DATA AND METADATA EXPLORATION USING SAS

Wisconsin Illinois SAS User Group

Charu Shankar
SAS Institute Inc. Canada
28 May 2014

PROGRAMMER RULE # 1

Know thy data

Investigators are interested in examining the occurrence of Type 2 diabetes in women of Pima Indian heritage who are at least 21 years old.

THE POPULATION OF INTEREST

The diagnostic, binary-valued, variable investigated is whether the patient shows signs of diabetes according to the World Health Organization (i.e, if the 2-hour post-load plasma glucose was at least 200 mg/dl at any survey examination or during routine medical care).

Metadata unlocks the value of data, and therefore requires management attention.” [Gartner 2011]

80-20 RULE

THE VARIABLES

Variable Name	Description
glucose	glucose
dbp	diastolic blood pressure
triceps	tricep skin fold thickness
insulin	2-hour serum insulin
pedigree	diabetes pedigree
Diabetes	1 = tested positive for diabetes 0 = tested negative for diabetes
ID	identification number

SAS data

Pregnancies	number of times pregnant
BMI	body mass index
age	age
ID	identification number

Excel data
Pima Indians
Diabetes.xls

AGENDA

- Part 1- METADATA - quick and easy way to know your data
- Part 2 - Powerful PROC SQL Dictionary Tables
- Part 3 - Analytical SAS Procedures to know your data
- Close
- Q&A

PART 1: METADATA
A. THE SAS® EXPLORER

- There is a lot of information available to you with a simple click of the mouse … OK, sometimes a double-click …
- Information about a SAS file.
- Information about the individual fields that make up the file.

PART 1 : METADATA
B. PROC CONTENTS …

PROC CONTENTS – looking at a single table

proc contents data=diabetes.pima out=test;

run;.

PROC CONTENTS OUTPUT

The CONTENTS Procedure

Data Set Name	DIABETES.PIMA	Observations	768
Member Type	DATA	Variables	10
Engine	V9	Indexes	0
Created	Thursday, January 31, 2013 11:08:13 AM	Observation Length	80
Last Modified	Friday, June 20, 2014 01:50:48 PM	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information

Data Set Page Size	8192
Number of Data Set Pages	8
First Data Page	1
Max Obs per Page	101
Obs in First Data Page	77
Number of Data Set Repairs	0
File name	C:\Users\cancxs\Desktop\wiilsu, jun 2014\managing the 80-20 rule with SAS\data\pima.sas7bdat
Release Created	9.0301M0
Host Created	X64_7PRO

Alphabetic List of Variables and Attributes

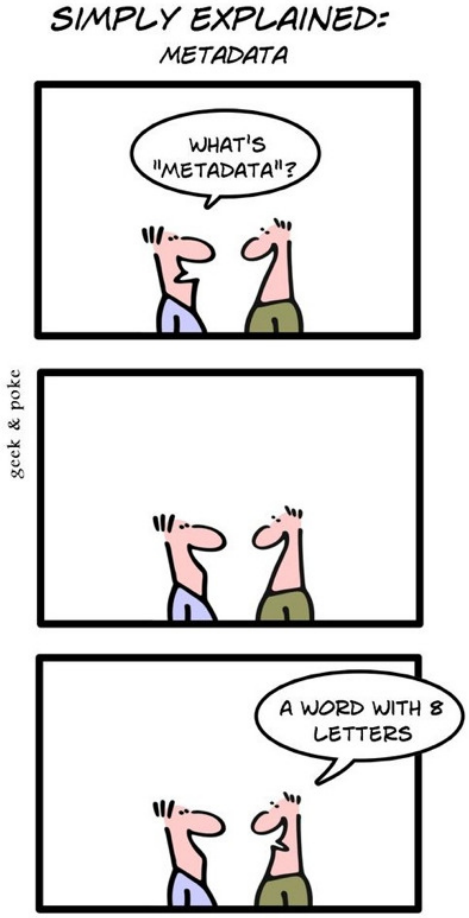
#	Variable	Type	Len
9	Age	Num	8
7	BMI	Num	8
10	Class	Num	8
4	DBP	Num	8
8	DiabetesPedigree	Num	8
6	Insulin	Num	8
3	PlasmaGluc	Num	8
2	Pregnancies	Num	8
5	Triceps	Num	8
1	id	Num	8

PART 2: METADATA PROC SQL DICTIONARY TABLES

1. Look up dictionary tables easily in SAS

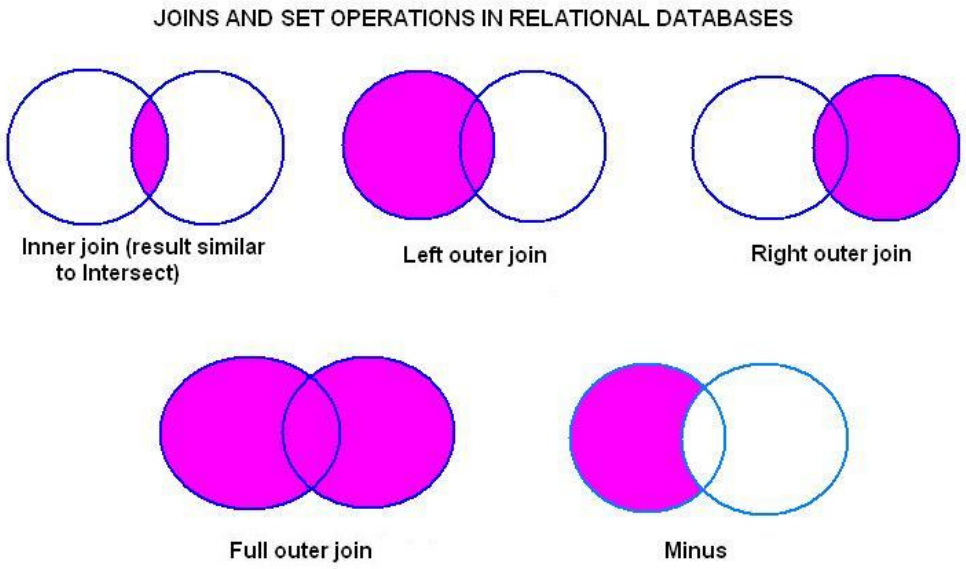
PART 2: METADATA PROC SQL DICTIONARY TABLES

2. Examine Metadata



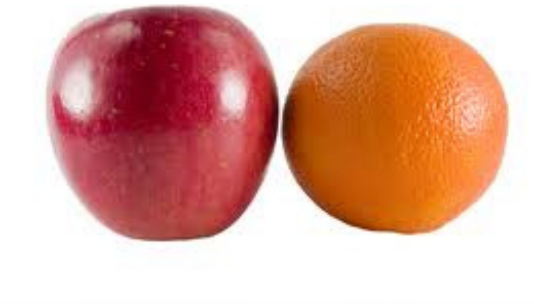
PART 2: METADATA PROC SQL DICTIONARY TABLES

3. Investigate common columns for joins



PART 2: METADATA PROC SQL DICTIONARY TABLES

4. An efficiency question-PROC SQL or SAS datastep?



PART 2: METADATA PROC SQL DICTIONARY TABLES

5. Locate changed variable names

PART 2: METADATA PROC SQL DICTIONARY TABLES

6. Reorder variables in dataset

PART 2: METADATA PROC SQL DICTIONARY TABLES

7. Isolate variable type conflicts

PART 3: DATA - PROC FREQ

Do you know your data, I mean really..

How do you find duplicates?

```
PROC FREQ   DATA=diabetes.pima/noprint ;
TABLES id /out=dupid(where=(count > 1));
run;
```

PART 3: DATA - PROC MEANS

What more do you want to know?

What's the largest value, smallest?

proc means data=diabetes.pima max min mean;
run;.

QUESTIONS

Wondering ….

Why did the Arizona Pima exceed diabetes rates of Mexican Pima by 5 times?

GREAT REFERENCES

Data on Pima Indians with diabetes symptoms is also available on the internet.

Topic	Web link
WHO Diabetes Definition	http://whqlibdoc.who.int/publications/2006/9241594934_eng.pdf
Stages of Diabetes	http://home.comcast.net/~cnmpat/bloodsugarstages.htm
Diabetes Diagnosis	http://diabetes.niddk.nih.gov/dm/pubs/diagnosis/index.htm
Diabetes Symptoms	http://diabetes.niddk.nih.gov/dm/pubs/insulinresistance/#symptoms
Sugar in the corn, sugar in the blood	http://www.youtube.com/watch?v=pN4HqWRybwk
New York Times Video	http://video.nytimes.com/video/2008/07/30/us/1194817478153/water-returns-to-the-pima.html

THANKS FOR ATTENDING
QUESTIONS???

Charu Shankar, SAS institute Inc.

BLOG http://blogs.sas.com/content/sastraining/author/charushankar/

LINKEDIN http://ca.linkedin.com/in/charushankar

TWITTER https://twitter.com/CharuSAS

EMAIL Charu.Shankar@sas.com