STAT 430
SAS Examples SAS1
==========
ssh abc@glue.umd.edu, tap sas913 (or sas82), sas
https://www.statlab.umd.edu/sasdoc/sashtml/onldoc.htm

1. How to get the SAS window. General Introduction.
2. Simple Examples.
3. Case Study: Grades of STAT 430.

1. How to get the SAS window. General Introduction.

SAS available in windows at all the WAM labs.

General info for the WAM Labs can be found at (some info out of date)
http://www.oit.umd.edu/wheretogo/

Two important locations: EPSL library 1st floor (windows),
Parking Garage, Ground floor (windows).

From windows: Use SAS Logo by clicking
Start/Programs/Applications/SAS

Unix at WAM/GLUE labs:
1. tap sas913 2. sas

To Get SAS From unix machines NOT in WAM/Glue labs (e.g. Math Dept):
ssh abc@glue.umd.edu, tap sas913 (or sas82), sas

Useful windows:
1. Program Editor. Here we write the SAS program.
2. Log window. Records everything about the program and gives error messages, cpu time, # observations, etc. It retains all the SAS programs of the session. So if you need to copy a portion of a SAS code you eliminated from the Program Editor, you have it on record in the Log window.
3. Output window. Gives the SAS results.

To Run:
1. Write SAS code in SAS-Program Editor.
2. Annotated information in SAS-Log window.
3. Select SAS code.
4. Go to "Run" in Program Editor
5. Click "Submit Cliboard" (If click "Submit" the code disappears!)
6. Find the results in the Output window.

Comments:
To comment use asterisk and end with semicolon
*This Is A Comment;
Or use the delimiters /*This Is A Comment*/

Variable names:
Variable names must start with a letter, be not more than 8 characters in length, cannot contain blanks, no commas and semicolons etc.

Basic structure of SAS programs:
a. DATA step;
b. PROC PROCNAME options;
    STATEMENT / statement option;
    STATEMENT /statement options;
       .
       .
       .
     STATEMENT /statement options;
c. RUN;

2. Simple Examples.

Example 1: One Variable

OPTION PS=40 LS=70;

```
DATA TEST;
INPUT EXAM1;
DATALINES;
80
70
30
60
97
76
;
PROC MEANS;
RUN;
```

Output:

The MEANS Procedure

Analysis Variable : EXAM1

N	Mean	Std Dev	Minimum	Maximum
6	68.8333333	22.6134178	30.0000000	97.0000000

Example 2: Two Variables

```
OPTION PS=40 LS=70;

DATA TEST;
INPUT EXAM1 EXAM2;
DATALINES;
80 75
78 57
65 90
77 83
85 79
100 89
;
```

```
PROC MEANS DATA=TEST;
RUN;
```

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
EXAM1	6	80.8333333	11.4789663	65.0000000	100.0000000
EXAM2	6	78.8333333	12.1394673	57.0000000	90.0000000

Example 3: Giving a Grade; Data Manipulation

Note: "$" is for character data!!!

```
DATA EXAMPLE;
INPUT SUBJECT GENDER $ EXAM1 EXAM2 EXAM3;
FINAL=(EXAM1+EXAM2+EXAM3)/3;
DATALINES;
20 M 45 67 50
 1 F 78 87 91
15 F 67 75 79
 7 M 90 92 61
12 M 54 45 82
 8 F 30 40 50
;
PROC MEANS DATA=EXAMPLE MAXDEC=1;
RUN;
```

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
SUBJECT	6	10.5	6.7	1.0	20.0
EXAM1	6	60.7	22.1	30.0	90.0
EXAM2	6	67.7	21.4	40.0	92.0
EXAM3	6	68.8	17.5	50.0	91.0
FINAL	6	65.7	17.4	40.0	85.3

Continue: We now state which variables we want, since SUBJECT is only an ID.
Also we add a title.

```
DATA EXAMPLE;
INPUT SUBJECT GENDER $ EXAM1 EXAM2 EXAM3;
FINAL=(EXAM1+EXAM2+EXAM3)/3;
DATALINES;
20 M 45 67 50
 1 F 78 87 91
15 F 67 75 79
 7 M 90 92 61
12 M 54 45 82
 8 F 30 40 50
;
PROC MEANS DATA=EXAMPLE MAXDEC=1;
TITLE 'SOME STATS';
VAR EXAM1 EXAM2 EXAM3 FINAL;
RUN;
```

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
EXAM1	6	60.7	22.1	30.0	90.0
EXAM2	6	67.7	21.4	40.0	92.0
EXAM3	6	68.8	17.5	50.0	91.0
FINAL	6	65.7	17.4	40.0	85.3

```
PROC PRINT DATA=EXAMPLE;
TITLE 'ID and Grades';
ID SUBJECT;
VAR EXAM1 EXAM2 EXAM3 FINAL;
RUN;
```

```
ID and Grades:

SUBJECT    EXAM1    EXAM2    EXAM3     FINAL

   20        45       67       50     54.0000
    1        78       87       91     85.3333
   15        67       75       79     73.6667
    7        90       92       61     81.0000
   12        54       45       82     60.3333
    8        30       40       50     40.0000

NOTE: The "ID SUBJECT" causes the the first variable in the
      printout to be "SUBJECT". Otherwise it will print "OBS".

Thus if we write:

PROC PRINT DATA=EXAMPLE;
TITLE 'ID and Grades';
VAR SUBJECT EXAM1 EXAM2 EXAM3 FINAL;
RUN;

we get "OBS" number:

  ID and Grades:

Obs    SUBJECT    EXAM1    EXAM2    EXAM3     FINAL

 1        20        45       67       50     54.0000
 2         1        78       87       91     85.3333
 3        15        67       75       79     73.6667
 4         7        90       92       61     81.0000
 5        12        54       45       82     60.3333
 6         8        30       40       50     40.0000

To print sorted output: First run (There is no output):

PROC SORT DATA=EXAMPLE;
BY SUBJECT;
RUN;
```

Then print:

```
PROC PRINT DATA=EXAMPLE;
TITLE 'ID and Grades';
ID SUBJECT;
VAR EXAM1 EXAM2 EXAM3 FINAL;
RUN;
```

The output is now sorted by SUBJECT (This is the strengthof SAS!!!):

ID and Grades:

SUBJECT	EXAM1	EXAM2	EXAM3	FINAL
1	78	87	91	85.3333
7	90	92	61	81.0000
8	30	40	50	40.0000
12	54	45	82	60.3333
15	67	75	79	73.6667
20	45	67	50	54.0000

Now giving a grade using "IF"

```
DATA EXAMPLE;
INPUT SUBJECT GENDER $ EXAM1 EXAM2 EXAM3;
FINAL=(EXAM1+EXAM2+EXAM3)/3;
IF FINAL GE 0 AND FINAL LT 41 THEN GRADE="F";
ELSE IF FINAL GE 41 AND FINAL LT 60 THEN GRADE="D";
ELSE IF FINAL GE 60 AND FINAL LT 73 THEN GRADE="C";
ELSE IF FINAL GE 73 AND FINAL LT 85 THEN GRADE="B";
ELSE IF FINAL GE 85 THEN GRADE="A";
DATALINES;
20 M 45 67 50
 1 F 78 87 91
```

```
15 F 67 75 79
 7 M 90 92 61
12 M 54 45 82
 8 F 30 40 50
;

*The above is the complete DATA statement. Now some PROCS;

PROC PRINT DATA=EXAMPLE;
TITLE 'GRADE';
ID SUBJECT;
VAR GENDER EXAM1 EXAM2 EXAM3 FINAL GRADE;
RUN;
```

GRADE:

SUBJECT	GENDER	EXAM1	EXAM2	EXAM3	FINAL	GRADE
20	M	45	67	50	54.0000	D
1	F	78	87	91	85.3333	A
15	F	67	75	79	73.6667	B
7	M	90	92	61	81.0000	B
12	M	54	45	82	60.3333	C
8	F	30	40	50	40.0000	F

```
*Finally, also sorting and giving the grade;

PROC SORT DATA=EXAMPLE;
BY SUBJECT;
RUN;

PROC PRINT DATA=EXAMPLE;
TITLE 'GRADE';
ID SUBJECT;
VAR GENDER EXAM1 EXAM2 EXAM3 FINAL GRADE;
RUN;
```

GRADE

SUBJECT	GENDER	EXAM1	EXAM2	EXAM3	FINAL	GRADE
1	F	78	87	91	85.3333	A
7	M	90	92	61	81.0000	B
8	F	30	40	50	40.0000	F
12	M	54	45	82	60.3333	C
15	F	67	75	79	73.6667	B
20	M	45	67	50	54.0000	D

Now sort by both GENDER and SUBJECT: FIRST BY GENDER THAN BY SUBJECT WITHIN GENDER!!!

```
PROC SORT DATA=EXAMPLE;
BY GENDER SUBJECT;
RUN;

PROC PRINT DATA=EXAMPLE;
TITLE 'GRADE';
ID SUBJECT;
VAR GENDER EXAM1 EXAM2 EXAM3 FINAL GRADE;
RUN;
```

GRADE

SUBJECT	GENDER	EXAM1	EXAM2	EXAM3	FINAL	GRADE
1	F	78	87	91	85.3333	A
8	F	30	40	50	40.0000	F
15	F	67	75	79	73.6667	B
7	M	90	92	61	81.0000	B
12	M	54	45	82	60.3333	C
20	M	45	67	50	54.0000	D

Now also frequency count for the variables GENDER and GRADE;

```
PROC FREQ DATA=EXAMPLE;
TABLES GENDER GRADE;
RUN;
```

GRADE

The FREQ Procedure

GENDER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
F	3	50.00	3	50.00
M	3	50.00	6	100.00

GRADE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	1	16.67	1	16.67
B	2	33.33	3	50.00
C	1	16.67	4	66.67
D	1	16.67	5	83.33
F	1	16.67	6	100.00

```
*We now add some options;
*PROC with the option ORDER=FREQ (Order by frequency);
PROC FREQ DATA=EXAMPLE ORDER=FREQ;
*Statement with the option NOCUM (No Cumulative Freq);
TABLES GENDER GRADE / NOCUM;
RUN;
```

```
                    The FREQ Procedure

GENDER    Frequency     Percent
-------------------------------
F                 3       50.00
M                 3       50.00


GRADE    Frequency     Percent
-------------------------------
B                2       33.33
A                1       16.67
C                1       16.67
D                1       16.67
F                1       16.67



Example 4: Problem 1-3, p. 21
-------------------------------

DATA TAXPROB;
INPUT SS SALARY AGE RACE $;
TAX= 0.3*SALARY;
FORMAT SS SSN11.; /*SSN11. is a built-in format for SSN's*/
DATALINES;
123874414 28000 35 W
646239182 29500 37 B
012437652 35100 40 W
018451357 26500 31 W
;

PROC MEANS DATA=TAXPROB N MEAN MAXDEC=0;
TITLE 'STATS FOR SALARY AND AGE';
VAR SALARY TAX AGE;
RUN;

 STATS FOR SALARY AND AGE
```

The MEANS Procedure

Variable	N	Mean
SALARY	4	29775
TAX	4	8933
AGE	4	36

```
PROC SORT DATA=TAXPROB;
BY SS;
RUN;

PROC PRINT DATA=TAXPROB;
TITLE 'SALARY AND TAXES';
ID SS;
VAR SALARY TAX AGE;
RUN;
```

SALARY AND TAXES

SS	SALARY	TAX	AGE
012-43-7652	35100	10530	40
018-45-1357	26500	7950	31
123-87-4414	28000	8400	35
646-23-9182	29500	8850	37

Without the FORMAT statement "FORMAT SS SSN11." we get:

SALARY AND TAXES

SS	SALARY	TAX	AGE
12437652	35100	10530	40
18451357	26500	7950	31

```
     123874414     28000     8400     35
     646239182     29500     8850     37

In addition, without the "ID SS" statement we get
(same, but instead of  SS we have Obs):

 SALARY AND TAXES

           Obs    SALARY     TAX     AGE

            1      35100    10530     40
            2      26500     7950     31
            3      28000     8400     35
            4      29500     8850     37



Example 5:
----------

What is wsrong with the folowing program?

DATA MISTAKE;
INPUT ID TOWN REGION YEAR BUDGET
VOTER TURNOUT <---- variable name has a blank, too long, no semicolon.
DATALINES; <---- no data.
;

PROC MEANS DATA=MISTAKE;
VAR ID REGION VOTER TURNOUT; <---- mean of ID is nonsense.
N,STD,MEAN; <---- go in PROC line, and without commas.
RUN;

==============================================

3. Case Study: Hypothetical Grades of STAT 430.
-----------------------------------------------

OPTION PS=40 LS=70;
```

```
NOTE: We define a new alphanumeric (character) variable GRADE
      with an extra space so that its value is two characters,
      the grade followed by space or by "+" or "-". Otherwise
      when printing SAS will omit the "+" amd "-" signs!!!

NOTE: Only the first 8 characters of the NAME are typed. The rest
      is eliminated.

DATA STAT430;
INPUT SS NAME $ FINAL MIDTERM PROJECT HW;
FORMULA=0.1*HW + .2*Project + .3*Midterm + .4*Final;
     IF FORMULA EQ  .                   THEN GRADE="AUDIT";
ELSE IF FORMULA GE  0 AND FORMULA LE 39 THEN GRADE="F    ";
ELSE IF FORMULA GT 39 AND FORMULA LE 54 THEN GRADE="D    ";
ELSE IF FORMULA GT 54 AND FORMULA LE 64 THEN GRADE="C-   ";
ELSE IF FORMULA GT 64 AND FORMULA LE 75 THEN GRADE="C    ";
ELSE IF FORMULA GT 75 AND FORMULA LE 78 THEN GRADE="B-   ";
ELSE IF FORMULA GT 78 AND FORMULA LE 86 THEN GRADE="B    ";
ELSE IF FORMULA GT 86                   THEN GRADE="A    ";
FORMAT SS SSN11.; /*SSN11. is a built-in format for SSN's*/;
DATALINES;
111111101 S1               70    80  89  90
111111102 S2               62    62  84 100
.......................................
.......................................
;

Better to read the data from a file:

DATA STAT430;
INFILE '/homes/bnk/STAT430';
INPUT SS NAME $ FINAL MIDTERM PROJECT HW;
FORMULA=0.1*HW + .2*Project + .3*Midterm + .4*Final;
     IF FORMULA EQ  .                   THEN GRADE="AUDIT";
ELSE IF FORMULA GE  0 AND FORMULA LE 39 THEN GRADE="F    ";
ELSE IF FORMULA GT 39 AND FORMULA LE 54 THEN GRADE="D    ";
ELSE IF FORMULA GT 54 AND FORMULA LE 64 THEN GRADE="C-   ";
ELSE IF FORMULA GT 64 AND FORMULA LE 75 THEN GRADE="C    ";
ELSE IF FORMULA GT 75 AND FORMULA LE 78 THEN GRADE="B-   ";
ELSE IF FORMULA GT 78 AND FORMULA LE 86 THEN GRADE="B    ";
```

```
ELSE IF FORMULA GT 86                    THEN GRADE="A    ";
FORMAT SS SSN11.; /*SSN11. is a built-in format for SSN's*/;
DATALINES;
```

Can arrange by NAME:

```
PROC SORT DATA=STAT430;
BY NAME;
RUN;

PROC PRINT DATA=STAT430;
TITLE 'Grades STAT 430';
ID NAME;
VAR SS FINAL MIDTERM PROJECT HW FORMULA GRADE;;
RUN;
```

NAME	SS	FINAL	MIDTERM	PROJECT	HW	FORMULA	GRADE
S1	111-11-1101	70	80	89	90	78.8	B
S10	111-11-1110	70	75	95	100	79.5	B
S11	111-11-1111	.	.	.	.	.	AUDIT
S12	111-11-1112	88	90	95	100	91.2	A

```
.............................
.............................
```

Can arrange by SSN:

```
PROC SORT DATA=STAT430;
BY SS;
RUN;

PROC PRINT DATA=STAT430;
TITLE 'Grades STAT 430';
ID SS;
VAR NAME FINAL MIDTERM PROJECT HW FORMULA GRADE;;
RUN;
```

SS	NAME	FINAL	MIDTERM	PROJECT	HW	FORMULA	GRADE
111-11-1101	S1	70	80	89	90	78.8	B

```
111-11-1102   S2       62       62        84     100     70.2    C
.......................................
.......................................
```

Can arrange by SSN without names:

```
PROC SORT DATA=STAT430;
BY SS;
RUN;

PROC PRINT DATA=STAT430;
TITLE 'Grades STAT 430';
ID SS;
VAR FINAL FORMULA GRADE;
RUN;
```

```
                        SS    FINAL    FORMULA    GRADE

               111-11-1101      80       78.8     B
               111-11-1102      72       70.2     C
               ................................
..............................................
```

Can arrange by SSN with FINAL and GRADE only:

```
PROC SORT DATA=STAT430;
BY SS;
RUN;

PROC PRINT DATA=STAT430;
TITLE 'Grades STAT 430';
ID SS;
VAR FINAL GRADE;
RUN;
```

```
                              SS    FINAL    GRADE

                     111-11-1101      80     B
```

```
                    111-11-1102      72     C
                ..................................
```

```
Can arrange by GRADE  with SSN and FINAL only:
PROC SORT DATA=STAT430;
BY GRADE;
RUN;

PROC PRINT DATA=STAT430;
TITLE 'Grades STAT';
ID GRADE;
VAR SS FINAL;
RUN;
```

GRADE	SS	FINAL
B	111-11-1101	80
C	111-11-1102	72

```
PROC MEANS DATA=STAT430 N MEAN MAXDEC=0;
TITLE 'STATS FOR STAT 430';
VAR FINAL MIDTERM PROJECT HW FORMULA;
RUN;
```

The MEANS Procedure

Variable	N	Mean
FINAL	27	67
MIDTERM	27	81
PROJECT	27	87
HW	27	93
FORMULA	27	78

```
PROC UNIVARIATE DATA=STAT430 NORMAL PLOT;
TITLE 'STATS FOR STAT 430';
VAR FINAL MIDTERM PROJECT HW FORMULA;
RUN;
```

Only record stats for the final grade from FORMULA!!!

The UNIVARIATE Procedure
Variable: FORMULA

Moments

N	27	Sum Weights	27
Mean	77.5962963	Sum Observations	2095.1
Std Deviation	12.3501318	Variance	152.525755
Skewness	-3.1860688	Kurtosis	13.8312058
Uncorrected SS	166537.67	Corrected SS	3965.66963
Coeff Variation	15.9158779	Std Error Mean	2.37678397

Basic Statistical Measures

Location		Variability	
Mean	77.59630	Std Deviation	12.35013
Median	79.10000	Variance	152.52575
Mode	82.50000	Range	68.20000
		Interquartile Range	8.20000

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	32.6476	Pr > |t|	<.0001
Sign	M	13.5	Pr >= |M|	<.0001
Signed Rank	S	189	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	92.4
99%	92.4
95%	91.2
90%	88.8
75% Q3	82.5
50% Median	79.1
25% Q1	74.3
10%	70.1
5%	68.4
1%	24.2
0% Min	24.2

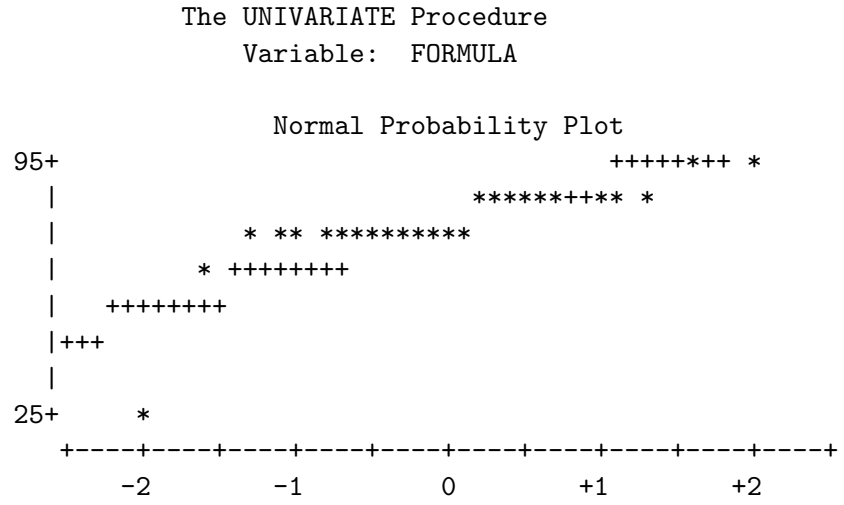
Missing Values

Missing Value	Count	Percent Of All Obs	Percent Of Missing Obs
.	4	12.90	100.00

Stem	Leaf	#	Boxplot
9	12	2	|
8	00122222789	11	+-----+
7	003346678899	12	*--+--*
6	8	1	|
5			
4			
3			
2	4	1	*

----+----+----+----+

Multiply Stem.Leaf by 10**+1



Some categorical results (Can get the same with horizontal PROC CHART as we do below):

```
PROC FREQ DATA=STAT430;
TABLES FORMULA GRADE;
RUN;
```

The FREQ Procedure

GRADE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	5	16.13	5	16.13
AUDIT	4	12.90	9	29.03
B	11	35.48	20	64.52
B-	4	12.90	24	77.42
C	6	19.35	30	96.77
F	1	3.23	31	100.00

Now Plots:

```
PROC CHART DATA=STAT430;
VBAR GRADE;
RUN;
```

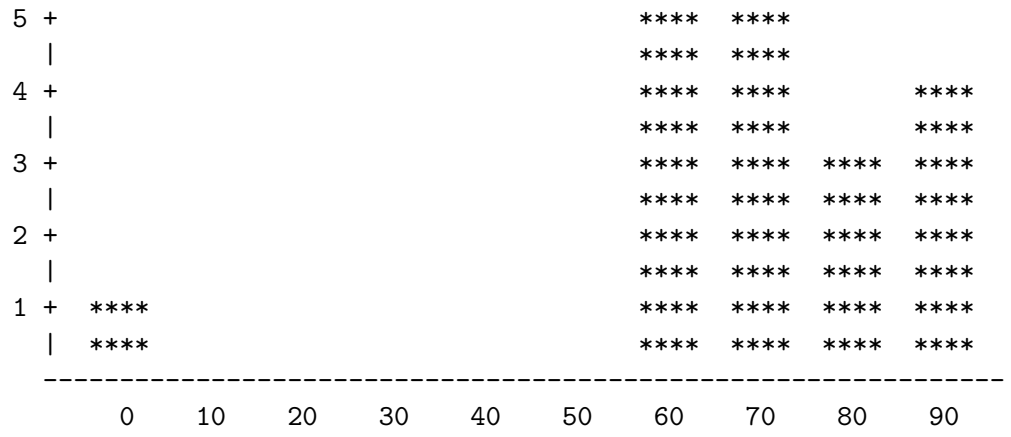
```
Frequency

11 +                          *****
   |                          *****
10 +                          *****
   |                          *****
 9 +                          *****
   |                          *****
 8 +                          *****
   |                          *****
 7 +                          *****
   |                          *****
 6 +                          *****               *****
   |                          *****               *****
 5 +      *****               *****               *****
   |      *****               *****               *****
 4 +      *****     *****     *****     *****     *****
   |      *****     *****     *****     *****     *****
 3 +      *****     *****     *****     *****     *****
   |      *****     *****     *****     *****     *****
 2 +      *****     *****     *****     *****     *****
   |      *****     *****     *****     *****     *****
 1 +      *****     *****     *****     *****     *****     *****
   |      *****     *****     *****     *****     *****     *****
   ---------------------------------------------------------------
            A       AUDIT       B         B-        C         F

                                  GRADE
```

```
PROC CHART DATA=STAT430;
HBAR GRADE;
RUN;
```

```
GRADE                                              Cum.              Cum.
                                             Freq  Freq  Percent  Percent
        |
A       |**********                             5     5    16.13    16.13
        |
AUDIT   |********                               4     9    12.90    29.03
        |
B       |**********************                11    20    35.48    64.52
        |
B-      |********                               4    24    12.90    77.42
        |
C       |************                           6    30    19.35    96.77
        |
F       |**                                     1    31     3.23   100.00
        |
        ----+---+---+---+---+--
            2   4   6   8   10

                  Frequency
```

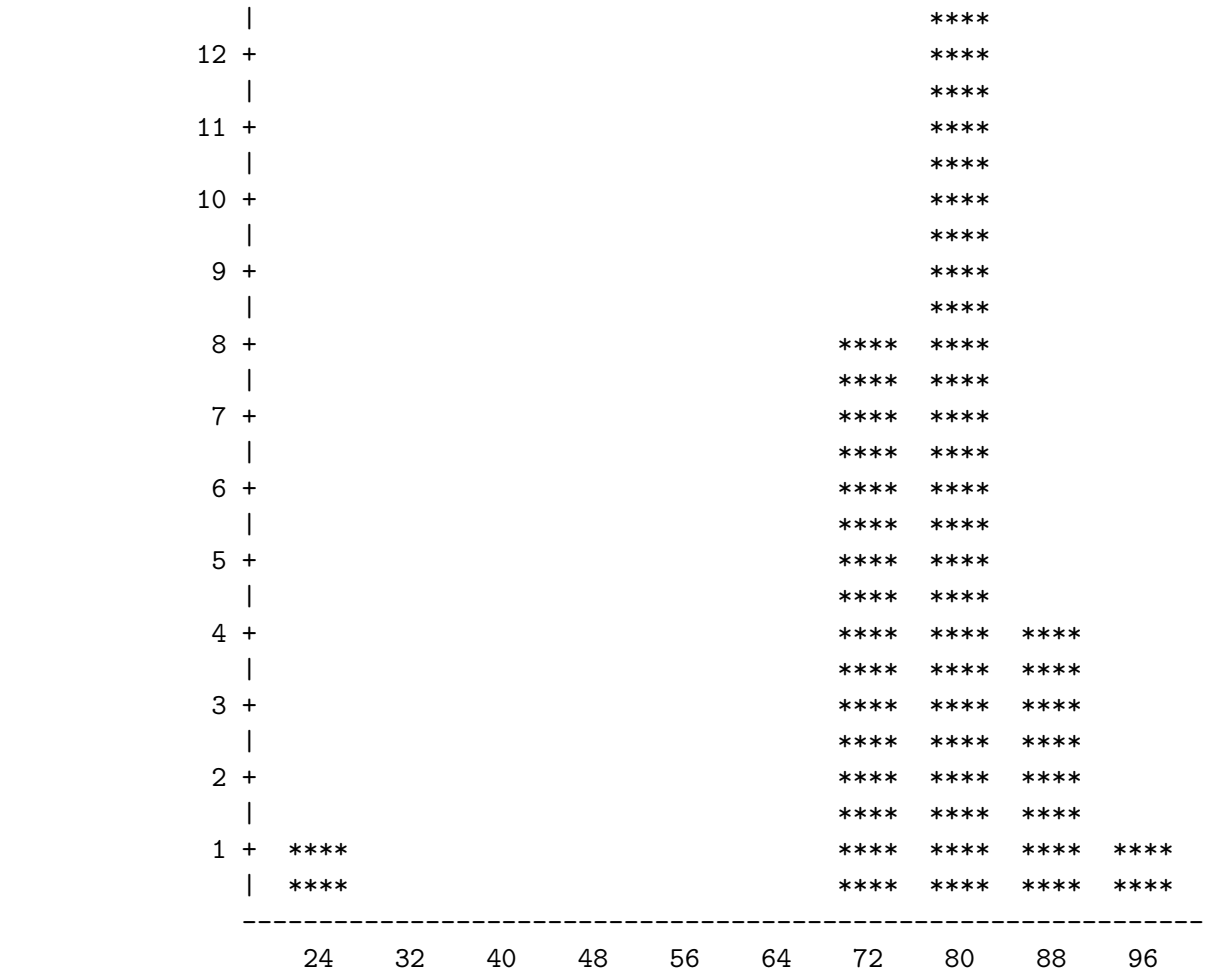
```
PROC CHART DATA=STAT430;
VBAR FINAL/ LEVELS=10;
RUN;
```

```
Frequency

11 +                                       ****
   |                                       ****
10 +                                       ****
   |                                       ****
 9 +                                       ****
   |                                       ****
 8 +                                 ****  ****
   |                                 ****  ****
 7 +                                 ****  ****
   |                                 ****  ****
 6 +                                 ****  ****
   |                                 ****  ****
```




```
PROC CHART DATA=STAT430;
VBAR FORMULA/ LEVELS=10;
RUN;
```

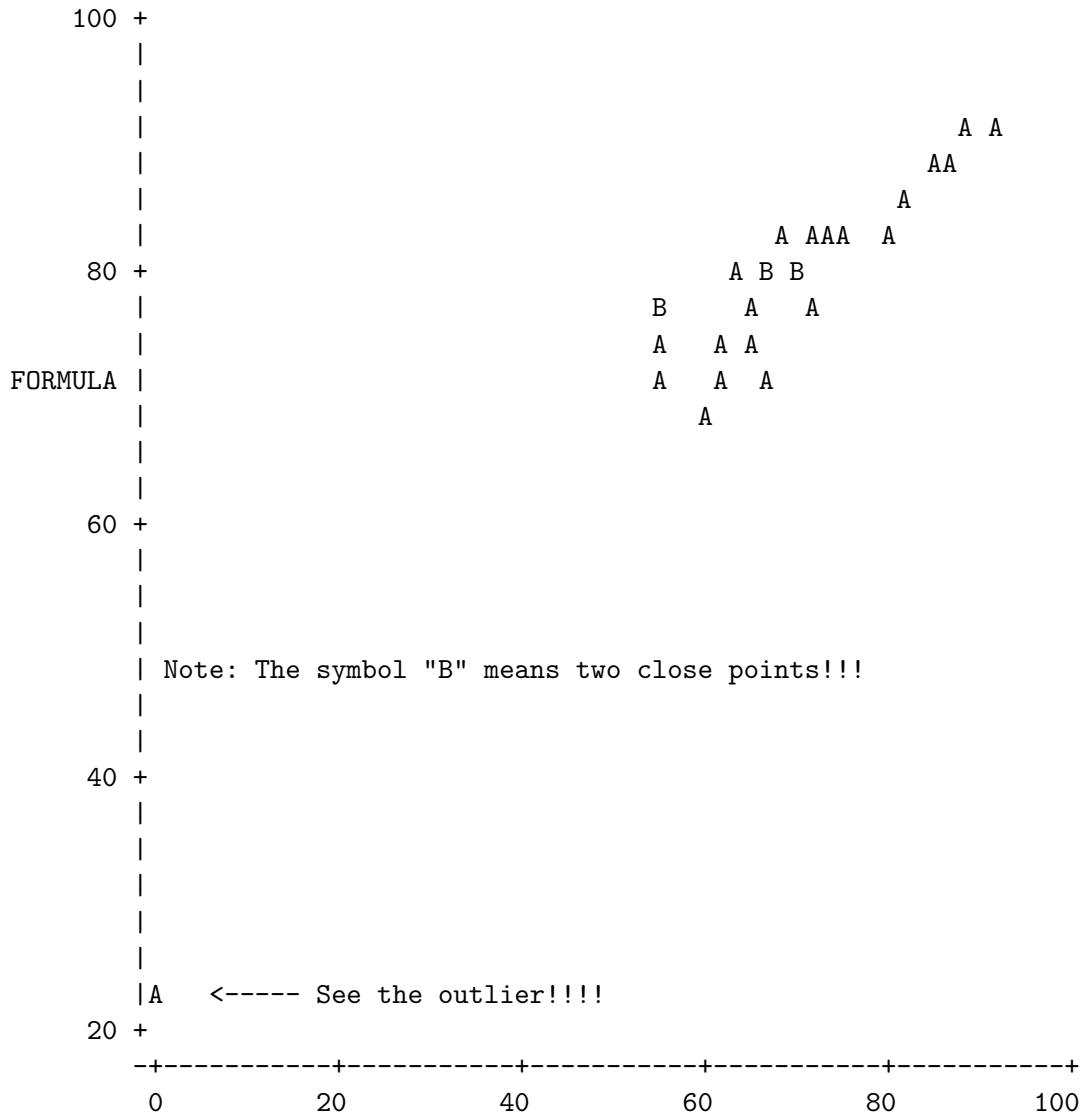
```
  Frequency

  13 +                                                ****
```



FORMULA Midpoint

Plot FORMULA (y) vs FINAL (x):

```
PROC PLOT DATA=STAT430;
PLOT FORMULA*FINAL;
RUN;
```

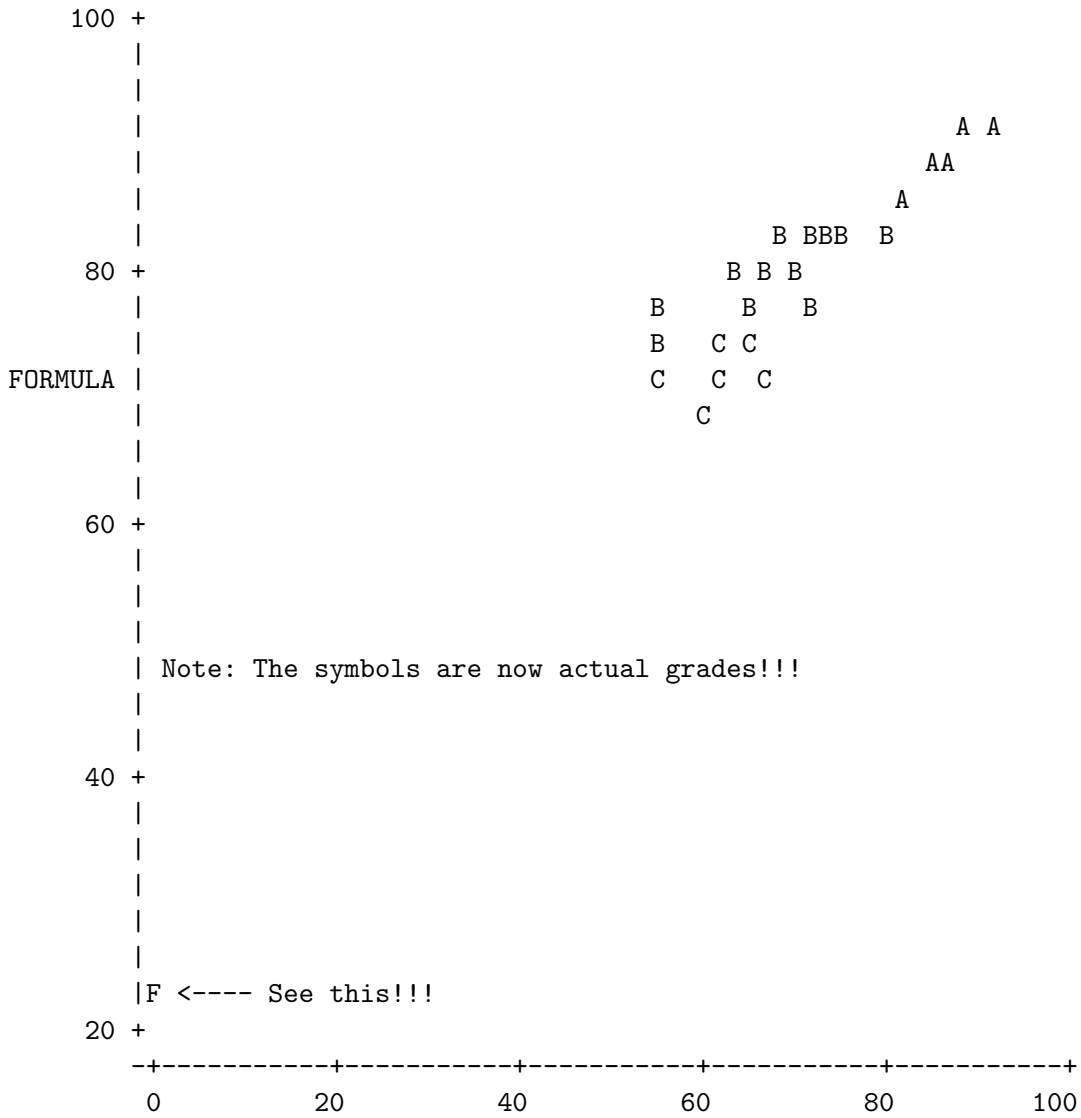


NOTE: 4 obs had missing values.

Can plot FORMULA vs FINAL by grade: See how the high B's and A's "run away" from the rest of the grades!!! But the B- and C are somewhat confused.

PROC PLOT DATA=STAT430;

```
PLOT FORMULA*FINAL=grade;
RUN;
```




```
NOTE: 4 obs had missing values.  3 obs hidden.
```