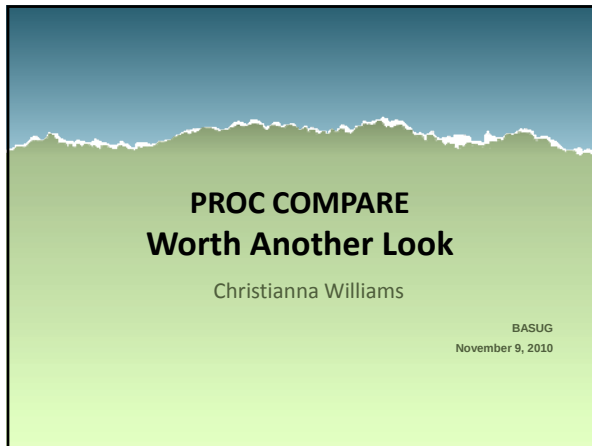




PROC COMPARE Worth Another Look

Christianna Williams

BASUG
November 9, 2010

PROC COMPARE

- Part of Base SAS for a very long time
- Does just what the name implies...
- Worth getting to know for many data management tasks
 - Evaluate newly collected data compared to existing file
 - Test whether updates have occurred as expected
 - Compare algorithms for variable construction/data manipulation
 - Prepare for MERGE/JOIN of two large data sets with some(??) variable names in common

PROC COMPARE 2

PROC COMPARE

- PROC COMPARE can compare data sets in several ways
 - Compare data set attributes
 - Compare variable existence and attributes
 - Compare variable values
- Compare 2 data sets at a time
 - Can also compare a data set to itself
- Are there other ways to do this stuff?
 - Sure, but...

PROC COMPARE 3

The DATA – Nursing Home Compare

- Nursing Home Ratings
 - Includes ~16,000 US Nursing homes
 - Published on internet
 - Updated every month
- For this talk...
 - Washington State Nursing homes only
 - Small subset of variables
 - Comparing ratings data sets for two months
 - July 2009: **WA_JUL2009**
 - February 2010: **WA_FEB2010**

PROC COMPARE 4

Plain Vanilla PROC COMPARE

```
PROC COMPARE
  BASE=in.WA_jul2009
  COMPARE=in.WA_feb2010 ;
RUN;
```

- All I am specifying is my “base” data set and my “compare” data set.
 - BASE = DATA
 - You must specify COMPARE data set*

**one exception... we'll get to that later.*

PROC COMPARE 5

Key features of Output

- DATA Set Summary
- Variables Summary Listing of Common Variables
 - With conflicting types
 - With differing attributes
- Observation Summary
- Values Comparison Summary
- Variables with Unequal Values
- Value Comparison Results

PROC COMPARE 6

Plain Vanilla PROC COMPARE

```
PROC COMPARE
BASE=in.WA_jul2009
COMPARE=in.WA_feb2010 ;
RUN;
```

PROC COMPARE

7

Plain Vanilla PROC COMPARE

The COMPARE Procedure
Comparison of IN.WA_JUL2009 with IN.WA_FEB2010
(Method=EXACT)

Data Set Summary

Dataset	Created	Modified	NVar	NObs	Label
IN.WA_JUL2009	20FEB10	20FEB10	22	234	WA: July 2009
IN.WA_FEB2010	20FEB10	20FEB10	22	231	WA: February 2010

Variables Summary

Number of Variables in Common: 17.
 Number of Variables in IN.WA_JUL2009 but not in IN.WA_FEB2010: 5.
 Number of Variables in IN.WA_FEB2010 but not in IN.WA_JUL2009: 5.
 Number of Variables with Conflicting Types: 1.
 Number of Variables with Differing Attributes: 4.

PROC COMPARE

8

Plain Vanilla PROC COMPARE

Listing of Common Variables with Conflicting Types

Variable	Dataset	Type	Length	Label
hospbased	WA_JUL2009	Char	3	provider in hospital (Y/N)
	WA_FEB2010	Num	8	Provider in Hospital (1)

Listing of Common Variables with Differing Attributes

Variable	Dataset	Type	Length	Format	Label
SurveyDate	WA_JUL2009	Num	8	MMDDYY10.	Cycle 1-Survey Date
	WA_FEB2010	Num	8	DATE9.	Cycle 1-Survey Date
bedcert	WA_JUL2009	Num	4		Total certified beds
	WA_FEB2010	Num	8		Total certified beds
PROVNAME	WA_JUL2009	Char	50		Provider Name
	WA_FEB2010	Char	55		Provider Name
OCCUPY	WA_JUL2009	Num	8		% of beds occupied
	WA_FEB2010	Num	8		restot/bedcert x 100

PROC COMPARE

9

Plain Vanilla PROC COMPARE

Observation Summary

Observation	Base	Compare
First Obs	1	1
First Unequal	1	1
Last Unequal	231	231
Last Match	231	231
Last Obs	234	.

Number of Obs in Common: 231.
 Number of Obs in IN.WA_JUL2009 but not in IN.WA_FEB2010: 3.
 Total Number of Observations Read from IN.WA_JUL2009: 234.
 Total Number of Observations Read from IN.WA_FEB2010: 231.

Number of Obs with Some Compared Variables Unequal: 229.
 Number of Obs with All Compared Variables Equal: 2.

PROC COMPARE

10

Plain Vanilla PROC COMPARE

Values Comparison Summary

Number of Variables Compared with All Observations Equal: 2.
 Number of Variables Compared with Some Observations Unequal: 14.
 Number of Variables with Missing Value Differences: 3.
 Total Number of Values which Compare Unequal: 1942.
 Maximum Difference: 529.

PROC COMPARE

11

Key features of Output

- DATA Set Summary
- Variables Summary Listing of Common Variables
 - With conflicting types
 - With differing attributes
- Observation Summary
- Values Comparison Summary
- Variables with Unequal Values
- Value Comparison Results

Not relevant when not comparing matching observations

What if Data Sets are Identical?

```
PROC COMPARE
BASE = in.WA_jul2009
COMPARE = in.WA_jul2009 ;
RUN;
```

- Sometimes you “hope” that the data sets you are comparing are identical...
— *In this case...they better be!* 😊

PROC COMPARE

13

What if Data Sets are Identical?

Comparison of IN.WA_JUL2009 with IN.WA_JUL2009

Data Set Summary				
Dataset	Created	Modified	NVar	NObs
WA_JUL2009	06FEB10:17:32:51	06FEB10:17:32:51	22	234
WA_JUL2009	06FEB10:17:32:51	06FEB10:17:32:51	22	234

Variables Summary

Number of Variables in Common: 22.

Observation Summary

Number of Observations in Common: 234.

Total Number of Observations Read from IN.WA_JUL2009: 234.

Total Number of Observations Read from IN.WA_JUL2009: 234.

Number of Obs with Some Compared Variables Unequal: 0.

Number of Obs with All Compared Variables Equal: 234.

NOTE: No unequal values found. All values compared exactly equal.

PROC COMPARE

14

When All You Care About is CONTENTS

```
PROC COMPARE
BASE=in.WA_jul2009 COMPARE=in.WA_feb2010
NOVALUES
LISTVAR;
RUN;
```

- NOVALUES suppresses comparison of variable values
- LISTVAR provides lists of variables that differ
 - Existence
 - Attributes

PROC COMPARE

15

Effect of LISTVAR option

Listing of Variables in WA_JUL2009 but not in WA_FEB2010

Variable	Type	Length	Label
revis_c1	Num	8	Cycle 1 - Number of Revisits
revis_c2	Num	8	Cycle 2 - Number of Revisits
revis_c3	Num	8	Cycle 3 - Number of Revisits
defscore_rank	Num	8	Rank for Variable defscore
zipcode	Char	5	Zip code of Provider

Listing of Variables in WA_FEB2010 but not in WA_JUL2009

Variable	Type	Length	Label
defscore0104	Num	8	Total Wtd Survey Deficiency Score
def5star0104	Num	8	5-star ranking system for total score
ZIP	Char	5	Zip code of Provider
COUNTY	Char	3	SSA GEOGRAPHIC CODE of provider
CONTROL	Char	2	Nature of organization for provider

PROC COMPARE

16

Even More Concise – BRIEF option

```
PROC COMPARE
BASE=in.WA_jul2009 COMPARE=in.WA_feb2010
NOVALUES
BRIEF;
RUN;
```

- BRIEF suppresses several pieces
 - DATA Set Summary
 - Variables Summary
 - Observation Summary
 - Values Comparison Summary

PROC COMPARE

17

Effect of BRIEF option

- Short and sweet!

The COMPARE Procedure
Comparison of IN.WA_JUL2009 with IN.WA_FEB2010

NOTE: Data set IN.WA_JUL2009 contains 3 observations not in IN.WA_FEB2010.

NOTE: Values of the following 14 variables compare unequal: PROVNUM SurveyDate numcycles defscore def5star bedcert MDS5star rnstf5 staff5star PROVNAME CITY RESTOT OCCUPY Comp5star

PROC COMPARE

18

When All You Care About is CONTENTS - Better

```
PROC COMPARE
  BASE=in.WA_jul2009 COMPARE=in.WA_feb2010
  NOVALUES BRIEF LISTVAR;
RUN;
```

- Listing of variables in one data set and not the other
- Listing of common variables with conflicting types
- Listing of common variables with differing attributes
- NOTES (log):
 - Number of observations in one data set and not the other
 - List of variables with unequal values

19

When All You Care About is CONTENTS - LOG

- Yet another option

```
PROC COMPARE
  BASE=in.WA_jul2009
  COMPARE=in.WA_feb2010
  NOVALUES WARNING NOPRINT;
RUN;
```

- WARNING option generates a warning about any differences found (to the LOG)
- NOPRINT suppresses all printed output

PROC COMPARE

20

When All You Care About is CONTENTS - LOG

```
WARNING: Data set WA_JUL2009 contains 5 variables not
in WA_FEB2010.
WARNING: Data set WA_FEB2010 contains 5 variables not
in WA_JUL2009.
WARNING: 1 variables are numeric in one data set but
character in the other.
WARNING: 4 variables have conflicting attributes in
the data sets.
WARNING: Data set WA_JUL2009 contains 3 obs not in
WA_FEB2010.
WARNING: Values of the following 14 variables compare
unequal: PROVNUM SurveyDate numcycles defscore
def5star MDS5star rnstf5 staff5star PROVNAME CITY
BEDCERT RESTOT OCCUPY Comp5star
WARNING: The data sets WA_JUL2009 and WA_FEB2010
contain unequal values.
```

Also an ERROR option

21

PROC COMPARE As Match-MERGE

- Use the ID statement
- Focus more on comparing *Values*

```
PROC COMPARE BASE=in.WA_jul2009
  COMPARE=in.WA_feb2010
  MAXPRINT = (2,20);
  ID provnum ;
RUN;
```

of Value differences
for any one variable

Total # of value
differences

PROC COMPARE

22

PROC COMPARE As Match-MERGE

- Sections of output that don't depend on match are identical to what we saw earlier
 - DATA Set Summary
 - List of Common Variables with Conflicting Types
 - List of Common Variables with Differing Attributes

PROC COMPARE

23

PROC COMPARE with ID statement

- Variables Summary shows ID variable(s)

```
Variables Summary
Number of Vars in Common: 17.
Number of Vars in WA_JUL2009 but not in WA_FEB2010: 5.
Number of Vars in WA_FEB2010 but not in WA_JUL2009: 5.
Number of Vars with Conflicting Types: 1.
Number of Vars with Differing Attributes: 4.

Number of ID Variables: 1.
```

PROC COMPARE

24

PROC COMPARE with ID statement

Observation Summary

Observation	Base	Compare	ID
First Obs	1	1	PROVNUM=505004
First Unequal	1	1	PROVNUM=505004
Last Unequal	234	231	PROVNUM=50A263
Last Obs	234	231	PROVNUM=50A263

Number of Observations in Common: 229.
 Number of Obs in WA_JUL2009 but not in WA_FEB2010: 5.
 Number of Obs in WA_FEB2010 but not in WA_JUL2009: 2.
 Total Number of Obs Read from WA_JUL2009: 234.
 Total Number of Obs Read from WA_FEB2010: 231.
 Number of Obs with Some Compared Vars Unequal: 226.
 Number of Obs with All Compared Vars Equal: 3.

PROC COMPARE 25

PROC COMPARE with ID statement

Values Comparison Summary

Number of Vars Compared with All Obs Equal: 4.
 Number of Vars Compared with Some Obs Unequal: 11.
 Number of Vars with Missing Value Differences: 2.

Total Number of Values that Compare Unequal: 1027.

Maximum Difference: 467.

PROC COMPARE 26

PROC COMPARE with ID statement

Summary Info on Value Differences

Variables with Unequal Values

Variable	Type	Len1	Len2	Ndif	MaxDif	MissDif
SurveyDate	NUM	8	8	131	467	0
defscore	NUM	8	8	214	238	0
def5star	NUM	8	8	64	2.000	0
bedcert	NUM	4	8	15	40.000	0
MDS5star	NUM	4	4	123	3.000	0
rnstf5	NUM	8	8	69	2.000	4
staff5star	NUM	8	8	73	3.000	4
PROVNAME	CHAR	50	55	4		0
RESTOT	NUM	8	8	118	24.000	0
OCCUPY	NUM	8	8	122	30.000	0
Comp5star	NUM	8	8	94	3.000	0

PROC COMPARE 27

PROC COMPARE with ID statement

Value Comparison Results for Variables

Cycle 1 - Survey Date					
	Base	Compare			
PROVNUM	SurveyDate	SurveyDate	Diff.	% Diff	
505016	08/07/08	04JUN2009	301.0000	1.6957	
505024	10/23/08	28AUG2009	309.0000	1.7332	

NOTE: The MAXPRINT=2 printing limit has been reached for the variable SurveyDate. No more values will be printed for this comparison.

Total wtd Survey Deficiency Score (3 cycles)					
	Base	Compare			
PROVNUM	defscore	defscore	Diff.	% Diff	
505004	174.3333	169.6667	-4.6667	-2.6769	
505009	81.8333	59.6667	-22.1667	-27.0876	

PROC COMPARE 28

Notes about ID statement

- Both data sets should be sorted by ID variable(s)
- You can use NOTSORTED but...
 - Then COMPARE does 1:1 matching
 - If not 1:1 match on ID – an ERROR is generated
- If not sorted and NOTSORTED not specified...
 - WARNING generated
 - ID statement ignored

PROC COMPARE 29

ID Statement + LISTOBS and LISTEQUALVAR

```
PROC COMPARE BASE=in.wa_jul2009
COMPARE=in.wa_feb2010
LISTOBS LISTEQUALVAR NOVALUES;

ID provnum ;
RUN;
```

- LISTOBS will show observations unique to each data set
- LISTEQUALVAR will list variables for which all observations have equal values

PROC COMPARE 30

LISTOBS option**Comparison Results for Observations**

```
Obs 112 in WA_JUL2009 not found in WA_FEB2010:PROVNUM=505333.
Obs 118 in WA_JUL2009 not found in WA_FEB2010:PROVNUM=505345.
Obs 171 in WA_JUL2009 not found in WA_FEB2010:PROVNUM=505426.
Obs 206 in WA_JUL2009 not found in WA_FEB2010:PROVNUM=505486.
Obs 225 in WA_FEB2010 not found in WA_JUL2009:PROVNUM=505518.
Obs 226 in WA_FEB2010 not found in WA_JUL2009:PROVNUM=505519.
Obs 231 in WA_JUL2009 not found in WA_FEB2010:PROVNUM=50A195.
```

PROC COMPARE

31

LISTEQUALVAR option**Variables with All Equal Values**

Variable	Type	Len1	Len2	Label
STATE	CHAR	2	2	State abbrev of Provider
numcycles	NUM	8	8	# cycles of deficiency data
CITY	CHAR	28	28	City of Provider
statename	CHAR	20	20	State Name

PROC COMPARE

32

The LISTxxxx Options

LISTALL	List all VARIABLES & OBSERVATIONS unique to one data set
LISTVAR	List all VARIABLES unique to one data set
LISTOBS	List all OBSERVATIONS unique to one data set
LISTBASE	List all variables and observations in BASE DATA SET and not in Compare data set
LISTCOMP	List all variables and observations in COMPARE DATA SET and not in BASE data set
LISTBASEVAR	List all variables in BASE DATA SET and not in Compare data set
LISTCOMPVAR	List all variables in COMPARE DATA SET and not in BASE data set
LISTBASEOBS	List all observations in BASE DATA SET and not in Compare data set
LISTCOMPOBS	List all observation in COMPARE DATA SET and not in BASE data set

33

COMPARE-ing Only Selected Variables

```
PROC COMPARE BASE=in.WA_jul2009
COMPARE=in.WA_feb2010
ALLSTATS;
ID provnum ;
VAR def5star staff5star;
RUN;
```

- You may just care about differences in a few variables – VAR statement
- ALLSTATS will provide summary statistics

PROC COMPARE

34

PROC COMPARE with VAR statement

- Variables Summary shows VAR variable(s)

Variables Summary

```
Number of Vars in Common: 17.
Number of Vars in WA_JUL2009 but not in WA_FEB2010: 5.
Number of Vars in WA_FEB2010 but not in WA_JUL2009: 5.
Number of Vars with Conflicting Types: 1.
Number of Vars with Differing Attributes: 4.
Number of ID Variables: 1.
Number of VAR Statement Variables: 2.
```

PROC COMPARE

35

PROC COMPARE with VAR statement

Observation	Summary	Base	Compare	ID
First Obs		1	1	PROVNUM=505004
First Unequal		2	2	PROVNUM=505009
Last Unequal		234	231	PROVNUM=50A263
Last Obs		234	231	PROVNUM=50A263

```
Number of Observations in Common: 229.
Number of Obs in WA_JUL2009 but not in WA_FEB2010: 5.
Number of Obs in WA_FEB2010 but not in WA_JUL2009: 2.
Total Number of Obs Read from WA_JUL2009: 234.
Total Number of Obs Read from WA_FEB2010: 231.
Number of Obs with Some Compared Vars Unequal: 108.
Number of Obs with All Compared Vars Equal: 121.
```

PROC COMPARE

36

PROC COMPARE with VAR statement

- All Comparisons are based on just VARs

Values Comparison Summary**w/o VAR**

Number of Vars Compared with All Obs Equal: 4.
 Number of Vars Compared with Some Obs Unequal: 11.
 Number of Vars with Missing Value Differences: 2.
 Total Number of Values that Compare Unequal: 1027.
 Maximum Difference: 467.

Values Comparison Summary**w/ VAR**

Number of Vars Compared with All Obs Equal: 0.
 Number of Vars Compared with Some Obs Unequal: 2.
 Number of Vars with Missing Value Differences: 1.
 Total Number of Values that Compare Unequal: 137.
 Maximum Difference: 3.

ALLSTATS option**Value Comparison Results for Variables**

5-star ranking for Staffing					
PROVNUM	Base	Compare			
	staff5star	staff5star	Diff.	% Diff	
505016	3.0000	4.0000	1.0000	33.3333	
505027	5.0000	4.0000	-1.0000	-20.0000	
505042	1.0000	2.0000	1.0000	100.0000	
N	224	224	222	222	
Mean	3.2455	3.3705	0.1081	11.3889	
Std	1.1704	1.0968	0.7999	45.7594	
Max	5.0000	5.0000	3.0000	300.0000	
Min	1.0000	1.0000	-3.0000	-75.0000	
StdErr	0.0782	0.0733	0.0537	3.0712	
t	41.5012	45.9948	2.0137	3.7083	
Prob> t	<.0001	<.0001	0.0453	0.0003	
Ndif	73	31.878%			
DifMeans	3.851%	3.709%	0.1250		
r, rsq	0.753	0.566			

Options for Handling Missing Values

NOMISSBASE	Missing values in BASE data set judged equal to ANY value in the comparison data set
NOMISSCOMP	Missing values in COMPARE data set judged equal to ANY value in the base data set
NOMISSING	Missing values in either data set judged equal to ANY value in the other data set

- NOMISSCOMP useful prior to UPDATE
 – BASE as Master, COMPARE as Transaction

PROC COMPARE

39

VAR and WITH Statements

```
PROC COMPARE
BASE=in.WA_jul2009
COMPARE=in.WA_feb2010 ALLSTATS;
ID provnum ;
VAR zipcode bedcert;
WITH zip bedcert;
RUN;
```

- Allows comparison of variables with different names
- Variables compared 1:1
- Could compare variables on SAME dataset

PROC COMPARE

40

VAR & WITH statements

- Variables summary gives count of VAR & WITH variables

Variables Summary

Number of Vars in Common: 17.
 Number of Vars in WA_JUL2009 but not in WA_FEB2010: 5.
 Number of Vars in WA_FEB2010 but not in WA_JUL2009: 5.
 Number of Vars with Conflicting Types: 1.
 Number of Vars with Differing Attributes: 4.
 Number of ID Variables: 1.
Number of VAR Statement Variables: 2.
Number of WITH Statement Variables: 2.

PROC COMPARE

41

VAR & WITH statements

- Shows which BASE variable is being compared to which COMPARE variable

Values Comparison Summary

Number of Vars Compared with All Obs Equal: 0.
 Number of Vars Compared with Some Obs Unequal: 2.
 Total Number of Values which Compare Unequal: 19.
 Maximum Difference: 40.
 All Variables Compared have Unequal Values

Variable	Type	Len	Compare	Len	Ndif	MaxDif
zipcode	CHAR	5	ZIP	5	4	
bedcert	NUM	4	BEDCERT	8	15	40.000

VAR & WITH statements

Values Comparison Results for Variables

		Zip code of Provider	
		Base Value	Compare Value
PROVNUM	zipcode	ZIP	
505211	98371	98372	
505280	98055	98057	
505296	98226	98225	
505469	98101	98104	

		Total certified beds	
		Tot medicare and/or medicaid certified beds	
		Base	Compare
PROVNUM	bedcert	BEDCERT	Diff. % Diff
505004	180.0000	140.0000	-40.0000 -22.2222
505010	95.0000	97.0000	2.0000 2.1053
505016	101.0000	105.0000	-16.0000 -13.2231

No STATS for Character vars

PROC COMPARE

43

More on VAR and WITH Statements

```
PROC COMPARE BASE=in.WA_jul2009
  COMPARE=in.WA_feb2010 ALLSTATS;
ID provnum ;
VAR bedcert bedcert;
WITH bedcert restot;
RUN;
```

- Compare one variable to two (or more)

PROC COMPARE

44

More on VAR and WITH Statements

- Compare variables in the same data set

```
PROC COMPARE BASE=WA_jul2009 NOVALUES;
VAR
  def5star defscore def5star;
WITH
  def5star0104 defscore0104 staff5star;
RUN;
```

PROC COMPARE

45

Generate an Observation-Oriented Report

- All previous examples have been *Variable-Oriented*

```
PROC COMPARE BASE=WA_jul2009
  COMPARE=WA_feb2010
  NOSUMMARY TRANSPOSE;
ID PROVNUM ;
VAR def5star staff5star ;
RUN;
```

PROC COMPARE

46

TRANSPOSE Option

PROVNUM=505009:				
Variable	Base Value	Compare	Diff.	% Diff
def5star	2.00000	3.00000	1.00000	50.00000

PROVNUM=505016:				
Variable	Base Value	Compare	Diff.	% Diff
staff5star	3.00000	4.00000	1.00000	33.33333

PROVNUM=505017:				
Variable	Base Value	Compare	Diff.	% Diff
def5star	4.00000	3.00000	-1.00000	-25.00000

PROVNUM=505027:				
Variable	Base Value	Compare	Diff.	% Diff
def5star	3.00000	4.00000	1.00000	33.33333
staff5star	5.00000	4.00000	-1.00000	-20.00000

PROVNUM=505350:				
Variable	Base Value	Compare	Diff.	% Diff
staff5star	.	3.00000	.	.

PROC COMPARE

47

Generate an Output Data Set of Differences

- An alternative for an observation-oriented report

```
PROC COMPARE BASE=in.WA_jul2009
  COMPARE=in.WA_feb2010
  OUT=outcomp
  OUTNOEQUAL OUTBASE OUTCOMP OUTDIF
  NOPRINT ;
ID provnum;
VAR def5star staff5star provname;
RUN;
```

Note Inclusion of Character variable

PROC COMPARE

48

Output Data Set of Differences

PROVNUM	_TYPE_	_OBS_	def5star	staff5	PROVNAME
505009	BASE	2	2	2	PARK RIDGE CARE CENTER
	COMPARE	2	3	2	PARK RIDGE CARE CENTER
	DIF	2	1	E	
505016	BASE	4	3	3	GRAYS HARBOR HEALTH&REHAB
	COMPARE	4	3	4	GRAYS HARBOR HEALTH&REHAB
	DIF	4	E	1	
505027	BASE	8	3	5	HEARTHSTONE, THE
	COMPARE	8	4	4	HEARTHSTONE, THE
	DIF	8	1	-1	
505033	BASE	9	4	4	ROCKWOOD SOUTH HILL
	COMPARE	9	4	4	ROCKWOOD RETIREMENT COMMUNITY
	DIF	9	E	E	XXXXXXXXXXXX

- OUTBASE, OUTCOMP, OUTDIF
- OUTNOEQUAL

PROC COMPARE

49

Other features to Explore

- ODS Objects
- OUTSTATS data set
- BY Statement
- CRITERION, METHOD & FUZZ
- SYSINFO macro variable

PROC COMPARE

50

Conclusions

- PROC COMPARE can be very handy
 - Data Cleaning
 - Preparing for Joins / Updates
 - Getting to know your (NEW) data
- An OLDIE but GOODIE!

PROC COMPARE

51

THANKS!

Christianna.S.Williams@gmail.com