Paper 270-2011

Mix & Match: Diversity in displaying data

Melissa Hill, MPH, Julie Kezik MS, Yale University, New Haven, CT, USA

ABSTRACT

A programmer is often asked to 'run some frequencies' or 'put together a quick report' in order to share results with a group. Just as every scientist has a preferred output style (ie: graph, table, figure, list, etc.) every programmer has a preferred way of getting there. The results of posing this question to 5 colleagues produced a variety of approaches including the use of PROC MEANS, PROC TABULATE, PROC FREQ and PROC BOXPLOT as well as visualization techniques in JMP.

INTRODUCTION

The purpose of this paper is to illustrate the range of procedures and syntax that can be used to create output that is visually diverse yet contains the same information. Five programmers in our group were provided with the same dataset and three tasks. In this paper we compare and contrast their approaches and finished products.

THE EXPERIMENT

THE PARTICIPANTS

Five SAS users at the Yale CPPEE participated in this exercise.

SAS User ID	Title	Primary use of SAS in their daily work	Years of experience with SAS
A	Programmer Analyst	Dataset maintenance and progress report creation for active field studies	1.5
B	Programmer Analyst	Dataset manipulation and cleaning for analysis of completed fieldwork	2
C	Research Associate	Database planning for new research and analysis support	8
D	Biostatistician	Data cleaning, processing and analysis support	3.5
E	Research Scientist	Complex data analysis	35

Table 1. Description of participants.

THE DATASET

Each participant was sent an email with identical instructions (Display 1) and dataset (Output 1).

Using any SAS procedure or tool, create a brief report (plot, table, frequency listing, etc.) that explores the following three aspects of the dataset provided

1. $PM_{2.5}$ measurements for each location (SITE_ID)
2. $PM_{2.5}$ measurements for each monitoring period (MP)
3. Trends in $PM_{2.5}$ measurements with respect to season (pump_start, pump_stop)

Display 1. Instructions provided to SAS users A-E.

The CONTENTS Procedure

Data Set Name	TMP1.SURVEY	Observations	330
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	Friday, October 07, 2011 01:35:19 PM	Observation Length	112
Last Modified	Friday, October 07, 2011 01:35:19 PM	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	YES
Label			
Data Representation	WINDOWS_32		
Encoding	wlatin1 Western (Windows)		

Variables in Creation Order

#	Variable	Type	Len	Format	Informat	Label
1	PM25	Num	8			PM 2.5 MEASURE
2	MP	Num	8			Week
3	SITE_ID	Num	8	11.	11.	STUDY SITE ID
4	ADDRESS	Char	50	$50.	$50.	address
5	City	Char	14	$14.	$14.	City
6	Zipcode	Char	5	$5.	$5.	Zipcode
7	PUMP_START	Num	8	MMDDYY8.		
8	PUMP_STOP	Num	8	MMDDYY8.		

Sort Information

Sortedby	MP
Validated	YES
Character Set	ANSI

Output1. Selected output from a PROC CONTENTS of the dataset provided to participants.

THE RESULTS

Each SAS users responses are outlined below. The syntax and output included were selected based on their interest and relevance.

SAS USER A

User A immediately accessed the data in JMP. He/she returned a JMP dataset and indicated that his/her responses could be accessed by running the scripts (Display 2). The scripts produced a series of three scatter plots (Output2) accomplishing all three tasks without providing a single frequency, percentage, or descriptive statistic.

SAS USER B

User B created a series of PROC FREQ and PROC MEANS (Display 3) to get a sense of the data and then created a series of plots using PROC GPLOT and PROC GCHART to depict the trend in measurement by week and month (Output 3).

SAS USER C

User C ran two simple PROC MEANS (Display 4) to execute tasks 1 and 2 and then used PROC BOXPLOT to complete task 3 (Output 4).

SAS USER D

After completing a PROC MEANS to accomplish task 1, SAS User D divided the exposure measurement into quartiles (Display 5) and created his/her output based on those quartiles. Using only PROC FREQ the programmer created tables and an elaborate stacked frequency plot using ODS in order to address tasks 2 and 3 (Output 5).

SAS USER E

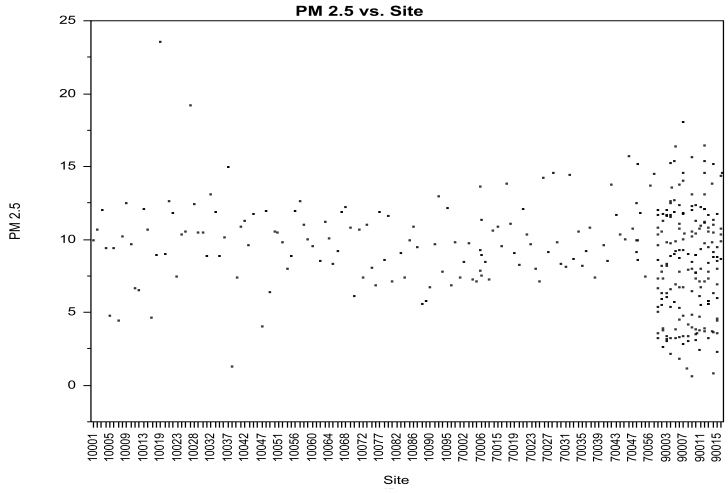
User E used his/her pre-existing knowledge of the data to divide the observations into 3 strata by creating the variable *stud_grp* (Display 6). He/she then used PROC FREQ and PROC TABULATE depicting the distribution of observations as well as $PM_{2.5}$ measurement over time (Output 6).

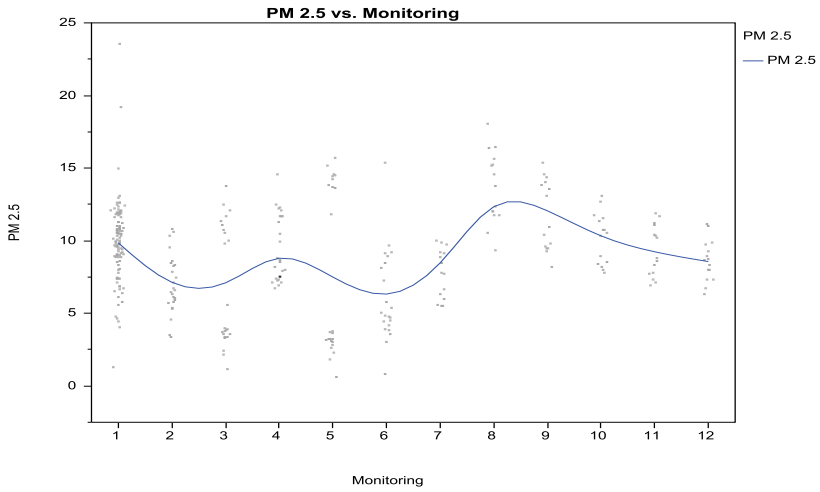
```
Graph Builder(
	Variables( X( :Site ID ), Y( :PM 2.5 ) ),
	Elements( Points( X, Y, Legend( 2 ), Jitter( 1 ) ) ),
	SendToReport(
		Dispatch( {}, "Site ID", ScaleBox, {Rotated Labels( "Automatic" )} )
Graph Builder(
	Lock Scales( 1 ),
	Variables( X( :Monitoring Period ), Y( :PM 2.5 ) ),
	Elements(
		Points( X, Y, Legend( 7 ), Jitter( 1 ) ),
		Smoother( X, Y, Legend( 8 ) )
	SendToReport(
		Dispatch(
			{},
			"Monitoring Period",
			ScaleBox,
			{Min( -0.5 ), Max( 11.5 ), Inc( 1 ), Minor Ticks( 0 ),
			Rotated Labels( "Automatic" )}
		Dispatch(
			{},
			"PM 2.5",
			ScaleBox,
			{Min( -2.5 ), Max( 25 ), Inc( 5 ), Minor Ticks( 1 ),
			Rotated Labels( "Automatic" )}
		Dispatch(
			{},
			"",
			ScaleBox,
			{Min( 0 ), Max( 0 ), Inc( 1 ), Minor Ticks( 0 ),
			Rotated Labels( "Automatic" )}
		Dispatch(
			{},
			"",
			ScaleBox( 2 ),
			{Min( 0 ), Max( 0 ), Inc( 1 ), Minor Ticks( 0 ),
			Rotated Labels( "Automatic" )}
Graph Builder(
	Lock Scales( 1 ),
	Variables( X( :Season ), Y( :PM 2.5 ) ),
	Elements(
		Points( X, Y, Legend( 7 ), Jitter( 1 ) ),
		Smoother( X, Y, Legend( 10 ) )
	SendToReport(
		Dispatch(
			{},
			"Season",
			ScaleBox,
			{Min( -0.5 ), Max( 3.5 ), Inc( 1 ), Minor Ticks( 0 ),
			Show Major Ticks( 0 ), Show Minor Ticks( 0 ),
			Rotated Labels( "Automatic" )}
		Dispatch(
			{},
			"PM 2.5",
			ScaleBox,
			{Min( -1.53378378378378 ), Max( 25.9662162162162 ), Inc( 5 ),
			Minor Ticks( 1 ), Rotated Labels( "Automatic" )}
		Dispatch(
			{},
			"",
			ScaleBox,
			{Min( 0 ), Max( 0 ), Inc( 1 ), Minor Ticks( 0 ),
			Rotated Labels( "Automatic" )}
		Dispatch(
			{},
			"",
			ScaleBox( 2 ),
			{Min( 0 ), Max( 0 ), Inc( 1 ), Minor Ticks( 0 ),
			Rotated Labels( "Automatic" )}
```

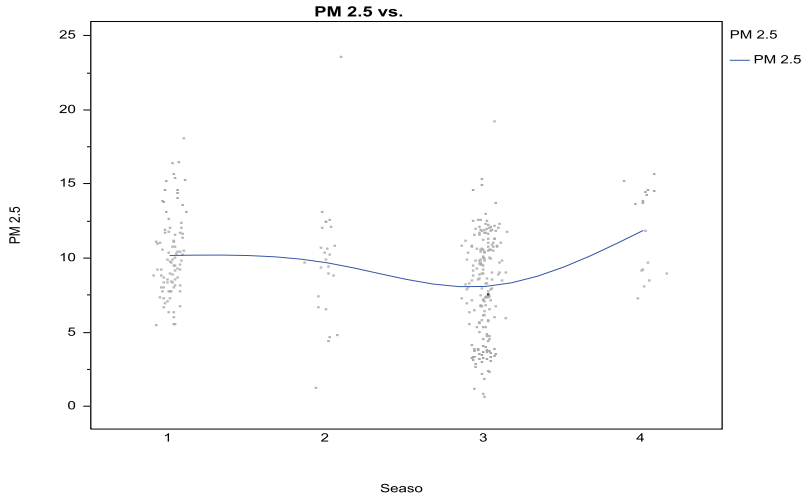
Display 2. Selected syntax from SAS User A.

Mix & Match: Diversity in displaying data, continued

Graph Builder







Output 2. Selected output from SAS User A.

Mix & Match: Diversity in displaying data, continued

.

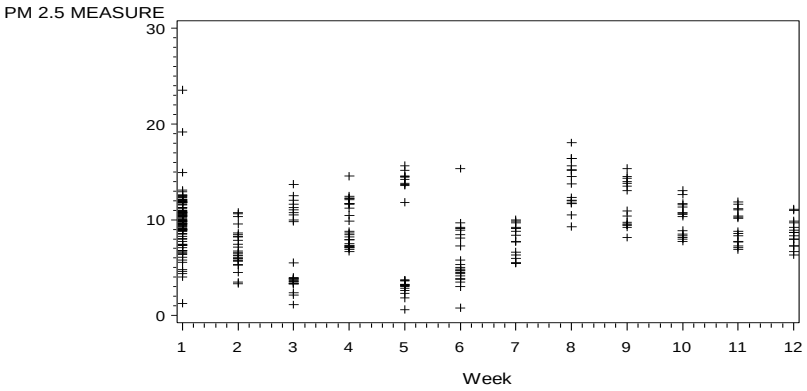
```
Proc sort data=report;
by site_id;
run;
ods rtf style=journal file= 'C:\Documents and Settings\gb297\Desktop\Geetanjoli\Report.rtf' ;
proc freq data=report;
tables site_id;
run;
proc means data=report;
var pm25;
class site_id;
run;
ods rtf close;
*PM2.5 measurements for each monitoring period;
ods rtf style=journal file= 'C:\Documents and Settings\gb297\Desktop\Geetanjoli\graph.rtf';
proc gplot data=report;
plot pm25*mp;
run;
data report2;
set report;
where mp=1;
run;
 proc gchart data=report2;
vbar pm25;
by mp;
run;
*seasonal trends;
data season;
set report;
month1 = month(pump_start);
month2 = month(pump_stop);
drop pump_start pump_stop;
run;
data season2;
set season;
if month1 = 1 and month2 = 1 then month = 1;
if month1 = 2 and month2 = 2 then month = 2;
if month1 = 3 and month2 = 3 then month = 3;
if month1 = 4 and month2 = 4 then month = 4;
if month1 = 5 and month2 = 5 then month = 5;
if month1 = 6 and month2 = 6 then month = 6;
if month1 = 7 and month2 = 7 then month = 7;
if month1 = 8 and month2 = 8 then month = 8;
if month1 = 9 and month2 = 9 then month = 9;
if month1 = 10 and month2 = 10 then month = 10;
if month1 = 11 and month2 = 11 then month = 11;
if month1 = 12 and month2 = 12 then month = 12;
run;
proc gplot data=season2;
plot pm25*month;
run;
ods rtf close;
```

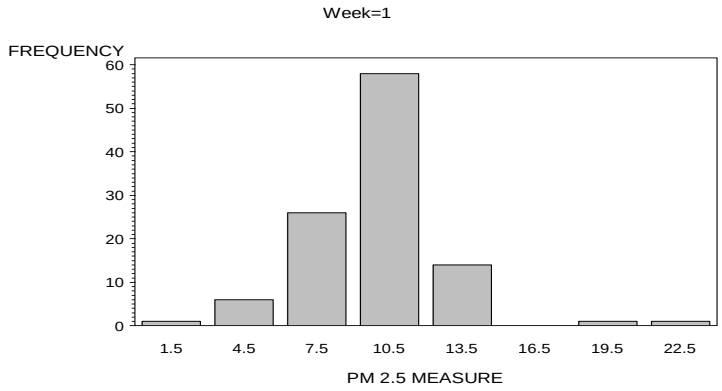
Display 3. Selected syntax from SAS User B.

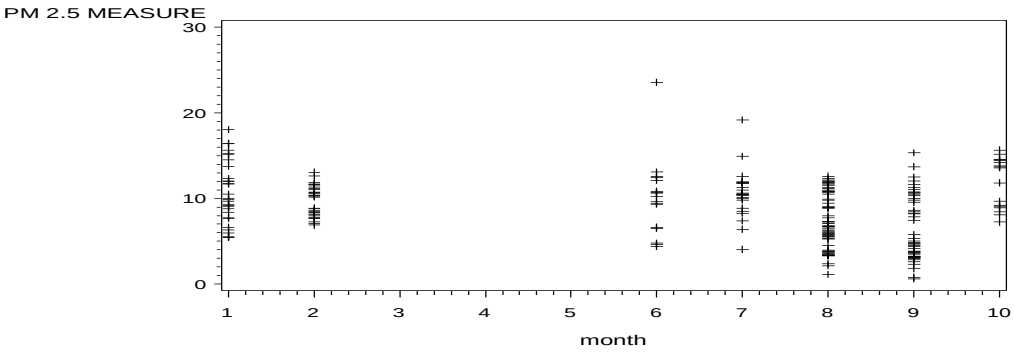
Mix & Match: Diversity in displaying data, continued

Analysis Variable : PM25 PM 2.5 MEASURE

STUDY SITE ID	N Obs	N	Mean	Std Dev	Minimum	Maximum
10001	1	1	9.8997337	.	9.8997337	9.8997337
10002	1	1	10.6591771	.	10.6591771	10.6591771
10003	1	1	12.0069729	.	12.0069729	12.0069729







Output 3. Selected output from SAS User B.

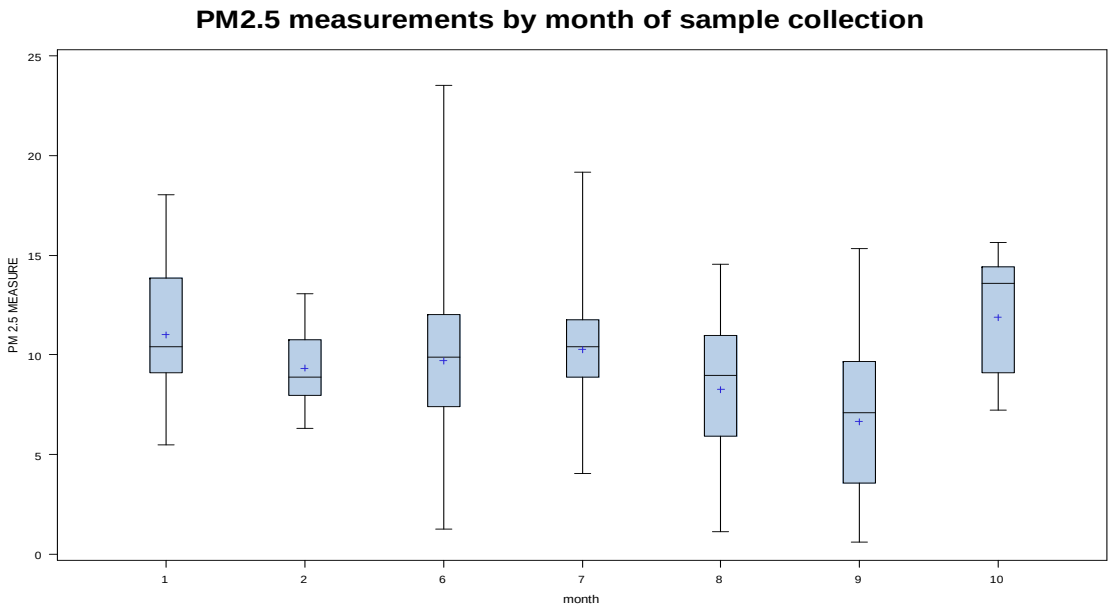
Mix & Match: Diversity in displaying data, continued

```
proc sort data = jkmh.survey; by site_id; run;
proc means data = jkmh.survey n nmiss min median max mean std;
var pm25 ;
by site_id;
run;
proc sort data = jkmh.survey; by mp; run;
proc means data = jkmh.survey n nmiss min median max mean std;
var pm25 ;
by mp;
run;data surveyb;set jkmh.survey;month = month (pump_start); run;
proc sort data = surveyb; by month; run;
proc boxplot data = surveyb;
plot pm25*month;
title 'PM2.5 measurements by month of sample collection';
run;
```

Display 4. Complete syntax from SAS User C.

```
------------------------------------ STUDY SITE ID=10001 ------------------------------------
                                   The MEANS Procedure

                          Analysis Variable : PM25 PM 2.5 MEASURE
             N
   N      Miss         Minimum          Median         Maximum            Mean         Std Dev
-------------------------------------------------------------------------------------------
   1         0       9.8997337       9.8997337       9.8997337       9.8997337               .
-------------------------------------------------------------------------------------------
.
------------------------------------------- Week=1 -------------------------------------------
                                   The MEANS Procedure
                          Analysis Variable : PM25 PM 2.5 MEASURE
             N
   N      Miss         Minimum          Median         Maximum            Mean         Std Dev
-------------------------------------------------------------------------------------------
 107         4       1.2439541       9.9706710      23.5358916       9.8858919       2.7741835
-------------------------------------------------------------------------------------------
```



Output 4. Selected output from SAS User C.

Mix & Match: Diversity in displaying data, continued

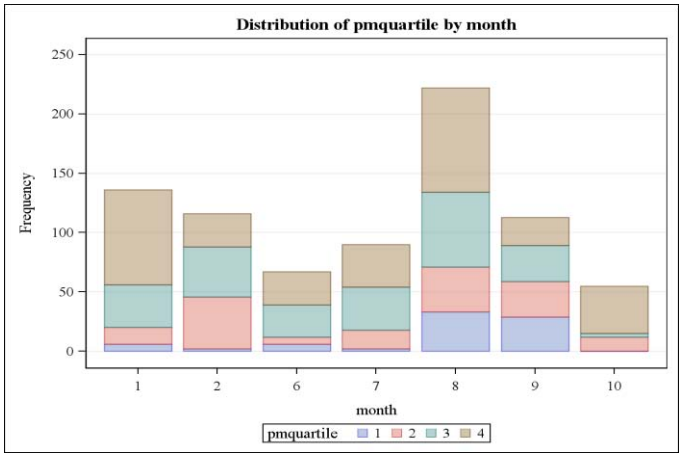
```
data b3; set b1;
pmquartile =.;
if 0.01 <=pm25 < 6.8 then pmquartile = 1;
if 6.8 <=pm25 < 9.2 then pmquartile = 2;
if 9.2 <=pm25 < 11.2 then pmquartile = 3;
if pm25 > = 11.2 then pmquartile = 4;
run;
proc freq data = b3; tables mp*pmquartile; run;
ods graphics on;
proc freq data=b3;
  tables  pmquartile*month/ trend measures cl
            plots=freqplot(twoway=stacked);
  test smdrc;
  exact trend / maxtime=60;
  weight pmquartile;
  title 'LOL';
run;
ods graphics off;ods rtf close;
```

Display 5. Selected syntax from SAS User D.

Mix & Match: Diversity in displaying data, continued

Table of MP by pmquartile					
MP(Week)	pmquartile				
Frequency Percent Row Pct Col Pct	1	2	3	4	Total
1	12 3.77 11.21 15.38	25 7.86 23.36 31.25	41 12.89 38.32 51.90	29 9.12 27.10 35.80	107 33.65
2	14 4.40 56.00 17.95	7 2.20 28.00 8.75	4 1.26 16.00 5.06	0 0.00 0.00 0.00	25 7.86
3	14 4.40 58.33 17.95	0 0.00 0.00 0.00	5 1.57 20.83 6.33	5 1.57 20.83 6.17	24 7.55
4	1 0.31 4.00 1.28	13 4.09 52.00 16.25	2 0.63 8.00 2.53	9 2.83 36.00 11.11	25 7.86
5	15 4.72 60.00 19.23	0 0.00 0.00 0.00	0 0.00 0.00 0.00	10 3.14 40.00 12.35	25 7.86
6	14 4.40 63.64 17.95	6 1.89 27.27 7.50	1 0.31 4.55 1.27	1 0.31 4.55 1.23	22 6.92
7	6 1.89 40.00 7.69	6 1.89 40.00 7.50	3 0.94 20.00 3.80	0 0.00 0.00 0.00	15 4.72
8	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 0.63 13.33 2.53	13 4.09 86.67 16.05	15 4.72
9	0 0.00 0.00 0.00	1 0.31 6.67 1.25	7 2.20 46.67 8.86	7 2.20 46.67 8.64	15 4.72
10	0 0.00 0.00 0.00	6 1.89 40.00 7.50	4 1.26 26.67 5.06	5 1.57 33.33 6.17	15 4.72
11	0 0.00 0.00 0.00	8 2.52 53.33 10.00	5 1.57 33.33 6.33	2 0.63 13.33 2.47	15 4.72
12	2 0.63 13.33 2.56	8 2.52 53.33 10.00	5 1.57 33.33 6.33	0 0.00 0.00 0.00	15 4.72
Total	78 24.53	80 25.16	79 24.84	81 25.47	318 100.00
Frequency Missing = 12					

Output 5A. Task 2 output from SAS User D.



Output 5B. Task 3 output from User D.

Mix & Match: Diversity in displaying data, continued

```
data x.survey2;set x.survey;
stud  grp=.;if(10000<site  id<70000) then stud_grp=1;
if(70000<=site_id<90000) then stud_grp=2;
if(site_id>90000) then stud_grp=3;
stud  year=year(pump  start);stud  mo=month(pump  start);
proc freq data=x.survey2;tables stud  year*stud_mo;run;
proc sort data=x.survey2;by site_id;run;
proc print data=x.survey2;var site_id pm25;run;
proc freq data=x.survey2;tables stud_grp mp;run;
proc tabulate data=x.survey2;
        class stud_grp mp;
        var pm25;
        table stud_grp mp, stud_grp*pm25*(mean*f=6.2 std*f=6.2 n*f=3.0);run;
proc tabulate data=x.survey2;
        class stud_year stud_mo;
        var pm25;
        table stud_mo,stud_year*pm25*(mean*f=6.2 std*f=6.2 n*f=3.0);run;
```

Display 6. Selected syntax from SAS User E.

The FREQ Procedure

Week

MP	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	111	33.64	111	33.64
2	25	7.58	136	41.21
3	25	7.58	161	48.79
4	25	7.58	186	56.36
5	25	7.58	211	63.94
6	23	6.97	234	70.91
7	16	4.85	250	75.76
8	16	4.85	266	80.61
9	16	4.85	282	85.45
10	16	4.85	298	90.30
11	16	4.85	314	95.15
12	16	4.85	330	100.00

	stud_year								
	2007			2010			2011		
	PM 2.5 MEASURE			PM 2.5 MEASURE			PM 2.5 MEASURE		
	Mean	Std	N	Mean	Std	N	Mean	Std	N
stud_mo									
1	.	.	.	.	.	.	11.02	3.32	45
2	.	.	.	.	.	.	9.33	1.74	45
6	.	.	.	9.70	4.17	25	.	.	.
7	.	.	.	10.27	2.61	31	.	.	.
8	9.48	0.92	10	8.13	3.28	85	.	.	.
9	9.42	1.79	30	3.89	2.46	30	.	.	.
10	11.88	2.93	17	.	.	.	.	.	.

Output 6. Selected output from SAS User E.

CONCLUSION

It is a rare occasion when a SAS programmer is limited to displaying data in only one way; various procedures can be used to produce similar tables and plots depicting the same information. In completing the three assigned tasks within a new dataset almost all users (User A excluded) performed some kind of exploratory procedure such as PROC CONTENTS, PROC MEANS, or PROC FREQ to garner a better understanding of the dataset before attempting to complete any of the specified tasks. Although not depicted in the syntax and output highlighted in this paper, it is interesting to note that the programmers felt a need to familiarize themselves with the dataset before determining their approach to the three tasks.

Similarly noteworthy is the fact that most programmers (again excluding User A) chose PROC MEANS for the first task. PROC MEANS allows the user to output a full range of descriptive statistics about a variable's distribution. The programmer can execute this procedure for more than one variable by including all variables of interest in the *var* statement and further explore the distributions within subgroups using *class* and *by* statements. This method is clearly recognizable in the syntax featured in Displays 3 and 4.

The second task was essentially a duplicate of task 1, but required that the programmer identify a different type of subgroup (monitoring period). As exhibited by User C, the exact syntax from task 1could have been recycled substituting the variable used in the *class* or *by* statement. However, the other three programmers working in SAS chose to use a new approach.

User B opted to use PROC GPLOT and PROC GCHART to create visual depictions of the data by monitoring period; a select few are featured in Output3. PROC GCHART is used with a VBAR statement to produce quality bar charts of frequency counts. Proc GPLOT displays the relationship between variables with a scatter plot.

Users D and E moved to PROC FREQ for task 2. PROC FREQ is a versatile procedure which can be used to create crosstab or list style output featuring the frequency of observations in each category of a variable or each cell of a cross-tabular display. PROC FREQ provides the user with a myriad of options which determine important factors such as if/how/where missing values will be included in the output as well as which types of percentages will be displayed. User D divided the outcome measure into quartiles (Display 5) and used PROC FREQ to output a crosstab of monitoring period by $PM_{2.5}$ quartile (Output 5A). User E also employed PROC FREQ (Display 6) but chose to create frequency listings for the variables *stud_grp* and *mp* (Output 6).

The third task called for a more dynamic exploration of the data in order to capture any trends in $PM_{2.5}$ measurement with respect to season. In completing task 3 we see the most variety in user approach. Users B and D continued working within the same procedures they had already used, this time broadening the scope of their utility by either creating new variables (User B) or by employing more options (User D). User B employed PROC GPLOT in the same fashion as he/she approached task 2, this time plotting $PM_{2.5}$ measurements over calendar month. User D continued to work in PROC FREQ, this time using the *plots* option in the *tables* statement to create a stacked frequency distribution of $PM_{2.5}$ quartile by calendar month. Using ODS to export the output into an rtf file allows the programmer to create colorful and easily interpreted displays which can effortlessly be moved and modified within the Windows environment.

Users C and E incorporated two new procedures in order to conquer task 3. User C worked with PROC BOXPLOT to fashion a monochromatic plot from which mean, median, confidence interval and standard deviation can all be compared across calendar months. PROC BOXPLOT is a procedure specifically designed to create box-and-whisker plots for comparison across subgroups. More complex statements such as the *Inset* or *Insetgroup* statements can be used to label and print additional values on the plot itself. User E moved into PROC TABULATE which creates journal style tables displaying a broad range of statistical measures for comparison among groups. In this case, the programmer built a three dimensional comparison of mean, standard deviation, and frequency of $PM_{2.5}$ observations across calendar year and month.

This exercise indicates that while programming and data analysis are highly scientific endeavors, one's approach to both of these tasks is highly personal. When a SAS User 'runs some frequencies' or 'puts together a quick report' he/she has a variety of tools at his/her disposal. How programmers use these tools is strongly influenced by their experience with SAS as well as their role in the management of data for research projects.

ACKNOWLEDGMENTS

We'd like to thank the following individuals at the Yale CPPEE for participating in our SAS experiment: Geetanjoli Banerjee, MPH, John Havens-McColgan MPH, Janneane Gent, PhD

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Melissa Hill, MPH
Enterprise: Yale University Center for Perinatal, Pediatric and Environmental Epidemiology
Address: One Church Street, 6th Floor
City, State ZIP: New Haven, CT 06510
Work Phone: 203-764-9375
Fax: 203-764-9378
E-mail: Melissa.hill@yale.edu

Name: Julie Kezik, MPH
Enterprise: Yale University Center for Perinatal, Pediatric and Environmental Epidemiology
Address: One Church Street, 6th Floor
City, State ZIP: New Haven, CT 06510
Work Phone: 203-764-9375
Fax: 203-764-9378
E-mail: julie.kezik@yale.edu