest			
Value	Age	Gender	Obs	Value	Age	Gender	Obs
110	68	X	2	142	35	X	12
110	68	F	1	144	23	M	13
120	28	X	4	144	23	X	14
120	28	F	3	150	45	M	15
130	55	X	6	150	45	X	16

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

Adding a BY or CLASS statement

A BY or CLASS statement can be added to obtain univariate statistics for subgroups.

When a BY statement is used, the data must first be sorted.

Syntax:

```
proc univariate;
  by variablename1 variablename2...variablename’n’;
  var variablename1 variablename2...variablename’n’;
run;

proc univariate;
  class variablename1 variablename2...variablename’n’;
  var variablename1 variablename2...variablename’n’;
run;
```

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

Other Options

FREQ – generates a frequency table of all the variable values (similar to PROC FREQ output)

PLOT - produces a stem-and-leaf plot, box plot, and normal probability plot

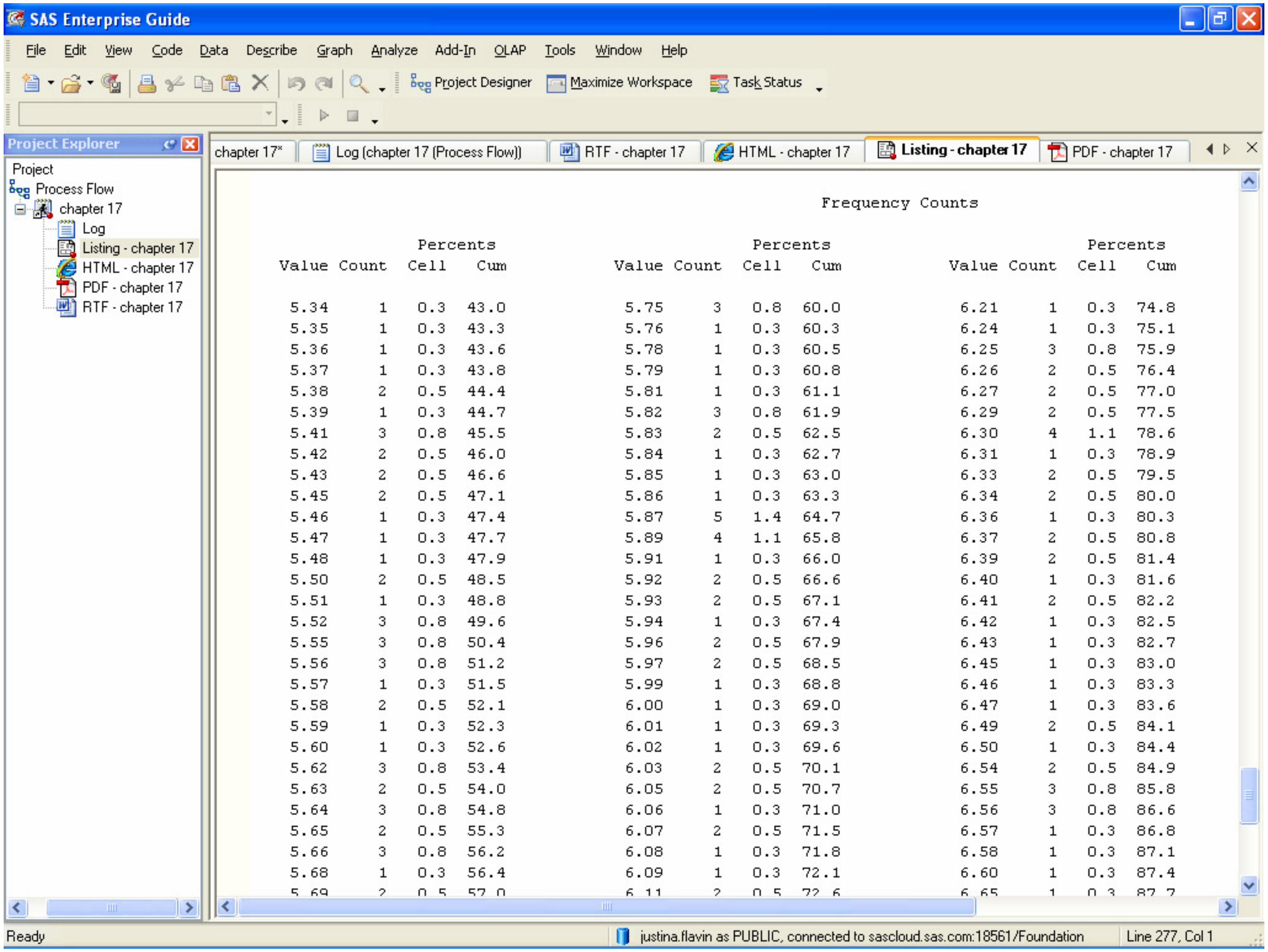
NORMAL – provides test for normality statistics

Syntax:

```
proc univariate freq plot normal;
   var variablename1 variablename2...variablename’n’;
run;
```

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

Here is a table of frequency counts produced by adding the FREQ option.



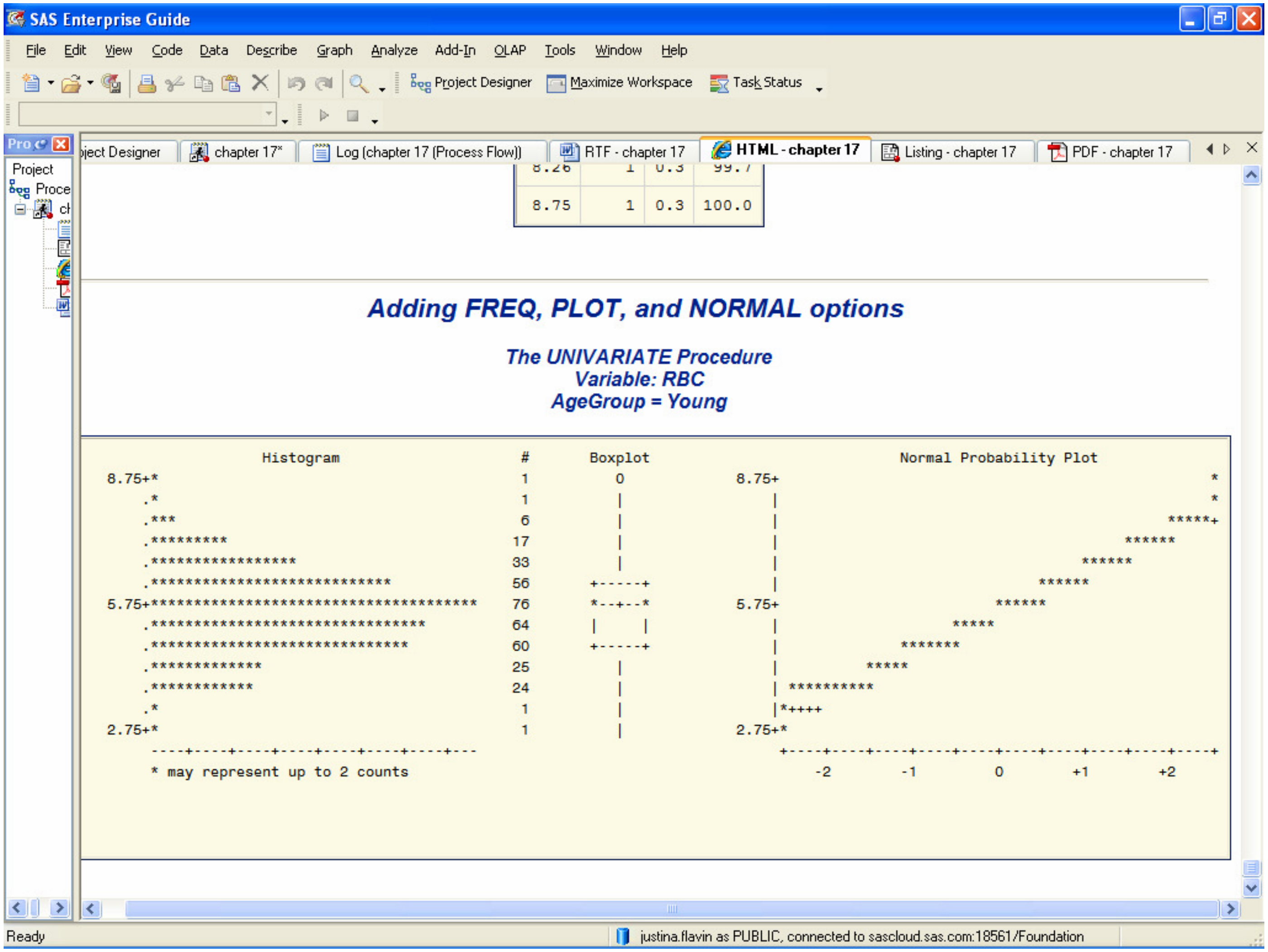
```
                                   Frequency Counts

                 Percents                          Percents                          Percents
 Value  Count   Cell   Cum        Value  Count   Cell   Cum        Value  Count   Cell   Cum

  5.34      1    0.3   43.0        5.75      3    0.8   60.0        6.21      1    0.3   74.8
  5.35      1    0.3   43.3        5.76      1    0.3   60.3        6.24      1    0.3   75.1
  5.36      1    0.3   43.6        5.78      1    0.3   60.5        6.25      3    0.8   75.9
  5.37      1    0.3   43.8        5.79      1    0.3   60.8        6.26      2    0.5   76.4
  5.38      2    0.5   44.4        5.81      1    0.3   61.1        6.27      2    0.5   77.0
  5.39      1    0.3   44.7        5.82      3    0.8   61.9        6.29      2    0.5   77.5
  5.41      3    0.8   45.5        5.83      2    0.5   62.5        6.30      4    1.1   78.6
  5.42      2    0.5   46.0        5.84      1    0.3   62.7        6.31      1    0.3   78.9
  5.43      2    0.5   46.6        5.85      1    0.3   63.0        6.33      2    0.5   79.5
  5.45      2    0.5   47.1        5.86      1    0.3   63.3        6.34      2    0.5   80.0
  5.46      1    0.3   47.4        5.87      5    1.4   64.7        6.36      1    0.3   80.3
  5.47      1    0.3   47.7        5.89      4    1.1   65.8        6.37      2    0.5   80.8
  5.48      1    0.3   47.9        5.91      1    0.3   66.0        6.39      2    0.5   81.4
  5.50      2    0.5   48.5        5.92      2    0.5   66.6        6.40      1    0.3   81.6
  5.51      1    0.3   48.8        5.93      2    0.5   67.1        6.41      2    0.5   82.2
  5.52      3    0.8   49.6        5.94      1    0.3   67.4        6.42      1    0.3   82.5
  5.55      3    0.8   50.4        5.96      2    0.5   67.9        6.43      1    0.3   82.7
  5.56      3    0.8   51.2        5.97      2    0.5   68.5        6.45      1    0.3   83.0
  5.57      1    0.3   51.5        5.99      1    0.3   68.8        6.46      1    0.3   83.3
  5.58      2    0.5   52.1        6.00      1    0.3   69.0        6.47      1    0.3   83.6
  5.59      1    0.3   52.3        6.01      1    0.3   69.3        6.49      2    0.5   84.1
  5.60      1    0.3   52.6        6.02      1    0.3   69.6        6.50      1    0.3   84.4
  5.62      3    0.8   53.4        6.03      2    0.5   70.1        6.54      2    0.5   84.9
  5.63      2    0.5   54.0        6.05      2    0.5   70.7        6.55      3    0.8   85.8
  5.64      3    0.8   54.8        6.06      1    0.3   71.0        6.56      3    0.8   86.6
  5.65      2    0.5   55.3        6.07      2    0.5   71.5        6.57      1    0.3   86.8
  5.66      3    0.8   56.2        6.08      1    0.3   71.8        6.58      1    0.3   87.1
  5.68      1    0.3   56.4        6.09      1    0.3   72.1        6.60      1    0.3   87.4
  5.69      2    0.5   57.0        6.11      2    0.5   72.6        6.65      1    0.3   87.7
```

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

Here are the plots produced by adding the PLOT option -stem-and-leaf plot, box plot, and normal probability plot



8.26	1	0.3	99.7
8.75	1	0.3	100.0

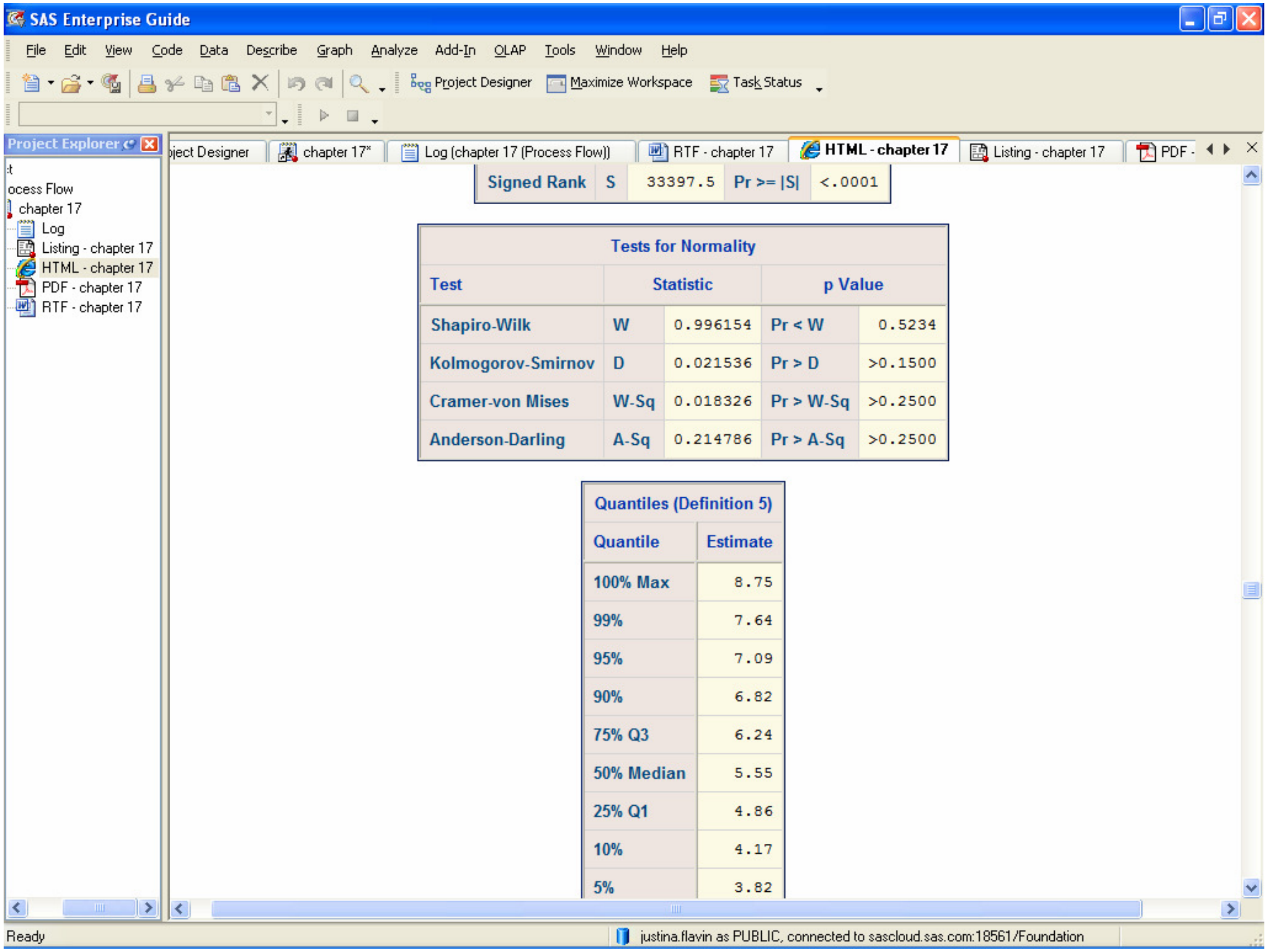
Adding FREQ, PLOT, and NORMAL options

The UNIVARIATE Procedure
Variable: RBC
AgeGroup = Young

```
                  Histogram                     #       Boxplot                             Normal Probability Plot
8.75+*                                          1          0             8.75+                                                  *
    .*                                          1          |                 |                                                  *
    .***                                        6          |                 |                                             *****+
    .*********                                 17          |                 |                                        ******
    .*****************                         33          |                 |                                   ******
    .****************************              56       +-----+              |                              ******
5.75+**************************************    76       *--+--*          5.75+                         ******
    .********************************          64       |     |              |                    *****
    .******************************            60       +-----+              |              *******
    .*************                             25          |                 |          *****
    .************                              24          |                 | **********
    .*                                          1          |                 |*++++
2.75+*                                          1          |             2.75+*
     ----+----+----+----+----+----+----+---                                   +----+----+----+----+----+----+----+----+----+----+
     * may represent up to 2 counts                                               -2        -1         0        +1        +2
```

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

Adding the NORMAL option adds the test for normality statistics


SAS Enterprise Guide
File Edit View Code Data Describe Graph Analyze Add-In OLAP Tools Window Help
Project Designer Maximize Workspace Task Status
Project Explorer
:t
ocess Flow
chapter 17
Log
Listing - chapter 17
HTML - chapter 17
PDF - chapter 17
RTF - chapter 17
ject Designer
chapter 17*
Log (chapter 17 (Process Flow))
RTF - chapter 17
HTML - chapter 17
Listing - chapter 17
PDF -
Ready
justina.flavin as PUBLIC, connected to sascloud.sas.com:18561/Foundation

Signed Rank	S	33397.5	Pr >= |S|	<.0001

Tests for Normality

Test	Statistic		p Value	
Shapiro-Wilk	W	0.996154	Pr < W	0.5234
Kolmogorov-Smirnov	D	0.021536	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.018326	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.214786	Pr > A-Sq	>0.2500

Quantiles (Definition 5)

Quantile	Estimate
100% Max	8.75
99%	7.64
95%	7.09
90%	6.82
75% Q3	6.24
50% Median	5.55
25% Q1	4.86
10%	4.17
5%	3.82

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

Adding Options to Generate Output Data Sets

The procedure and conventions for routing statistics to an output data set is the same as for PROC MEANS.

Syntax:

```
proc univariate noprint;
   var variablename1 variablename2...variablename’n’;
   output out=datasetname
         statistics-keyword1 = variablename1 variablename2...variablename’n’
         statistics-keyword2 = variablename1 variablename2...variablename’n’
         statistics-keyword’n’ = variablename1 variablename2...variablename’n’;
run;
```

Example:

```
proc univariate data=learn.blood noprint;
       class Gender AgeGroup;
       var rbc wbc chol;
       output out = summary
              mean = m_rbc m_wbc m_chol
              median = goat pig;
run;
```

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

10.6 Combining Detail and Summary Data

Data sets created from PROC MEANS, FREQ, and UNIVARIATE are often merged back onto the data sets from which they were derived to enable additional processing.

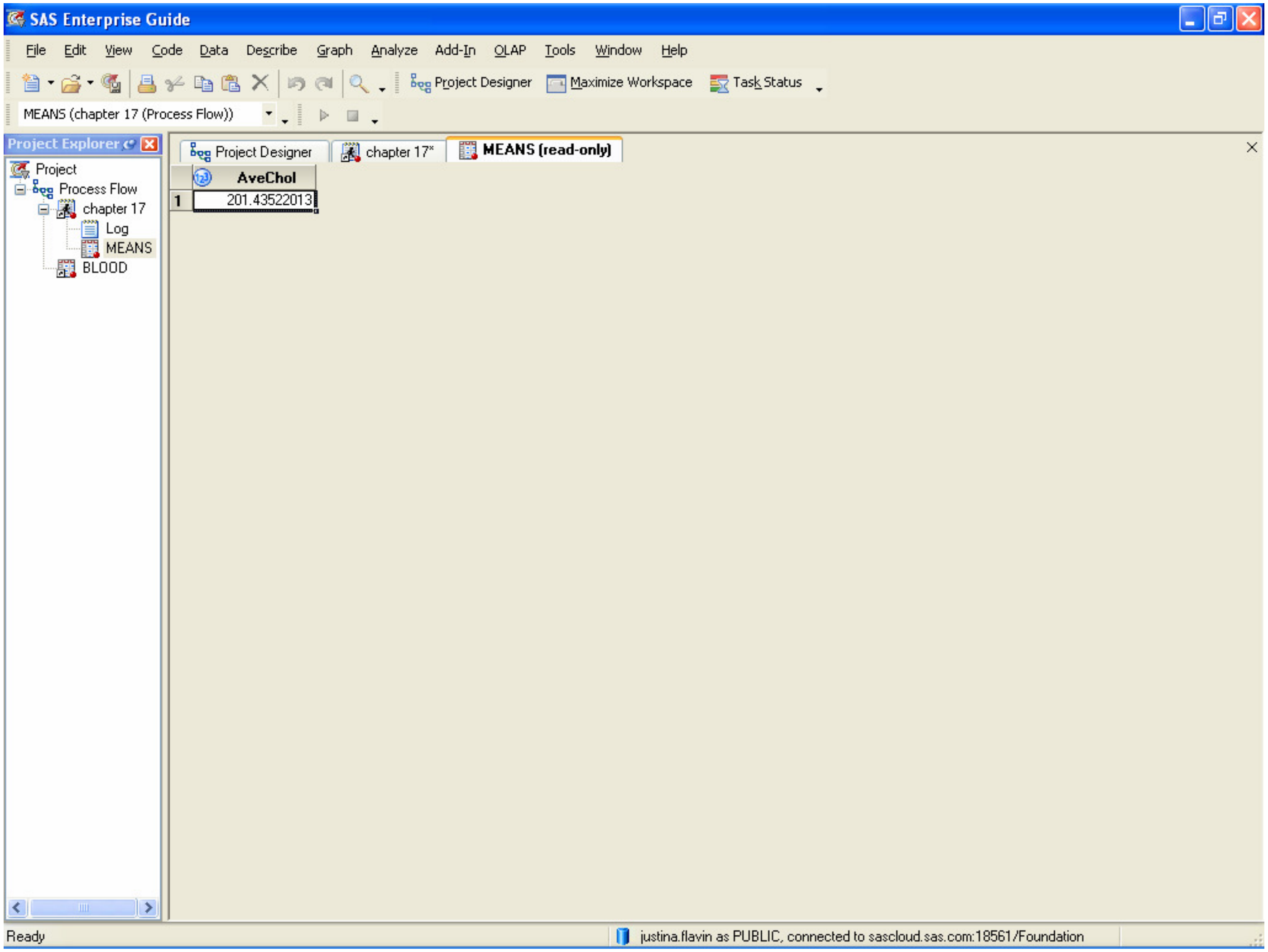
We will now look at Program 10-7.

```
*Program 10-7 Combining detail and summary data: Conditional SET statement - page 168;
proc means data=learn.blood noprint;
   var Chol;
   output out = means(keep=AveChol)
          mean = AveChol;
run;

data percent;
   set learn.blood(keep=Subject Chol);
   if _n_ = 1 then set means;
   PerChol = Chol / AveChol;
   format PerChol percent8.;
run;
```

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

First, PROC MEANS is used to create a data set that contains one observation with one variable, the mean value of Cholesterol.



SAS Enterprise Guide
File Edit View Code Data Describe Graph Analyze Add-In OLAP Tools Window Help
Project Designer
Maximize Workspace
Task Status
MEANS (chapter 17 (Process Flow))
Project Explorer
Project
Process Flow
chapter 17
Log
MEANS
BLOOD
Project Designer
chapter 17*
MEANS (read-only)
Ready
justina.flavin as PUBLIC, connected to sascloud.sas.com:18561/Foundation

	AveChol
1	201.43522013

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

```
data percent;
   set learn.blood(keep=Subject Chol);
   if _n_ = 1 then set means;
   PerChol = Chol / AveChol;
   format PerChol percent8.;
run;
```

This data step illustrates the use of the automatic variable _n_ that counts iterations of the data step.

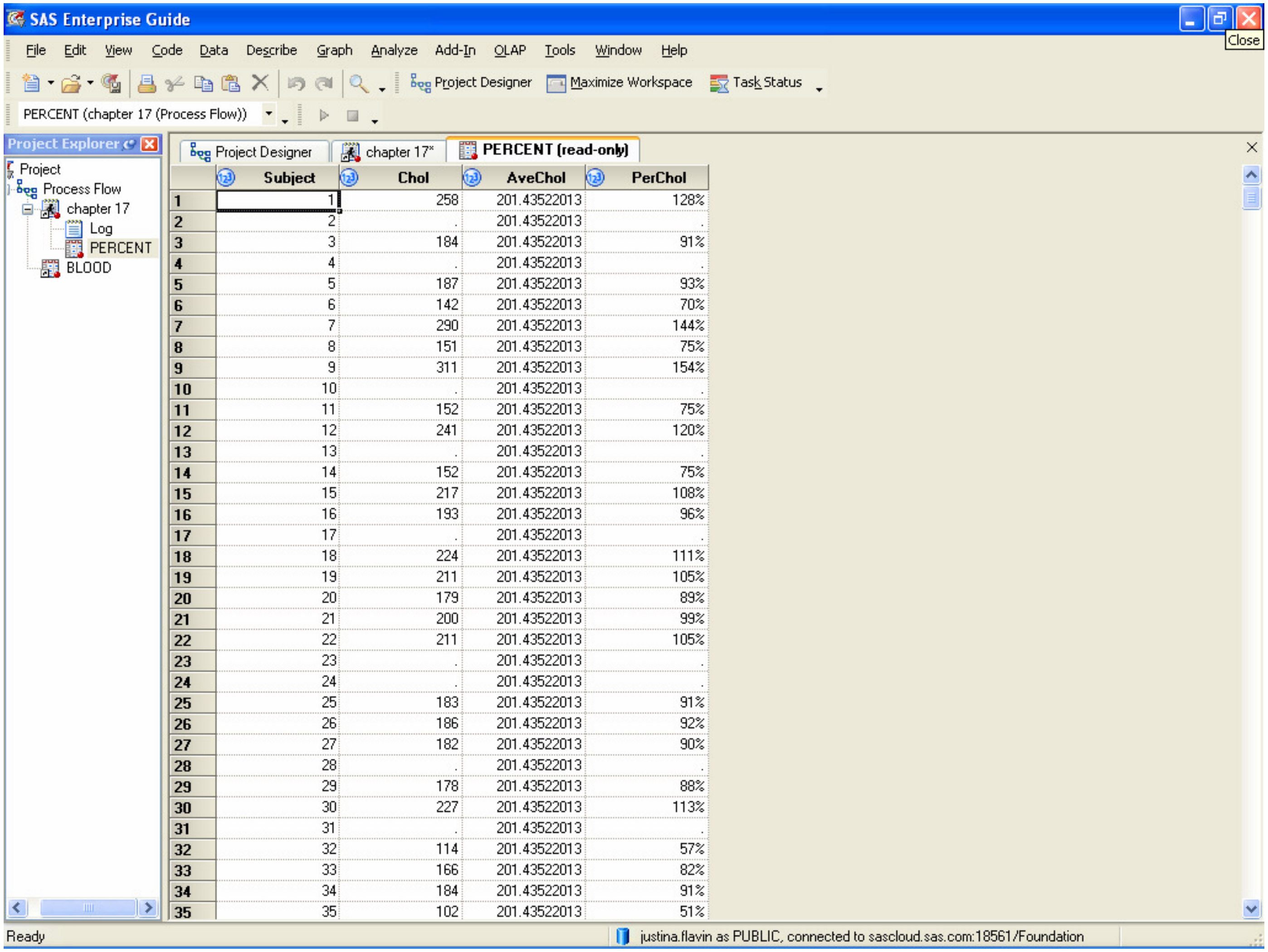
The first time SAS executes the data step code, _n_=1, so the MEANS data set is set into the data step and the variable AveChol is added.

MEANS is not added for any other iteration of the data step.

However, since variable values are retained, this has the effect of adding the value on every observation that is read in from LEARN.BLOOD.

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

Here is the data set PERCENT



	Subject	Chol	AveChol	PerChol
1	1	258	201.43522013	128%
2	2	.	201.43522013	.
3	3	184	201.43522013	91%
4	4	.	201.43522013	.
5	5	187	201.43522013	93%
6	6	142	201.43522013	70%
7	7	290	201.43522013	144%
8	8	151	201.43522013	75%
9	9	311	201.43522013	154%
10	10	.	201.43522013	.
11	11	152	201.43522013	75%
12	12	241	201.43522013	120%
13	13	.	201.43522013	.
14	14	152	201.43522013	75%
15	15	217	201.43522013	108%
16	16	193	201.43522013	96%
17	17	.	201.43522013	.
18	18	224	201.43522013	111%
19	19	211	201.43522013	105%
20	20	179	201.43522013	89%
21	21	200	201.43522013	99%
22	22	211	201.43522013	105%
23	23	.	201.43522013	.
24	24	.	201.43522013	.
25	25	183	201.43522013	91%
26	26	186	201.43522013	92%
27	27	182	201.43522013	90%
28	28	.	201.43522013	.
29	29	178	201.43522013	88%
30	30	227	201.43522013	113%
31	31	.	201.43522013	.
32	32	114	201.43522013	57%
33	33	166	201.43522013	82%
34	34	184	201.43522013	91%
35	35	102	201.43522013	51%

Ready

justina.flavin as PUBLIC, connected to sascloud.sas.com:18561/Foundation

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

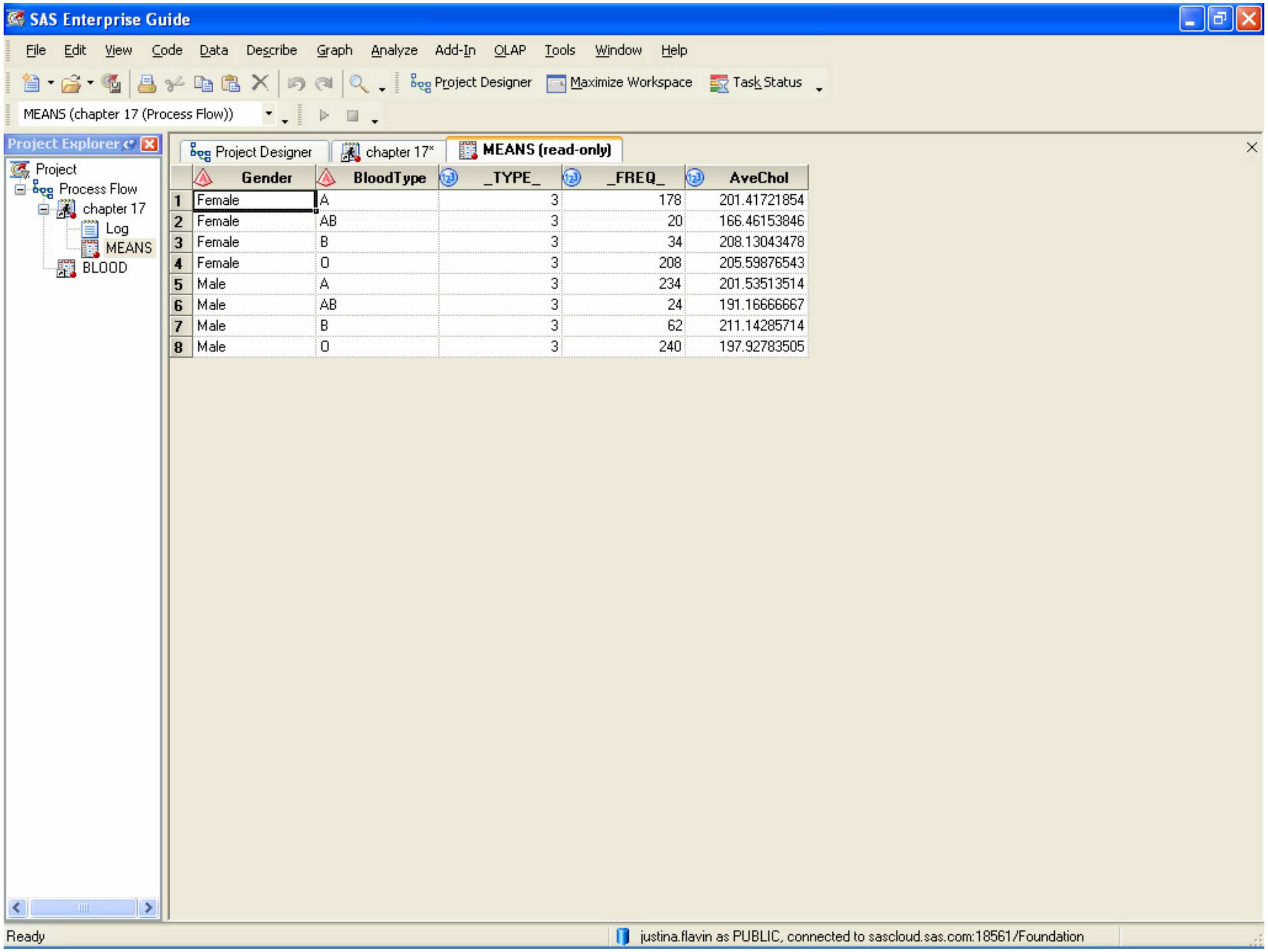
A merge can also be used to combine summary information back into a data set.

In this example, we calculate the mean value of cholesterol stratified by gender and bloodtype.

```
proc means data=learn.blood noprint nway;
   class gender bloodtype;
   var Chol;
   output out = means
          mean = AveChol;
run;
```

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

Here is the data set MEANS


SAS Enterprise Guide
File Edit View Code Data Describe Graph Analyze Add-In OLAP Tools Window Help
Project Designer
Maximize Workspace
Task Status
MEANS (chapter 17 (Process Flow))
Project Explorer
Project
Process Flow
chapter 17
Log
MEANS
BLOOD
Project Designer
chapter 17*
MEANS (read-only)
Ready
justina.flavin as PUBLIC, connected to sascloud.sas.com:18561/Foundation

	Gender	BloodType	_TYPE_	_FREQ_	AveChol
1	Female	A	3	178	201.41721854
2	Female	AB	3	20	166.46153846
3	Female	B	3	34	208.13043478
4	Female	O	3	208	205.59876543
5	Male	A	3	234	201.53513514
6	Male	AB	3	24	191.16666667
7	Male	B	3	62	211.14285714
8	Male	O	3	240	197.92783505

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

```
proc sort data=learn.blood out=blood;
  by gender bloodtype;
run;

data percent;
   merge blood means;
   by gender bloodtype;
   PerChol = Chol / AveChol;
   if chol > avechol then cat='H';
   else if . < chol < avechol then cat='L';
   else if chol ne . then cat='*';
run;
```

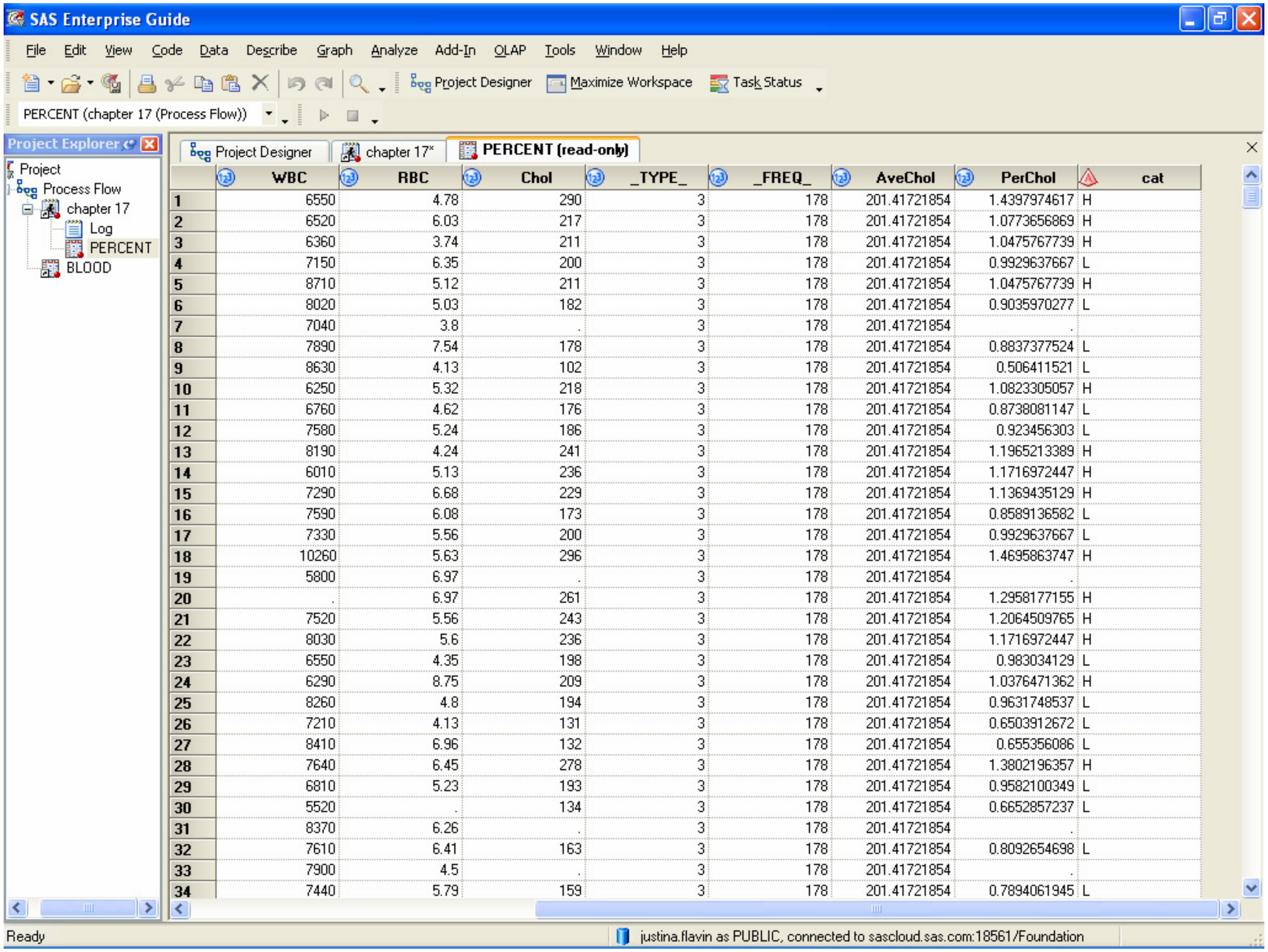
The data are sorted by gender and bloodtype and then the two data sets are merged together on gender and bloodtype.

The AveChol value on each record is reflective of the AveChol value for all subjects having the same Gender and bloodtype.

A new variable is also created to indicate if the chol value is higher or lower than the mean value.

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.

Here is the PERCENT data set.



	WBC	RBC	Chol	_TYPE_	_FREQ_	AveChol	PerChol	cat
1	6550	4.78	290	3	178	201.41721854	1.4397974617	H
2	6520	6.03	217	3	178	201.41721854	1.0773656869	H
3	6360	3.74	211	3	178	201.41721854	1.0475767739	H
4	7150	6.35	200	3	178	201.41721854	0.9929637667	L
5	8710	5.12	211	3	178	201.41721854	1.0475767739	H
6	8020	5.03	182	3	178	201.41721854	0.9035970277	L
7	7040	3.8	.	3	178	201.41721854	.	
8	7890	7.54	178	3	178	201.41721854	0.8837377524	L
9	8630	4.13	102	3	178	201.41721854	0.506411521	L
10	6250	5.32	218	3	178	201.41721854	1.0823305057	H
11	6760	4.62	176	3	178	201.41721854	0.8738081147	L
12	7580	5.24	186	3	178	201.41721854	0.923456303	L
13	8190	4.24	241	3	178	201.41721854	1.1965213389	H
14	6010	5.13	236	3	178	201.41721854	1.1716972447	H
15	7290	6.68	229	3	178	201.41721854	1.1369435129	H
16	7590	6.08	173	3	178	201.41721854	0.8589136582	L
17	7330	5.56	200	3	178	201.41721854	0.9929637667	L
18	10260	5.63	296	3	178	201.41721854	1.4695863747	H
19	5800	6.97	.	3	178	201.41721854	.	
20	.	6.97	261	3	178	201.41721854	1.2958177155	H
21	7520	5.56	243	3	178	201.41721854	1.2064509765	H
22	8030	5.6	236	3	178	201.41721854	1.1716972447	H
23	6550	4.35	198	3	178	201.41721854	0.983034129	L
24	6290	8.75	209	3	178	201.41721854	1.0376471362	H
25	8260	4.8	194	3	178	201.41721854	0.9631748537	L
26	7210	4.13	131	3	178	201.41721854	0.6503912672	L
27	8410	6.96	132	3	178	201.41721854	0.655356086	L
28	7640	6.45	278	3	178	201.41721854	1.3802196357	H
29	6810	5.23	193	3	178	201.41721854	0.9582100349	L
30	5520	.	134	3	178	201.41721854	0.6652857237	L
31	8370	6.26	.	3	178	201.41721854	.	
32	7610	6.41	163	3	178	201.41721854	0.8092654698	L
33	7900	4.5	.	3	178	201.41721854	.	
34	7440	5.79	159	3	178	201.41721854	0.7894061945	L

Ready

justina.flavin as PUBLIC, connected to sascloud.sas.com:18561/Foundation

Copyright 2009, Justina Flavin. All rights reserved. No part of this material may be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without the express written permission of the copyright holder.