PROC SUMMARY VS. RETAIN BY

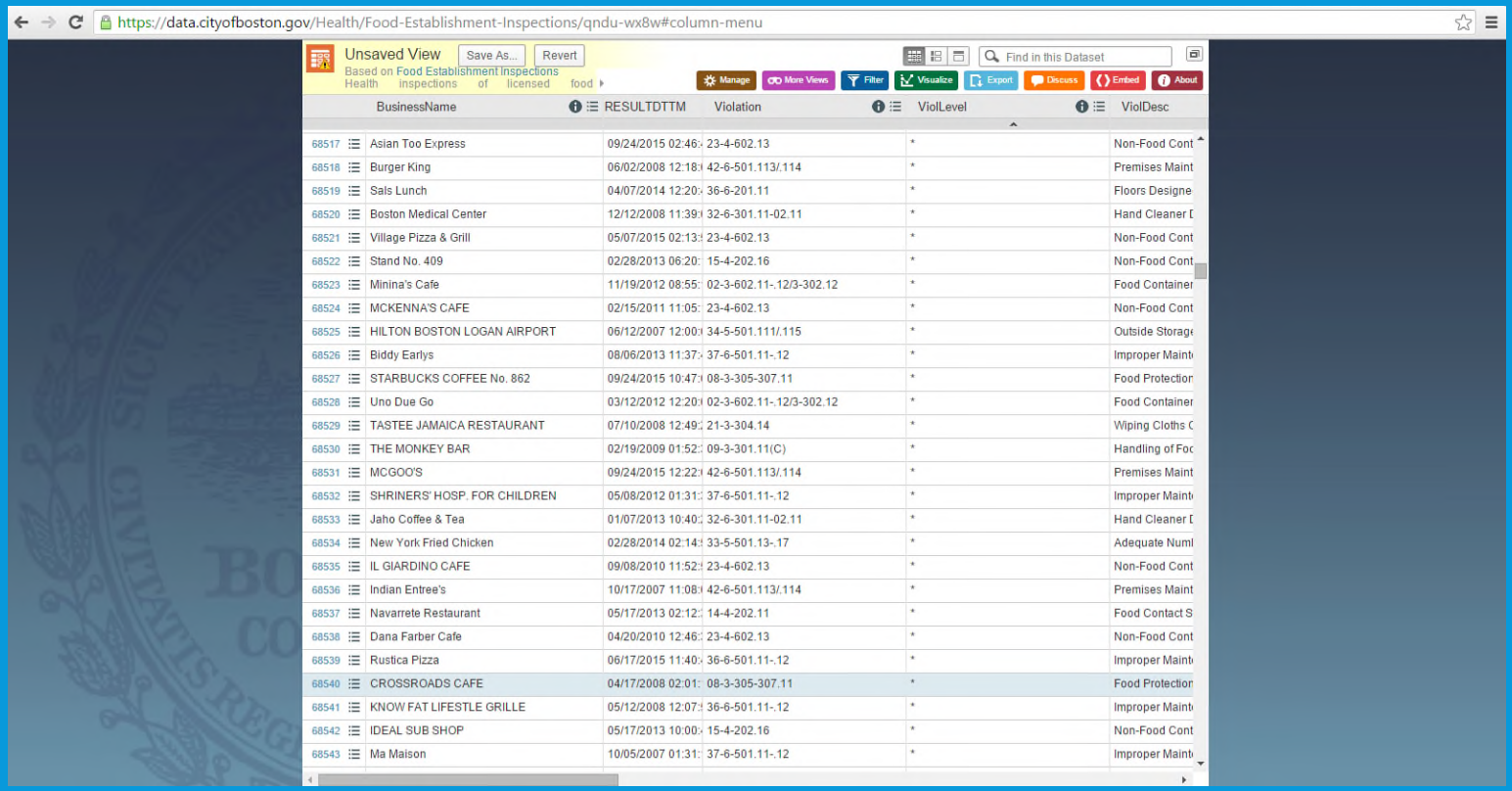
SAS Techniques for Summarizing Data

Neal Jawadekar

December 2015

BIG DATA FOR SOCIAL GOOD

CITY OF BOSTON – OPEN DATA



	BusinessName	RESULTDTTM	Violation	ViolLevel	ViolDesc
68517	Asian Too Express	09/24/2015 02:46:	23-4-602.13	*	Non-Food Cont
68518	Burger King	06/02/2008 12:18:	42-6-501.113/.114	*	Premises Maint
68519	Sals Lunch	04/07/2014 12:20:	36-6-201.11	*	Floors Designe
68520	Boston Medical Center	12/12/2008 11:39:	32-6-301.11-02.11	*	Hand Cleaner D
68521	Village Pizza & Grill	05/07/2015 02:13:	23-4-602.13	*	Non-Food Cont
68522	Stand No. 409	02/28/2013 06:20:	15-4-202.16	*	Non-Food Cont
68523	Minina's Cafe	11/19/2012 08:55:	02-3-602.11-.12/3-302.12	*	Food Container
68524	MCKENNA'S CAFE	02/15/2011 11:05:	23-4-602.13	*	Non-Food Cont
68525	HILTON BOSTON LOGAN AIRPORT	06/12/2007 12:00:	34-5-501.111/.115	*	Outside Storage
68526	Biddy Earlys	08/06/2013 11:37:	37-6-501.11-.12	*	Improper Maint
68527	STARBUCKS COFFEE No. 862	09/24/2015 10:47:	08-3-305-307.11	*	Food Protection
68528	Uno Due Go	03/12/2012 12:20:	02-3-602.11-.12/3-302.12	*	Food Container
68529	TASTEE JAMAICA RESTAURANT	07/10/2008 12:49:	21-3-304.14	*	Wiping Cloths C
68530	THE MONKEY BAR	02/19/2009 01:52:	09-3-301.11(C)	*	Handling of Foc
68531	MCGOO'S	09/24/2015 12:22:	42-6-501.113/.114	*	Premises Maint
68532	SHRINERS' HOSP. FOR CHILDREN	05/08/2012 01:31:	37-6-501.11-.12	*	Improper Maint
68533	Jaho Coffee & Tea	01/07/2013 10:40:	32-6-301.11-02.11	*	Hand Cleaner D
68534	New York Fried Chicken	02/28/2014 02:14:	33-5-501.13-.17	*	Adequate Numl
68535	IL GIARDINO CAFE	09/08/2010 11:52:	23-4-602.13	*	Non-Food Cont
68536	Indian Entree's	10/17/2007 11:08:	42-6-501.113/.114	*	Premises Maint
68537	Navarrete Restaurant	05/17/2013 02:12:	14-4-202.11	*	Food Contact S
68538	Dana Farber Cafe	04/20/2010 12:46:	23-4-602.13	*	Non-Food Cont
68539	Rustica Pizza	06/17/2015 11:40:	36-6-501.11-.12	*	Improper Maint
68540	CROSSROADS CAFE	04/17/2008 02:01:	08-3-305-307.11	*	Food Protection
68541	KNOW FAT LIFESTLE GRILLE	05/12/2008 12:07:	36-6-501.11-.12	*	Improper Maint
68542	IDEAL SUB SHOP	05/17/2013 10:00:	15-4-202.16	*	Non-Food Cont
68543	Ma Maison	10/05/2007 01:31:	37-6-501.11-.12	*	Improper Maint

BACKGROUND

- "Proc summary" and "retain by" are SAS methods that can be used to summarize information across a variable, which appears on multiple lines

- Both could be useful in analyzing Boston's publicly-available food inspection data

FOOD INSPECTION DATA

- In the food inspection data, there are multiple rows per restaurant inspection. However, you need to summarize this information to count the TOTAL number of inspection violations, per type, for each inspection.

DATASET

ID	Restaurant	Viollevel_1	Viollevel_2	Viollevel_3	Inspection Date
2	Subway	1	0	0	10/05/2013
1	Applebee's	0	0	1	05/07/2011
3	McDonald's	0	0	1	04/04/2012
2	Subway	0	0	1	10/05/2013
4	Taco Bell	0	1	0	02/17/2014
2	Subway	0	0	1	10/05/2013
1	Applebee's	1	0	0	05/07/2011

- "Proc summary" or "Retain by" could both help to summarize this data

ADVANTAGES OF EACH METHOD

- Each method has its own pros and cons

Proc Summary	Retain By
Succinct code	More code required
Intended for outputting summary stats of interest	Allows you easily retain variables that you may want to keep
Has many options (sum, mean, max, min, etc.)	Code intuitively helps you understand how SAS works

STEPS – RETAIN BY

(1) Sort

(2) Retain by

(3) Keep a count

(4) Output last record (of each ID)

SAS SYNTAX – RETAIN BY

#1

```
PROC SORT data = food_violation;
by ID;
run;
```

#2

```
data food_violation_counts;
set food_violation;
RETAIN ID  count_3    count_2    count_1 ;
BY ID;
if first.ID then
 DO;
  count_1 = 0;
  count_2 = 0;
  count_3 = 0;
 end;
```

#3

```
/* Counting number of each violation level */
COUNT_1 = count_1 + viollevel_1;
COUNT_2 = count_2 + viollevel_2;
COUNT_3 = count_3 + viollevel_3;
```

#4

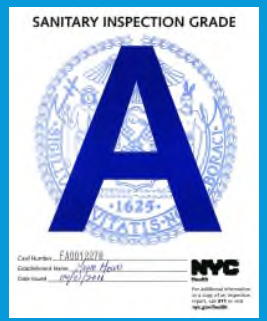
```
IF LAST.ID THEN OUTPUT;

keep ID count_3 count_2 count_1   Date
BusinessName   Address City State Zip;
run;
```

RESULT OF RETAIN BY

Restaurant	Count_1	Count_2	Count_3	Date	Address	City	State	Zip
Applebee's	1	0	1	5/7/2011	1 Elm St.	Boston	MA	02111
McDonald's	0	0	1	4/4/2012	2 Elm St.	Boston	MA	02111
Taco Bell	0	1	0	2/17/2014	3 Elm St.	Boston	MA	02111
Subway	1	0	2	10/5/2013	4 Elm St.	Boston	MA	02111

Perhaps you could use this information to give inspection grades!



PROC SUMMARY

- Proc summary is more succinct, but is more useful for outputting summary statistics
- No sort is needed before proc summary!

```
Proc summary data = food_violation1;
CLASS ID;
VAR  count_3    count_2    count_1;
Output out = food_violation2   sum= ;
run;
```

RESULT OF PROC SUMMARY

ID	Count_1	Count_2	Count_3
1	1	0	1
3	0	0	1
4	0	1	0
2	1	0	2

You could also use this to assign inspection grades!

CONCLUSION

- Your decision to use "proc summary" or "retain by" depends on what you're trying to accomplish, as well as your comfort level
- There are many other utilities/options of proc summary and the retain statement. Check them out!

QUESTIONS

njawadekar@predilytics.com