Data cleaning

Statistics 426: SAS Programming

Module 8

2021

Cleaning

Some techniques for cleaning data include

(1) Edit raw outside of SAS (in spreadsheet or other format types)
(2) Using the Explorer window in SAS using Viewtable (click Edit menu, select `edit mode`)
(3) Programmatically edit the dataset using the DATA step
(4) Programmatically edit the dataset using PROC SQL

Assignment statement

Assignment statements are used in the DATA step to update existing variables or to create new ones.

General form:
`variable = expression;`

The variable names an existing or new variable; an expression is a sequence of operators and operands that form a set of instructions that produce a value (like the WHERE statement where-expression)

Operands/operators review

Are constants (character, numeric or date) and variables (numeric or character); examples will show using assignment statement

`bonus = 500;` (numeric constant)
`gender='M';` (character constant)
`hire_date='01apr2008'd;` (date constant)
`newsalary=1.1*salary;` (variable)
`revenue=quantity*price;`
`newcountry=upcase(country);`
`logtime=log(time);`
`sqrttime=sqrt(time);`

Cleaning functions+

SAS functions

`UPCASE`: converts all letters in an argument to uppercase

General form:

`UPCASE(argument);`

The assignment statement is executed for every observation regardless of whether or not the value needs to be treated

Example:

```
Country=upcase(country);
```

Conditional statements (review)

IF-THEN statements: execute a SAS statement for observations that meet specific conditions.

General form:

```
IF expression THEN statement;
```

`Expression`: sequence of operands and operators that form a set of instructions
`Statement`: any executable statement, i.e. assignment statement

```
if employee_id = 123456 then salary=36000;
```

SAS functions

SAS routines that return a value that is determined from specified arguments. Some SAS functions manipulate character values, compute descriptive statistics or manipulate SAS date values.

General form:
`function-name(argument1, arg2, ...);`

SUM: sums the arguments in the function

General form: `sum(arg1, arg2, ...);`
Values must be numeric and missing values are ignored by some (not all) of the statistic functions

Date functions

Date functions can be used to extract information from SAS date values and/or create SAS date values

1/1/1959; SAS date value = -365 1/1/1960 = 0 1/1/1961 = 366

SAS date functions

Function	Description
TODAY()	Returns the current date as a SAS date value (number of days from 1/1/1960 to today)
MDY(month,day,year)	Returns a SAS date from numeric month, day and year values
YEAR(SASdate)	Extracts the year from a SAS date and returns a 4-digit value for the year
QTR(SASdate)	Extracts the quarter from a SAS date and returns a value from 1 to 4
MONTH(SASdate)	Extracts the month from a SAS date and returns a value from 1 to 12
DAY(SASdate)	Extracts the day of the month from a SAS date and returns a value from 1 to 31
WEEKDAY(SASdate)	Extracts the day of the week from a SAS date and returns a value from 1 to 7; 1=Sunday

SASdates

They can be used easily with the assignment statement to create new or update existing variables

```
bonusmonth=month(Hire_date);
annivbonus=mdy(month(hire_date),15,2008);
annivbonus=mdy(month(hire_date),day(hire_date),year(hire_date));
```

Cleaning part I

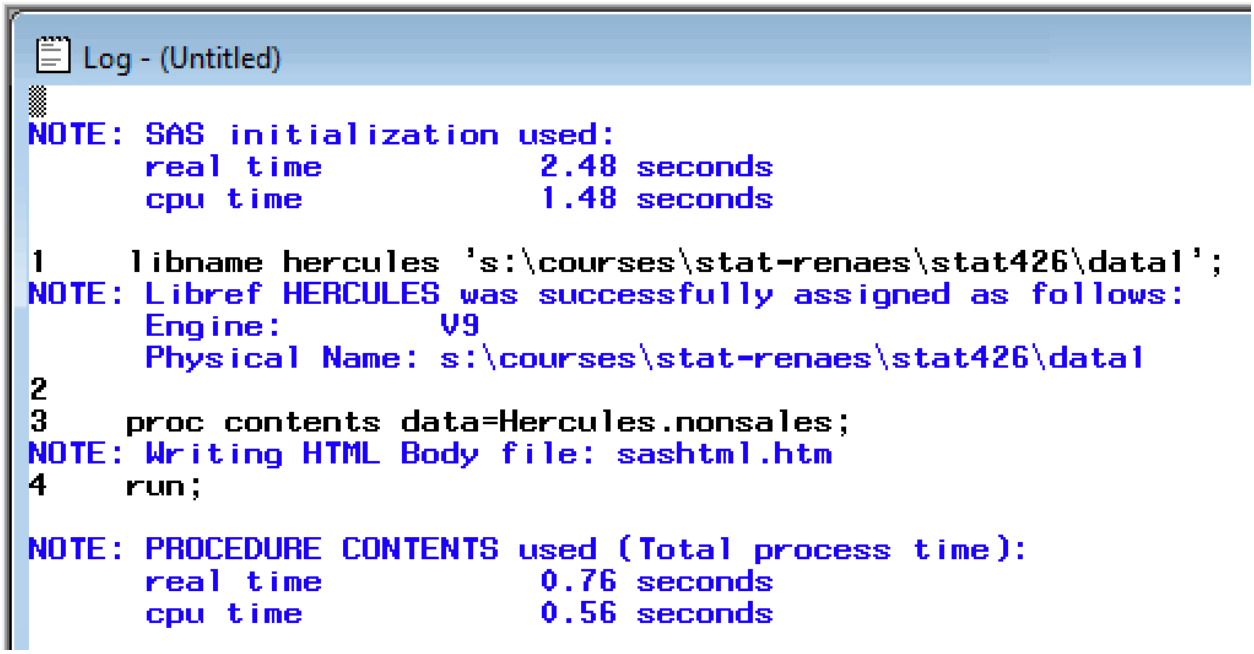
```
libname hercules 's:\courses\stat-renaes\stat426\data1';

proc contents data=Hercules.nonsales;
run;
```

Cleaning part I output

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format
8	Birth_Date	Num	8	DDMMYY10.
7	Country	Char	2	
1	Employee_ID	Num	8	12.
2	First	Char	12	
4	Gender	Char	1	
9	Hire_Date	Num	8	DDMMYY10.
6	Job_Title	Char	25	
3	Last	Char	18	
5	Salary	Num	8	

Cleaning part I log



Cleaning part II

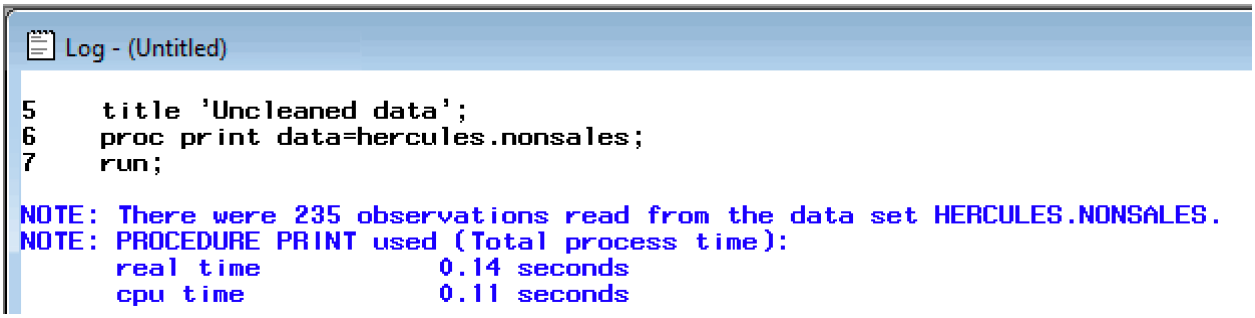
```
title 'Uncleaned data';
proc print data=hercules.nonsales;
run;
```

Cleaning part II output

Uncleaned data

Obs	Employee_ID	First	Last	Gender	Salary	Job_Title	Country	Birth_Date	Hire_Date
1	120101	Patrick	Lu	M	163040	Director	AU	18/08/1976	01/07/2003
2	120104	Kareen	Billington	F	46230	Administration Manager	au	11/05/1954	01/01/1981
3	120105	Liz	Povey	F	27110	Secretary I	AU	21/12/1974	01/05/1999
4	120106	John	Hornsey	M	.	Office Assistant II	AU	23/12/1944	01/01/1974
5	120107	Sherie	Sheedy	F	30475	Office Assistant III	AU	01/02/1978	21/01/1953
6	120108	Gladys	Gromek	F	27660	Warehouse Assistant II	AU	23/02/1984	01/08/2006
7	120108	Gabriele	Baker	F	26495	Warehouse Assistant I	AU	15/12/1986	01/10/2006
8	120110	Dennis	Entwisle	M	28615	Warehouse Assistant III	AU	20/11/1949	01/11/1979
9	120111	Ubaldo	Spillane	M	26895	Security Guard II	AU	23/07/1949	.
10	120112	Ellis	Glattback	F	26550		AU	17/02/1969	01/07/1990

Cleaning part II log



Cleaning part III

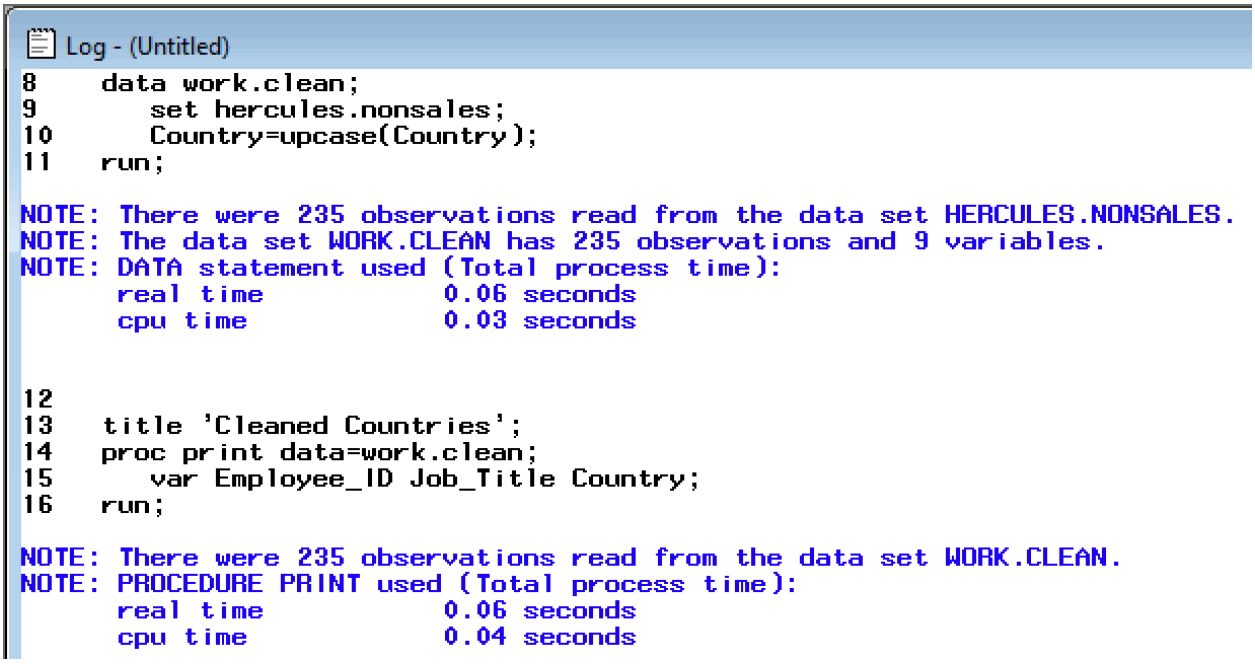
```
data work.clean;
   set hercules.nonsales;
   Country=upcase(Country);
run;

title 'Cleaned Countries';
proc print data=work.clean;
   var Employee_ID Job_Title Country;
run;
```

Cleaning part III output

Cleaned Countries

Obs	Employee_ID	Job_Title	Country
1	120101	Director	AU
2	120104	Administration Manager	AU
3	120105	Secretary I	AU
4	120106	Office Assistant II	AU
5	120107	Office Assistant III	AU
6	120108	Warehouse Assistant II	AU
7	120108	Warehouse Assistant I	AU
8	120110	Warehouse Assistant III	AU
9	120111	Security Guard II	AU
10	120112		AU

Cleaning part III log



```
Log - (Untitled)
8     data work.clean;
9        set hercules.nonsales;
10       Country=upcase(Country);
11    run;

NOTE: There were 235 observations read from the data set HERCULES.NONSALES.
NOTE: The data set WORK.CLEAN has 235 observations and 9 variables.
NOTE: DATA statement used (Total process time):
      real time           0.06 seconds
      cpu time            0.03 seconds


12
13    title 'Cleaned Countries';
14    proc print data=work.clean;
15       var Employee_ID Job_Title Country;
16    run;

NOTE: There were 235 observations read from the data set WORK.CLEAN.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.06 seconds
      cpu time            0.04 seconds
```

Cleaning part IV

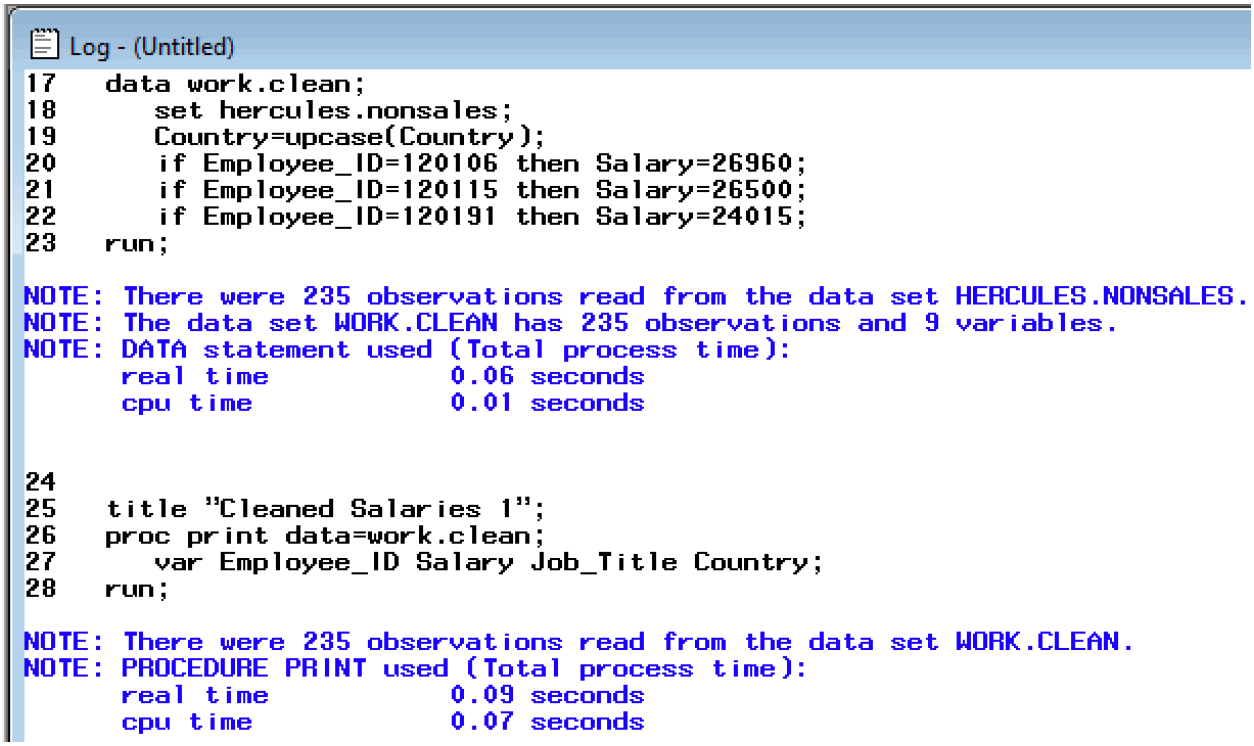
```
data work.clean;
   set hercules.nonsales;
   Country=upcase(Country);
   if Employee_ID=120106 then Salary=26960;
   if Employee_ID=120115 then Salary=26500;
   if Employee_ID=120191 then Salary=24015;
run;

title "Cleaned Salaries 1";
proc print data=work.clean;
   var Employee_ID Salary Job_Title Country;
run;
```

Cleaning part IV output

Cleaned Salaries 1

Obs	Employee_ID	Salary	Job_Title	Country
1	120101	163040	Director	AU
2	120104	46230	Administration Manager	AU
3	120105	27110	Secretary I	AU
4	120106	26960	Office Assistant II	AU
5	120107	30475	Office Assistant III	AU
6	120108	27660	Warehouse Assistant II	AU
7	120108	26495	Warehouse Assistant I	AU
8	120110	28615	Warehouse Assistant III	AU
9	120111	26895	Security Guard II	AU
10	120112	26550		AU

Cleaning part IV log



Cleaning part V

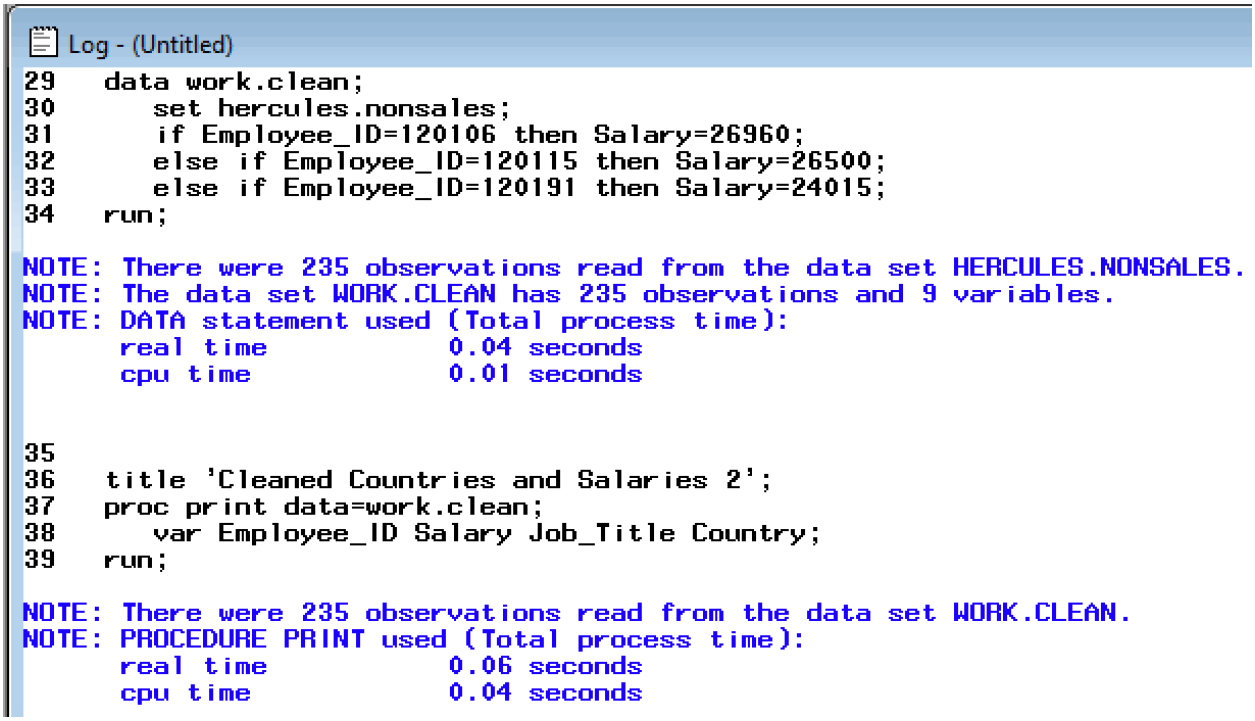
```
data work.clean;
   set hercules.nonsales;
   if Employee_ID=120106 then Salary=26960;
   else if Employee_ID=120115 then Salary=26500;
   else if Employee_ID=120191 then Salary=24015;
run;

title 'Cleaned Countries and Salaries 2';
proc print data=work.clean;
   var Employee_ID Salary Job_Title Country;
run;
```

Cleaning part V output

Cleaned Countries and Salaries 2

Obs	Employee_ID	Salary	Job_Title	Country
1	120101	163040	Director	AU
2	120104	46230	Administration Manager	au
3	120105	27110	Secretary I	AU
4	120106	26960	Office Assistant II	AU
5	120107	30475	Office Assistant III	AU
6	120108	27660	Warehouse Assistant II	AU
7	120108	26495	Warehouse Assistant I	AU
8	120110	28615	Warehouse Assistant III	AU
9	120111	26895	Security Guard II	AU
10	120112	26550		AU

Cleaning part V log



All-in-one DATA step

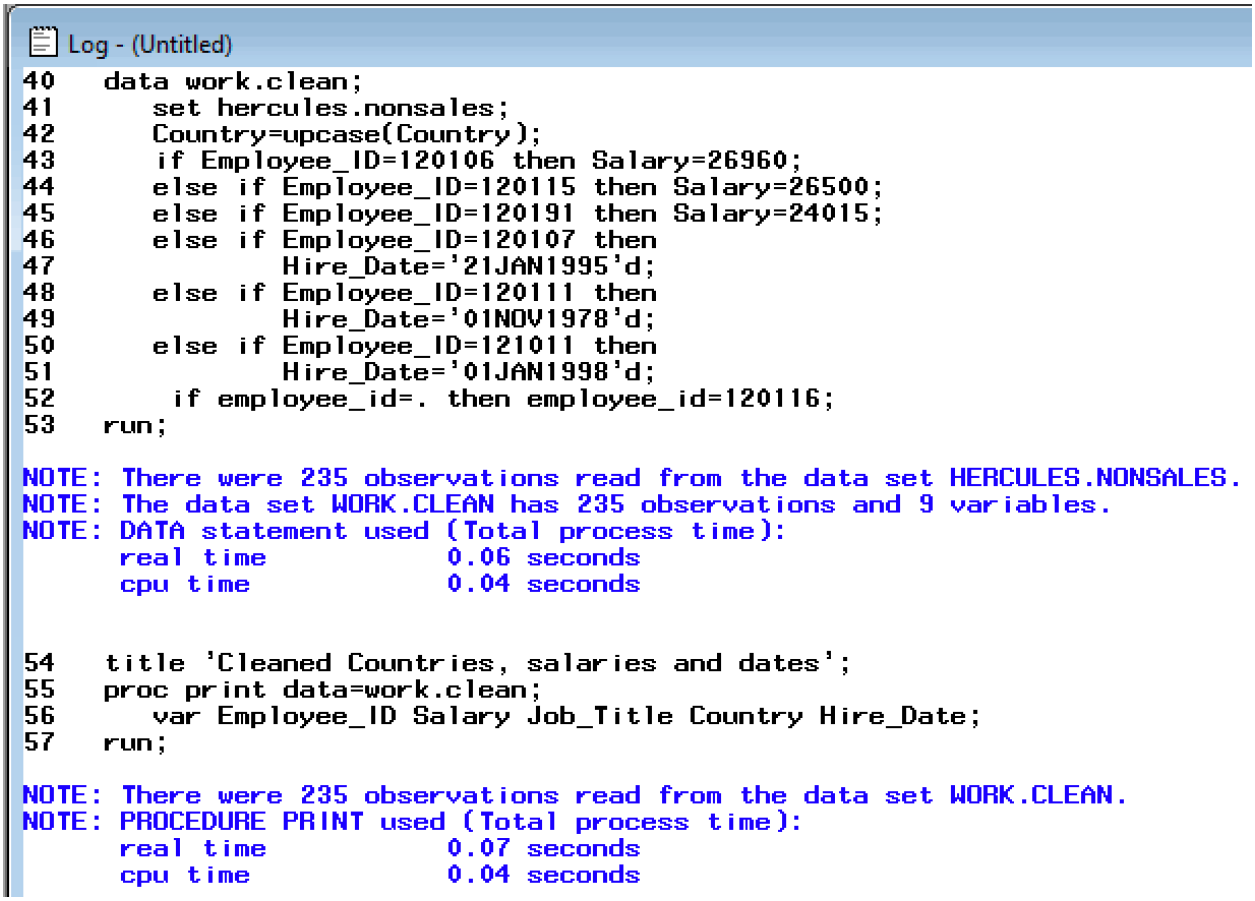
```
data work.clean;
   set hercules.nonsales;
   Country=upcase(Country);
   if Employee_ID=120106 then Salary=26960;
   else if Employee_ID=120115 then Salary=26500;
   else if Employee_ID=120191 then Salary=24015;
   else if Employee_ID=120107 then
           Hire_Date='21JAN1995'd;
   else if Employee_ID=120111 then
           Hire_Date='01NOV1978'd;
   else if Employee_ID=121011 then
           Hire_Date='01JAN1998'd;
    if employee_id=. then employee_id=120116;
run;
title 'Cleaned Countries, salaries and dates';
proc print data=work.clean;
   var Employee_ID Salary Job_Title Country Hire_Date;
run;
title;
```

All-in-one output

Cleaned Countries, salaries and dates

Obs	Employee_ID	Salary	Job_Title	Country	Hire_Date
1	120101	163040	Director	AU	01/07/2003
2	120104	46230	Administration Manager	AU	01/01/1981
3	120105	27110	Secretary I	AU	01/05/1999
4	120106	26960	Office Assistant II	AU	01/01/1974
5	120107	30475	Office Assistant III	AU	21/01/1995
6	120108	27660	Warehouse Assistant II	AU	01/08/2006
7	120108	26495	Warehouse Assistant I	AU	01/10/2006
8	120110	28615	Warehouse Assistant III	AU	01/11/1979
9	120111	26895	Security Guard II	AU	01/11/1978
10	120112	26550		AU	01/07/1990

All-in-one log



```
Log - (Untitled)
40   data work.clean;
41      set hercules.nonsales;
42      Country=upcase(Country);
43      if Employee_ID=120106 then Salary=26960;
44      else if Employee_ID=120115 then Salary=26500;
45      else if Employee_ID=120191 then Salary=24015;
46      else if Employee_ID=120107 then
47              Hire_Date='21JAN1995'd;
48      else if Employee_ID=120111 then
49              Hire_Date='01NOV1978'd;
50      else if Employee_ID=121011 then
51              Hire_Date='01JAN1998'd;
52       if employee_id=. then employee_id=120116;
53   run;

NOTE: There were 235 observations read from the data set HERCULES.NONSALES.
NOTE: The data set WORK.CLEAN has 235 observations and 9 variables.
NOTE: DATA statement used (Total process time):
      real time           0.06 seconds
      cpu time            0.04 seconds


54   title 'Cleaned Countries, salaries and dates';
55   proc print data=work.clean;
56      var Employee_ID Salary Job_Title Country Hire_Date;
57   run;

NOTE: There were 235 observations read from the data set WORK.CLEAN.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.07 seconds
      cpu time            0.04 seconds
```