Data Validation

Statistics 426: SAS Programming

Module 7

2021

Data validation

Data errors: occur when data values are not appropriate for the SAS statements that are specified in a program. SAS detects data errors during program execution. When a data error is detected, SAS writes a note to the log *and continues to execute the program*

Syntax errors occur when program statements do not conform to the rules of the SAS language. Examples are misspelled words, unmatched quotes, missing commas, invalid operations, incorrect data type for specified command, occasionally missing data values, and many more

Validating data

in general, SAS procedures analyze data, produce output or manage SAS files. Some SAS procedures can be used to detect invalid data

During the processing of every DATA step, SAS automatically creates the following variables
- The `_N_` variable: counts the number of times the DATA step begins to iterate
- The `_ERROR_` variable: signals the occurrence of an error caused by the data during execution
- 0 = no errors exist
- 1 = at least one error occurred

Missing and invalid values

PROC PRINT: can show missing values and invalid values (like a date – as in hire date that happens before the employee's birthdate). The WHERE statement will be used in PROC PRINT

When trying to use WHERE statement to find information about dates, you'd either have to know the SAS date value (the number of days from 1/1/1960 to the date in question) or use a SAS date constant.

To write a SAS date constant, enclose the date (in the ddMMMYYYY format) in quotation marks, followed by the letter d.

Ex: January 1, 1974

To write as a SAS date constant: `'01jan1974'd`

```
WHERE Hire_Date < '01jan1974'd;
```

Missing or invalid values of categorical values

PROC FREQ: produces one-way to n-way frequency tables. Can show if categorical variables have missing or invalid values

General form of PROC FREQ

```
PROC FREQ data=SASdataset <options>;
    TABLES variable(s) / <options>;
RUN;
```

1st line options: NLEVELS option displays a table that provides the number of distinct values for each variable named in the TABLES statement

TABLES statement options: noprint – suppresses the frequency tables

Quantitative values in acceptable range and if values are missing

PROC MEANS: can show if quantitative values are in an acceptable range and if values are missing

General form of PROC MEANS

```
PROC MEANS data=SASdataset <options> <statistics>;
    VAR variable(s);
RUN;
```

The statistics option is used to display specify statistics:
- Default: N, mean, standard deviation, minimum and maximum
- Can specify: N, NMISS (number of missing observations), MIN, MAX
- For more statistics, search for "SAS proc means statistics"

Data within acceptable ranges

PROC UNIVARIATE: produces summary reports that display descriptive statistics. Can show if quantitative variables are not within acceptable ranges

Produces basic statistical measures, tests for location, quantiles, extreme observations and missing values.

General form of PROC UNIVARIATE

```
PROC UNIVARIATE data=SASdataset;
    VAR variable(s);
run;
```

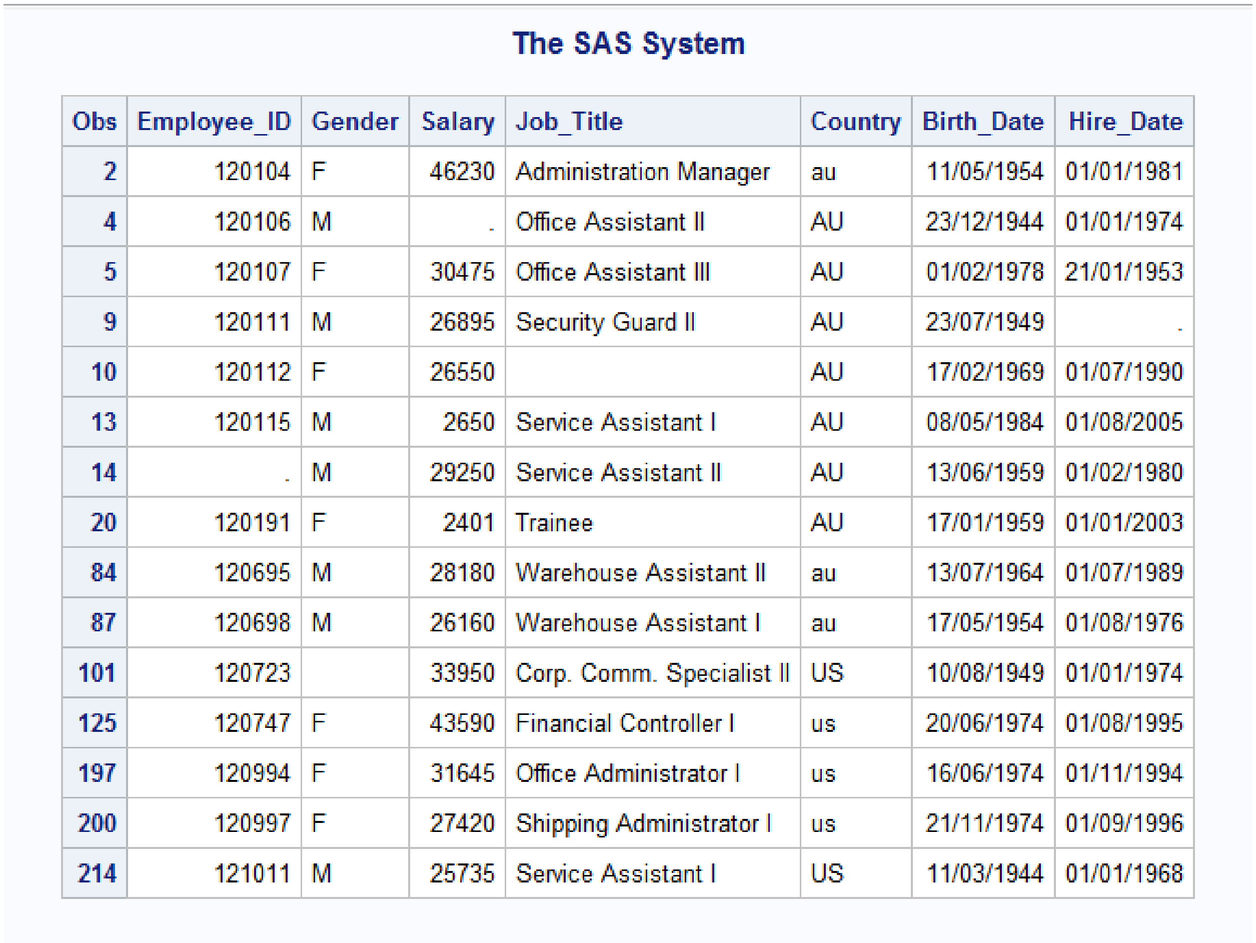
Drawbacks of the procedures

One issue with using some of the PROCs is that there are some redundancies between all the methods but you will get a decent picture of your dataset using the four listed in this module

proc print validation

```
libname hercules 's:\courses\stat-renaes\stat426\data1';

proc print data=hercules.nonsales;
run;

proc print data=hercules.nonsales;
   var Employee_ID Gender Salary Job_Title
        Country Birth_Date Hire_Date;
   where Employee_ID = . or
        Gender not in ('F','M') or
        Salary not between 24000 and 500000 or
        Job_Title = ' ' or
        Country not in ('AU','US') or
        Birth_Date > Hire_Date or
        Hire_Date < '01JAN1974'd;
run;
```

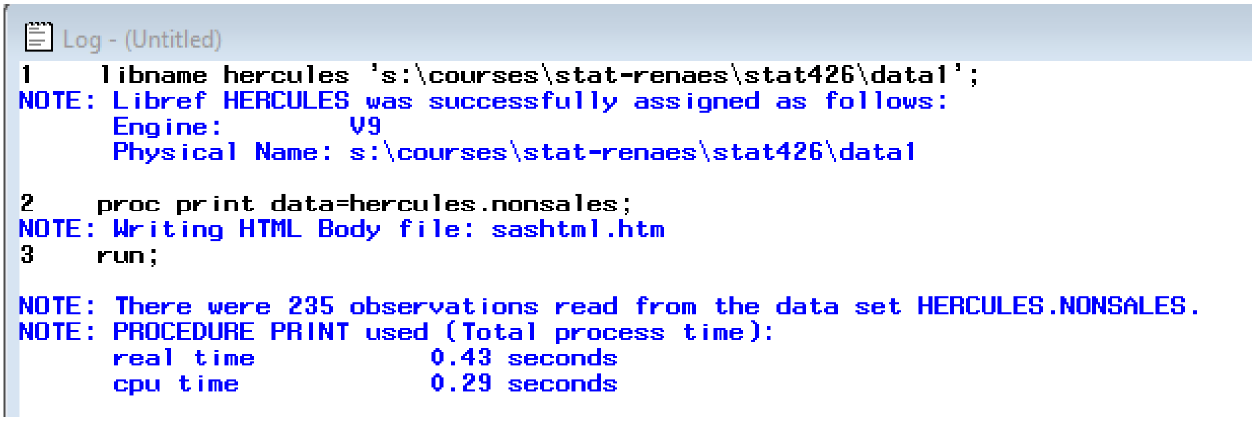
proc print validation output



The SAS System

Obs	Employee_ID	Gender	Salary	Job_Title	Country	Birth_Date	Hire_Date
2	120104	F	46230	Administration Manager	au	11/05/1954	01/01/1981
4	120106	M	.	Office Assistant II	AU	23/12/1944	01/01/1974
5	120107	F	30475	Office Assistant III	AU	01/02/1978	21/01/1953
9	120111	M	26895	Security Guard II	AU	23/07/1949	.
10	120112	F	26550		AU	17/02/1969	01/07/1990
13	120115	M	2650	Service Assistant I	AU	08/05/1984	01/08/2005
14	.	M	29250	Service Assistant II	AU	13/06/1959	01/02/1980
20	120191	F	2401	Trainee	AU	17/01/1959	01/01/2003
84	120695	M	28180	Warehouse Assistant II	au	13/07/1964	01/07/1989
87	120698	M	26160	Warehouse Assistant I	au	17/05/1954	01/08/1976
101	120723		33950	Corp. Comm. Specialist II	US	10/08/1949	01/01/1974
125	120747	F	43590	Financial Controller I	us	20/06/1974	01/08/1995
197	120994	F	31645	Office Administrator I	us	16/06/1974	01/11/1994
200	120997	F	27420	Shipping Administrator I	us	21/11/1974	01/09/1996
214	121011	M	25735	Service Assistant I	US	11/03/1944	01/01/1968

proc print validation log

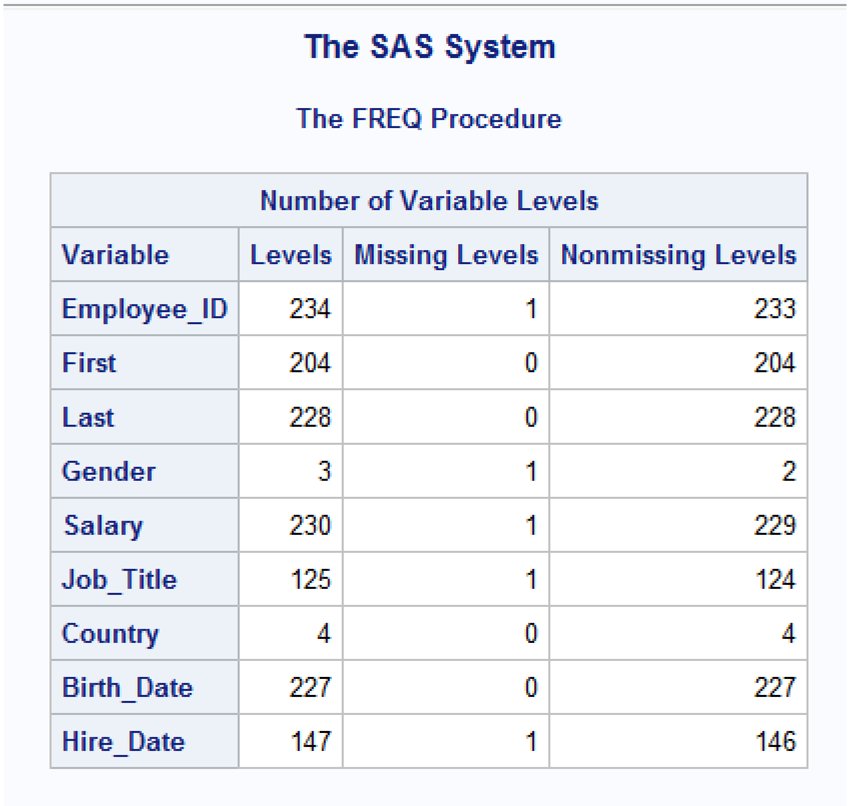


```
1     libname hercules 's:\courses\stat-renaes\stat426\data1';
NOTE: Libref HERCULES was successfully assigned as follows:
      Engine:        V9
      Physical Name: s:\courses\stat-renaes\stat426\data1

2     proc print data=hercules.nonsales;
NOTE: Writing HTML Body file: sashtml.htm
3     run;

NOTE: There were 235 observations read from the data set HERCULES.NONSALES.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.43 seconds
      cpu time            0.29 seconds
```

proc freq validation

```
proc freq data=hercules.nonsales nlevels;
   tables Gender Country Employee_ID;
run;

proc freq data=hercules.nonsales nlevels;
   tables Gender Country Employee_ID / noprint;
run;

proc freq data=hercules.nonsales nlevels;
   tables _all_ / noprint;
run;
```

proc freq validation output



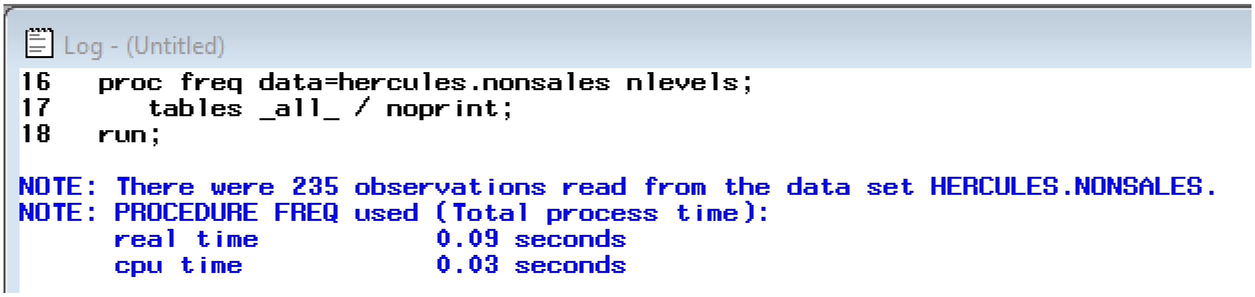
The SAS System

The FREQ Procedure

Number of Variable Levels

Variable	Levels	Missing Levels	Nonmissing Levels
Employee_ID	234	1	233
First	204	0	204
Last	228	0	228
Gender	3	1	2
Salary	230	1	229
Job_Title	125	1	124
Country	4	0	4
Birth_Date	227	0	227
Hire_Date	147	1	146

proc freq validation log

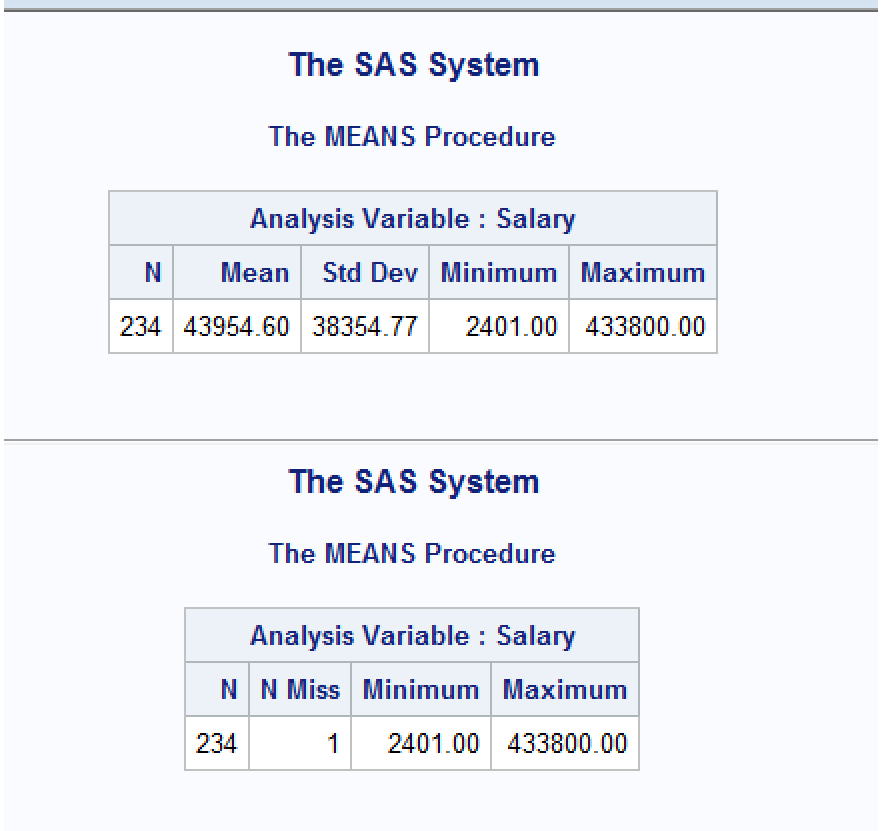


```
16   proc freq data=hercules.nonsales nlevels;
17      tables _all_ / noprint;
18   run;

NOTE: There were 235 observations read from the data set HERCULES.NONSALES.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.09 seconds
      cpu time            0.03 seconds
```

proc means validation

```
proc means data=hercules.nonsales;
   var Salary;
run;

proc means data=hercules.nonsales n nmiss min max;
   var Salary;
run;
```

proc means validation output



The SAS System

The MEANS Procedure

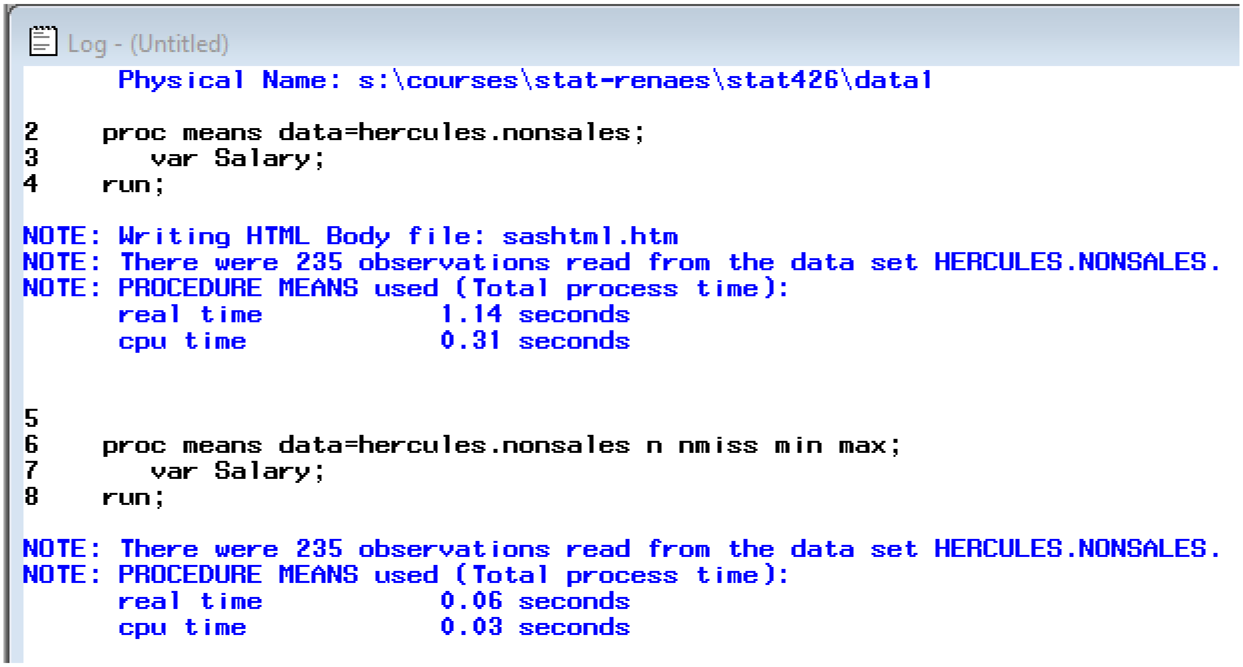
Analysis Variable : Salary				
N	Mean	Std Dev	Minimum	Maximum
234	43954.60	38354.77	2401.00	433800.00

The SAS System

The MEANS Procedure

Analysis Variable : Salary			
N	N Miss	Minimum	Maximum
234	1	2401.00	433800.00

proc means validation log

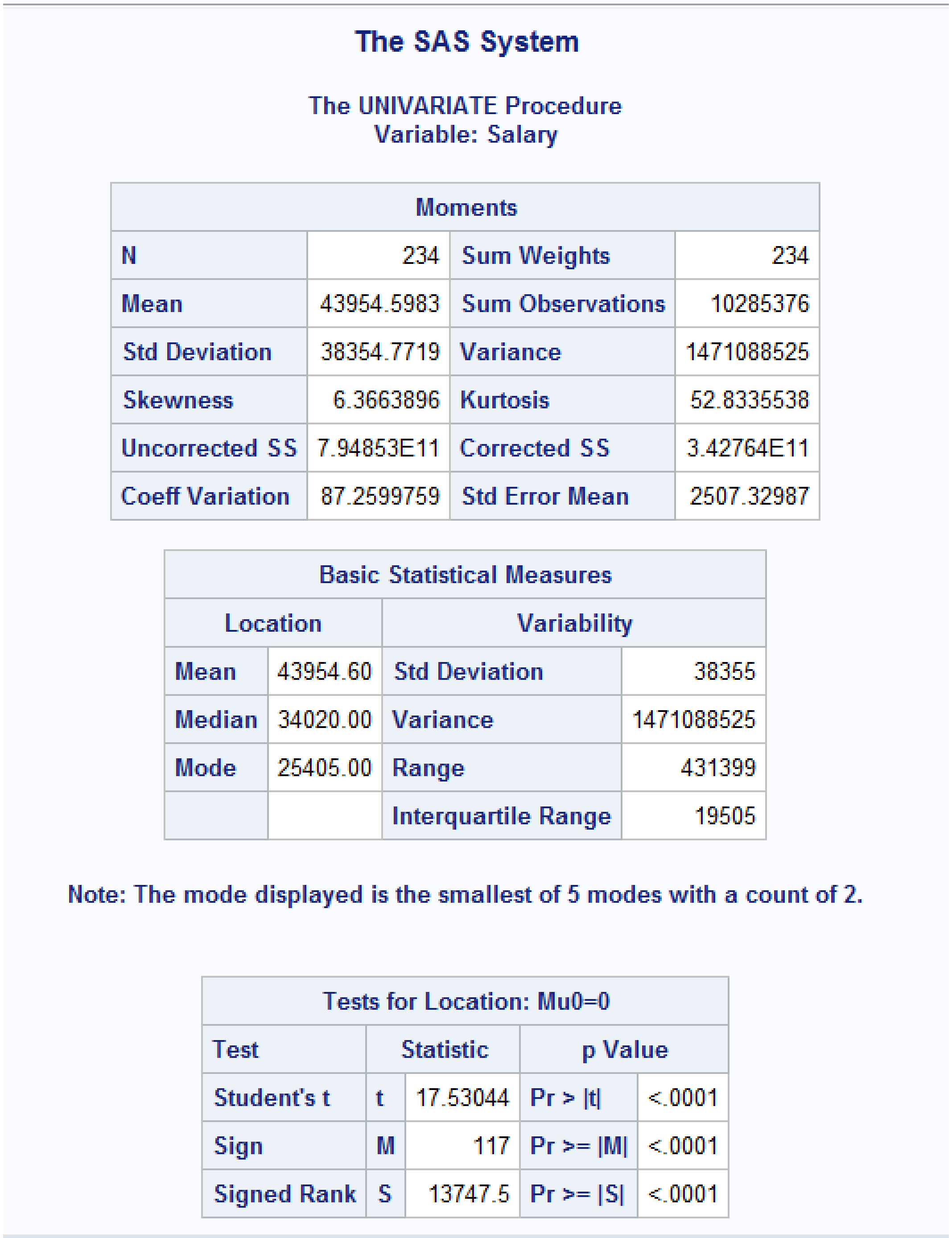


```
      Physical Name: s:\courses\stat-renaes\stat426\data1

2     proc means data=hercules.nonsales;
         var Salary;
3
4     run;

NOTE: Writing HTML Body file: sashtml.htm
NOTE: There were 235 observations read from the data set HERCULES.NONSALES.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           1.14 seconds
      cpu time            0.31 seconds


5
6     proc means data=hercules.nonsales n nmiss min max;
7        var Salary;
8     run;

NOTE: There were 235 observations read from the data set HERCULES.NONSALES.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.06 seconds
      cpu time            0.03 seconds
```

proc univariate validation

```
proc univariate data=hercules.nonsales;
   var Salary;
run;
```

proc univariate validation output 1



The SAS System

The UNIVARIATE Procedure
Variable: Salary

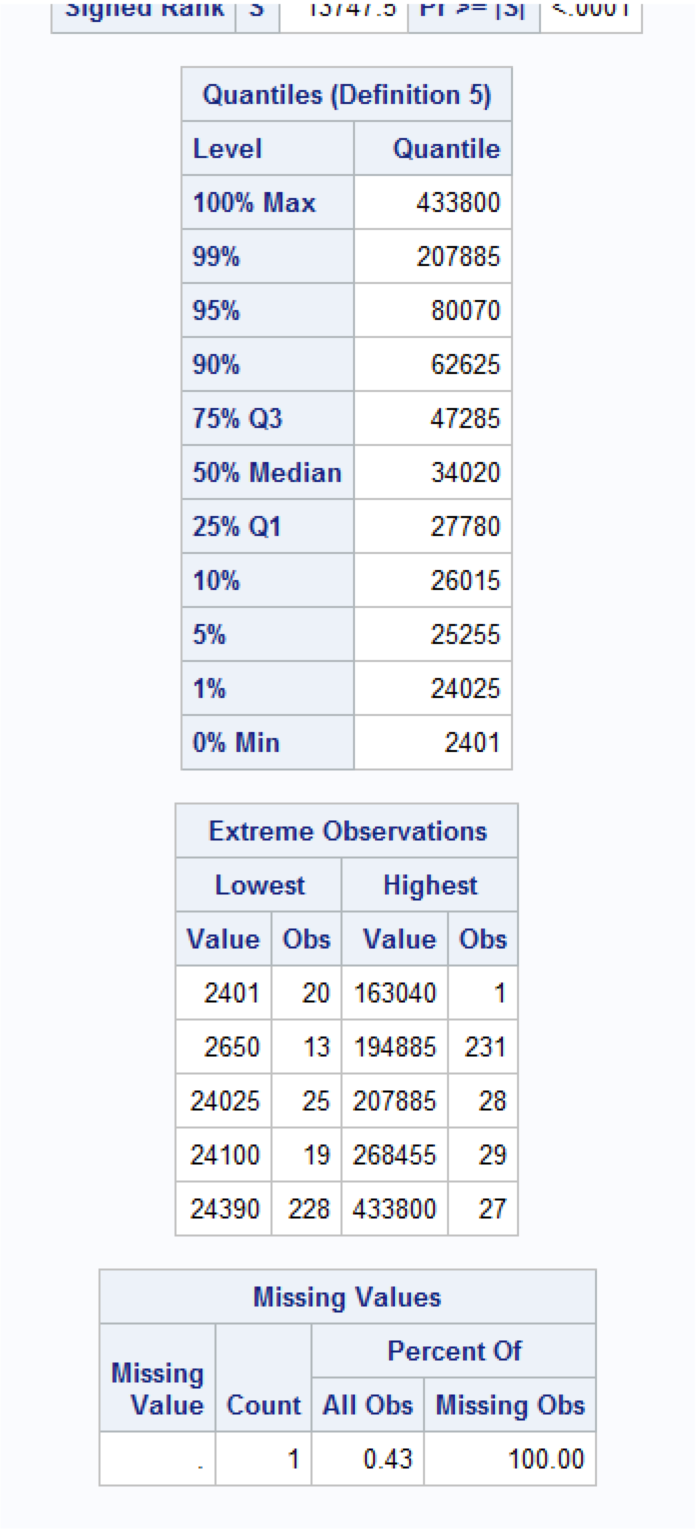
Moments			
N	234	Sum Weights	234
Mean	43954.5983	Sum Observations	10285376
Std Deviation	38354.7719	Variance	1471088525
Skewness	6.3663896	Kurtosis	52.8335538
Uncorrected SS	7.94853E11	Corrected SS	3.42764E11
Coeff Variation	87.2599759	Std Error Mean	2507.32987

Basic Statistical Measures			
Location		Variability	
Mean	43954.60	Std Deviation	38355
Median	34020.00	Variance	1471088525
Mode	25405.00	Range	431399
		Interquartile Range	19505

Note: The mode displayed is the smallest of 5 modes with a count of 2.

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	17.53044	Pr > |t|	<.0001
Sign	M	117	Pr >= |M|	<.0001
Signed Rank	S	13747.5	Pr >= |S|	<.0001

proc univariate validation output 2



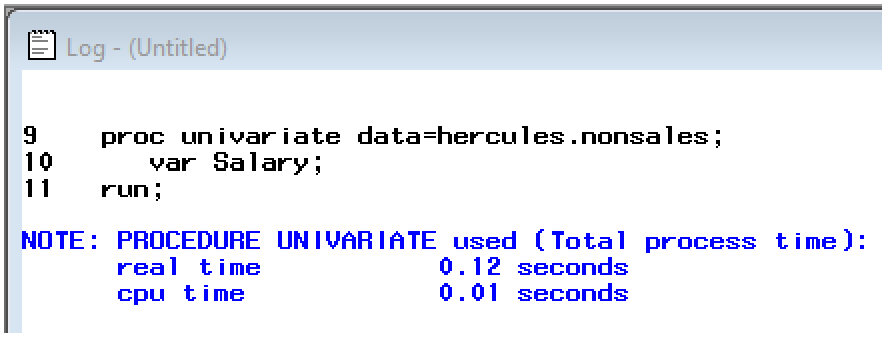
Signed Rank	S	13747.5	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	433800
99%	207885
95%	80070
90%	62625
75% Q3	47285
50% Median	34020
25% Q1	27780
10%	26015
5%	25255
1%	24025
0% Min	2401

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
2401	20	163040	1
2650	13	194885	231
24025	25	207885	28
24100	19	268455	29
24390	228	433800	27

Missing Values			
		Percent Of	
Missing Value	**Count**	**All Obs**	**Missing Obs**
.	1	0.43	100.00

proc univariate validation log



```
9     proc univariate data=hercules.nonsales;
10       var Salary;
11    run;

NOTE: PROCEDURE UNIVARIATE used (Total process time):
      real time           0.12 seconds
      cpu time            0.01 seconds
```