SESUG Paper BB-51-2017

Data Analysis and Storytelling – Communicating Analytical Results with Clarity, Precision and Efficiency

Kirk Paul Lafler, Software Intelligence Corporation, Spring Valley, California

Abstract

The data analysis process involves the gathering and collection, cleansing, transforming, and modeling of data from various sources. The purpose is to discover, evaluate, understand and derive useful information from the data to support decision-making. Unfortunately, and all too often, data analysts omit a very crucial step – the development of a narrative, or story, of the data analysis process and outcome. This omission not only fails to bring context, insight and interpretation of the data analysis results to stakeholders, it neglects to bring meaning, relevance and interest to the "key" points of the data analysis results. This presentation describes the importance, considerations and steps needed in developing a compelling narrative, along with the necessary visual analytics, to communicate a convincing point-of-view to help persuade others to understand the complexities associated with the data analysis results.

Introduction

Considerable effort and resources are expended by data analysts in the performance of analyzing data to discover, evaluate, understand and derive meaningful and useful information to assist the decision-making process. Unfortunately, the data analysis process doesn't always include a narrative, or storyline, to convey information about the discovery, evaluation, and understanding of the data analysis results. The omission or failure to bring context, insight and interpretation of the data analysis results to stakeholders, denies an audience meaning, relevance and interest to the "key" points of the data analysis results. This paper describes the importance, considerations and steps involved in performing a comprehensive data analysis along with the development of a compelling narrative to persuade others to understand the complexities of their data analysis results. Examples are illustrated using the SASHELP.HEART data set which consists of 5,209 observations and seventeen columns, illustrated below.

Obs	Status	DeathCause	AgeCHDdiag	Sex	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	AgeAtDeath	Cholesterol	Chol_Status	BP_Status	Weight_Status	Smoking_Status
1	Dead	Other	.	Female	29	62.50	140	78	124	121	0	55	.		Normal	Overweight	Non-smoker
2	Dead	Cancer	.	Female	41	59.75	194	92	144	183	0	57	181	Desirable	High	Overweight	Non-smoker
3	Alive		.	Female	57	62.25	132	90	170	114	10	.	250	High	High	Overweight	Moderate (6-15)
4	Alive		.	Female	39	65.75	158	80	128	123	0	.	242	High	Normal	Overweight	Non-smoker
5	Alive		.	Male	42	66.00	156	76	110	116	20	.	281	High	Optimal	Overweight	Heavy (16-25)
6	Alive		.	Female	58	61.75	131	92	176	117	0	.	196	Desirable	High	Overweight	Non-smoker
7	Alive		.	Female	36	64.75	136	80	112	110	15	.	196	Desirable	Normal	Overweight	Moderate (6-15)
8	Dead	Other	.	Male	53	65.50	130	80	114	99	0	77	276	High	Normal	Normal	Non-smoker
9	Alive		.	Male	35	71.00	194	68	132	124	0	.	211	Borderline	Normal	Overweight	Non-smoker
10	Dead	Cerebral Vascular Disease	.	Male	52	62.50	129	78	124	106	5	82	284	High	Normal	Normal	Light (1-5)
11	Alive		.	Male	39	66.25	179	76	128	133	30	.	225	Borderline	Normal	Overweight	Very Heavy (> 25)
12	Alive		57	Male	33	64.25	151	68	108	118	0	.	221	Borderline	Optimal	Overweight	Non-smoker
13	Alive		55	Male	33	70.00	174	90	142	114	0	.	188	Desirable	High	Overweight	Non-smoker
14	Alive		79	Male	57	67.25	165	76	128	118	15	.	.		Normal	Overweight	Moderate (6-15)
15	Alive		66	Male	44	69.00	155	90	130	105	30	.	292	High	High	Normal	Very Heavy (> 25)
16	Alive		.	Female	37	64.50	134	76	120	108	10	.	196	Desirable	Normal	Normal	Moderate (6-15)
17	Alive		.	Male	40	66.25	151	72	132	112	30	.	192	Desirable	Normal	Overweight	Very Heavy (> 25)
18	Dead	Cancer	56	Male	56	67.25	122	72	120	87	15	72	194	Desirable	Normal	Underweight	Moderate (6-15)
19	Alive		.	Female	42	67.75	162	96	138	119	1	.	200	Borderline	High	Overweight	Light (1-5)
20	Dead	Coronary Heart Disease	74	Male	46	66.50	157	84	142	116	30	76	233	Borderline	High	Overweight	Very Heavy (> 25)
21	Alive		.	Female	37	66.25	148	78	110	112	15	.	192	Desirable	Optimal	Overweight	Moderate (6-15)
22	Alive		.	Female	45	64.00	147	74	120	119	5	.	209	Borderline	Normal	Overweight	Light (1-5)
23	Alive		.	Female	59	65.75	156	74	156	122	0	.	200	Borderline	High	Overweight	Non-smoker
24	Alive		.	Female	36	63.75	122	84	132	102	0	.	184	Desirable	Normal	Normal	Non-smoker
25	Alive		.	Female	50	67.50	185	88	150	136	15	.	228	Borderline	High	Overweight	Moderate (6-15)
26	Alive		.	Female	35	66.00	123	76	132	93	0	.	150	Desirable	Normal	Normal	Non-smoker
27	Alive		.	Male	42	72.25	182	78	136	113	0	.	221	Borderline	Normal	Overweight	Non-smoker
28	Dead	Coronary Heart Disease	71	Female	49	60.50	153	110	196	140	5	73	221	Borderline	High	Overweight	Light (1-5)
29	Alive		68	Male	40	70.00	189	78	124	124	0	.	319	High	Normal	Overweight	Non-smoker
30	Alive		.	Female	41	61.75	139	72	116	124	0	.	194	Desirable	Optimal	Overweight	Non-smoker
31	Dead	Unknown	.	Female	59	67.75	153	82	172	113	0	79	263	High	High	Overweight	Non-smoker
32	Alive		68	Male	40	70.00	195	76	132	128	20	.	205	Borderline	Normal	Overweight	Heavy (16-25)
33	Alive		.	Female	41	62.00	114	78	112	98	15	.	267	High	Optimal	Normal	Moderate (6-15)
34	Alive		.	Female	39	63.00	144	80	120	120	0	.	196	Desirable	Normal	Overweight	Non-smoker
35	Alive		43	Male	33	66.50	172	106	146	127	0	.	247	High	High	Overweight	Non-smoker
36	Alive		.	Male	41	69.25	159	96	142	107	0	.	209	Borderline	High	Normal	Non-smoker
37	Dead	Coronary Heart Disease	67	Female	49	61.00	142	92	138	127	30	75	276	High	High	Overweight	Very Heavy (> 25)
38	Alive		.	Male	51	69.50	181	98	144	122	20	.	223	Borderline	High	Overweight	Heavy (16-25)
39	Dead	Cancer	.	Male	43	65.50	172	78	118	131	10	63	150	Desirable	Optimal	Overweight	Moderate (6-15)
40	Alive		.	Male	48	66.75	142	72	108	105	30	.	292	High	Optimal	Normal	Very Heavy (> 25)

Data Analysis and Data Storytelling

Data analysis involves inspecting, cleansing, transforming and discovering useful information from a variety of data sources. Considerable effort should be spent on the quality of data and the data cleaning effort. From determining the frequency counts, the minimum, maximum, mean, standard deviation, and variance can all lend credence to the assessment of data quality. Once this is accomplished, the process of storytelling attempts to express and communicate complex ideas, data and statistics with clarity, precision and efficiency. Storytelling uses visuals, graphs and charts to help an audience gain greater insight and which supports the underlying data. The following table illustrates eight tips for better data analysis and storytelling.

Tips for Better Data Analysis and Storytelling
Tip #1: Storyboard a detailed outline of your thoughts and ideas.
Tip #2: Communicate creatively, clearly and understandably.
Tip #3: Create a "water cooler moment" using headlines, tweets and images.
Tip #4: Stimulate your audience's senses with a compelling story.
Tip #5: Talk to and engage your audience with a relatable human interest story.
Tip #6: Make your story memorable and impactful.
Tip #7: Determine the best type of visuals to use.
Tip #8: Create an impactful message using graphs, charts and other visuals.

The Base SAS® software provides users with a number of powerful procedures to help with data analysis activities. The following table illustrates the names of popular Base SAS procedures along with their purpose.

Procedure	Purpose
PROC FREQ	**Produces one-way to n-way frequency and cross-tabular results in a tabular layout.**
PROC MEANS	**Produces data summaries by computing descriptive statistics across all observations and within groups of observations.**
PROC PRINT	**Produces data in a simple, detail and organized layout.**
PROC SQL	**Produces detail, summary and statistical results in an organized layout.**
PROC SUMMARY	**Like PROC MEANS, this procedure computes descriptive statistics across all observations and within groups of observations.**
PROC TABULATE	**Produces descriptive statistics in a tabular layout.**
PROC UNIVARIATE	**Produces descriptive statistics including moments, quantiles (or percentiles), frequency tables, and extreme values.**

Data Analysis Programming Techniques

Data analysis involves the process of inspecting, cleaning, transforming and discovering information from a variety of structured and unstructured data sources. Data is collected, processed and analyzed to answer questions and make decisions. A few data analysis techniques will be explored to illustrate popular programming techniques. In the first data analysis example, a PROC FREQ with the NLEVELS option is specified to determine the number of distinct groups (or levels) that exist for the SEX and STATUS variables.

```
proc freq data=sashelp.heart
         (keep=sex ageatstart height weight status)
          nlevels ;
  tables sex status ;
run ;
```

The FREQ Procedure

Number of Variable Levels	
Variable	Levels
Sex	2
Status	2

Sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	2873	55.15	2873	55.15
Male	2336	44.85	5209	100.00

Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Alive	3218	61.78	3218	61.78
Dead	1991	38.22	5209	100.00

In the next data analysis example, a PROC FREQ is specified to illustrate a two-way interaction table (or cross-tabulation) between the SEX and STATUS variables.

```
proc freq data=sashelp.heart(keep=sex status) ;
  tables sex * status ;
run ;
```

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of Sex by Status			
	Status		
Sex	Alive	Dead	Total
Female	1977 37.95 68.81 61.44	896 17.20 31.19 45.00	2873 55.15
Male	1241 23.82 53.13 38.56	1095 21.02 46.88 55.00	2336 44.85
Total	3218 61.78	1991 38.22	5209 100.00

In the next data analysis example, a PROC MEANS is specified to illustrate the descriptive statistics N, MIN, MAX, MEAN, Standard Deviation, and Variance for the SEX and STATUS variables.

```
proc means data=sashelp.heart
  (keep=sex height weight status)
        n min max mean std var ;
  class sex status ;
run ;
```

The MEANS Procedure

Sex	Status	N Obs	Variable	N	Minimum	Maximum	Mean	Std Dev	Variance
Female	Alive	1977	Height	1976	54.7500000	70.7500000	62.7057186	2.4342883	5.9257597
			Weight	1974	85.0000000	300.0000000	138.9701114	24.3559226	593.2109663
	Dead	896	Height	893	51.5000000	69.7500000	62.2779955	2.4679958	6.0910035
			Weight	895	67.0000000	300.0000000	146.7229050	29.4343633	866.3817455
Male	Alive	1241	Height	1241	56.0000000	76.5000000	67.8890008	2.7498566	7.5617116
			Weight	1241	111.0000000	260.0000000	167.2312651	24.3621968	593.5166346
	Dead	1095	Height	1093	56.5000000	76.0000000	67.2021958	2.6664703	7.1100639
			Weight	1093	99.0000000	276.0000000	167.7328454	26.3139330	692.4230685

In the final data analysis example, a PROC UNIVARIATE with a CLASS statement, for grouping the results by the SEX variable, is specified to illustrate a slew of descriptive statistics including Moments, Basic Measures, Test for Location, Quantiles, Extreme Observations, and Missing Values for the HEIGHT and WEIGHT variables.

```
proc univariate data=sashelp.heart ;
 class sex ;
 var height weight ;
run ;
```

The UNIVARIATE Procedure
Variable: Height
Sex = Female

Moments			
N	2869	Sum Weights	2869
Mean	62.5725863	Sum Observations	179520.75
Std Deviation	2.45241118	Variance	6.01432058
Skewness	-0.0145049	Kurtosis	0.1164692
Uncorrected SS	11250326.7	Corrected SS	17249.0714
Coeff Variation	3.91930608	Std Error Mean	0.04578551

Basic Statistical Measures			
Location		Variability	
Mean	62.57259	Std Deviation	2.45241
Median	62.50000	Variance	6.01432
Mode	62.50000	Range	19.25000
		Interquartile Range	3.25000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	1366.645	Pr > |t|	<.0001
Sign	M	1434.5	Pr >= |M|	<.0001
Signed Rank	S	2058508	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	70.75
99%	68.25
95%	66.75
90%	65.75
75% Q3	64.25
50% Median	62.50
25% Q1	61.00
10%	59.50
5%	58.50
1%	56.75
0% Min	51.50

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
51.50	2250	69.50	3862
53.75	3509	69.75	3206
54.75	3024	70.00	458
54.75	2772	70.50	3463
55.00	4632	70.75	1736

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	4	0.14	100.00

The UNIVARIATE Procedure
Variable: Height
Sex = Male

Moments			
N	2334	Sum Weights	2334
Mean	67.5673736	Sum Observations	157702.25
Std Deviation	2.73213662	Variance	7.46457051
Skewness	-0.0293866	Kurtosis	0.07201722
Uncorrected SS	10672941.7	Corrected SS	17414.843
Coeff Variation	4.04357381	Std Error Mean	0.05655252

Basic Statistical Measures			
Location		Variability	
Mean	67.56737	Std Deviation	2.73214
Median	67.50000	Variance	7.46457
Mode	68.50000	Range	20.50000
		Interquartile Range	4.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	1194.772	Pr > |t|	<.0001
Sign	M	1167	Pr >= |M|	<.0001
Signed Rank	S	1362473	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	76.50
99%	73.75
95%	72.00
90%	71.00
75% Q3	69.50
50% Median	67.50
25% Q1	65.50
10%	64.25
5%	63.25
1%	61.25
0% Min	56.00

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
56.00	2800	75.5	419
56.50	5038	75.5	679
58.75	4749	76.0	1894
58.75	4222	76.0	3356
59.25	4750	76.5	3883

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	2	0.09	100.00

Data Storytelling – Develop a Compelling Narrative

Data storytelling should communicate data insights using data analysis results, visuals, and a strong narrative. The online Oxford Dictionary defines storytelling as,

A narrative consisting of a "spoken or written account of connected events; a story."

Source: https://en.oxforddictionaries.com/definition/narrative

In Bessler's (2012) paper (see References), organizations must communicate graphics, charts and other images effectively. Bessler offers the following storytelling insights and suggestions.

- ✓ Deliver image plus precise numbers
- ✓ Provide ordering – Show them what's important
- ✓ Subset the content where appropriate
- ✓ Provide a reliable usable legend
- ✓ Suppress and avoid graphic frills – Let your data talk

Successful data storytelling should seek an objectiveness and balance in its narrative. The following suggestions should help to develop a balanced narrative.

- ✓ Avoid introducing Bias into your analysis, statistics, and visualizations
- ✓ Label Axis to avoid ambiguity
- ✓ Make graphic dimensions match data dimensions
- ✓ Use standardized units

Finally, data storytelling should avoid censorship. The following suggestions offer guidelines to consider.

- ✓ Describe missing data and how you dealt with missing
- ✓ Describe outliers and out-of-range values
- ✓ Describe intervals and other important elements

Know Your Audience

When conducting data analysis and data storytelling, always keep your audience in mind. At a minimum, ask yourself these questions.

- ✓ Who is my audience?
- ✓ Who are the decision makers in the audience?
- ✓ Who are the novices in the audience?
- ✓ Who are the generalists in the audience?
- ✓ Who are the experts in the audience?
- ✓ Who are the executives in the audience?

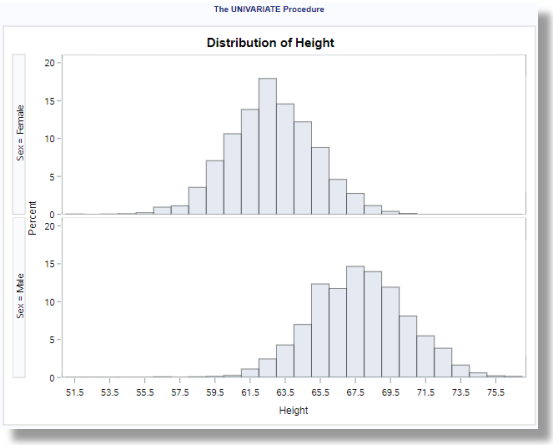
Compelling Visualizations to Help Tell a Story

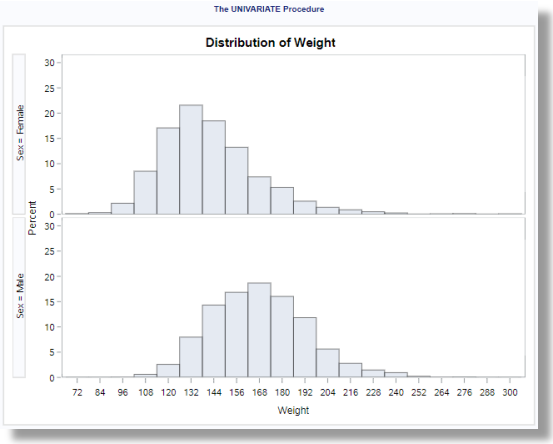
The Base SAS software offers users with powerful procedures to help with the data storytelling narrative. The following table identifies three procedures to help develop compelling visualizations, along with their purpose.

Procedure	Purpose
PROC SGPANEL	**Produces a panel of graph cells representing the values of one or more classification variables.**
PROC SGPLOT	**Produces one or more plots, overlays, histograms, and regression plots using HBAR, HBOX, HISTOGRAM, HLINE, NEEDLE, REG, SCATTER, SERIES, VBAR, VBOX, VECTOR, VLINE and other statements.**
PROC UNIVARIATE	**Produces descriptive statistics including moments, quantiles (or percentiles), frequency tables, extreme values, and histograms.**

A few popular procedures that are used with the data storytelling process are illustrated, below. In the first data visualization example, a PROC UNIVARIATE with CLASS, VAR and HISTOGRAM statements are specified to determine the number of distinct groups (or levels) that exist for the HEIGHT and SEX variables.

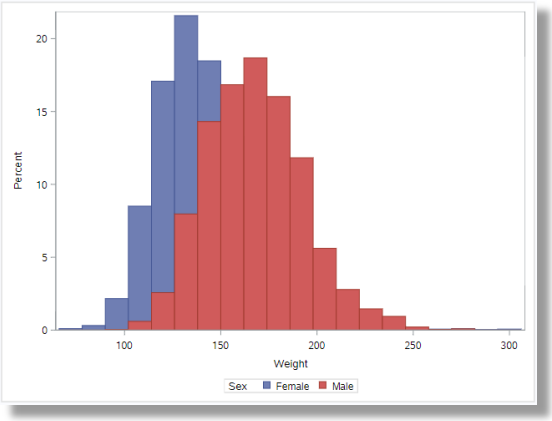
```
proc univariate data=sashelp.heart noprint ;
  class sex ;
  var height weight ;
  histogram height weight / nrows=2 ;
  ods select histogram ;
run ;
```





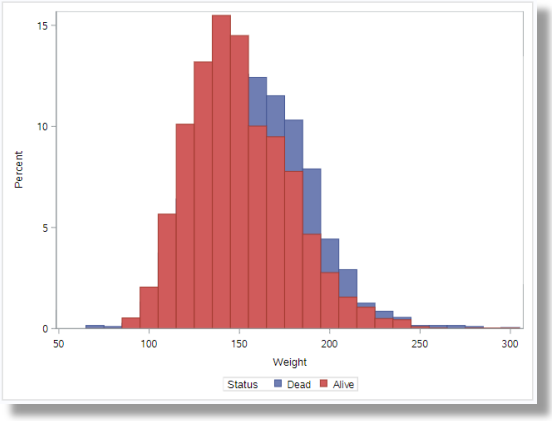
In the next data visualization example, a PROC SGPLOT with a HISTOGRAM statement is specified to display a vertical bar chart for the SEX and WEIGHT variables.

```
proc sgplot data=sashelp.heart ;
  histogram weight / group=sex ;
run ;
```



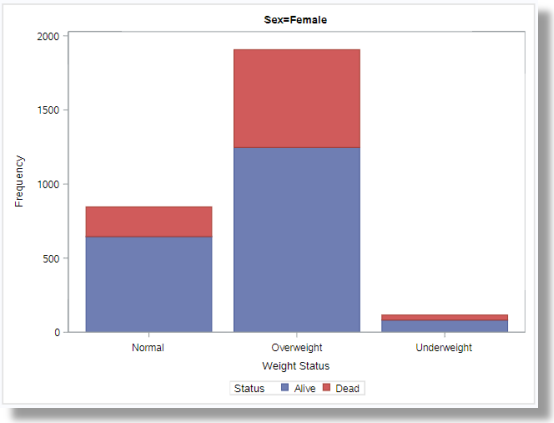
In the next data visualization example, a PROC SGPLOT with a HISTOGRAM statement is specified to display a vertical bar chart for the STATUS and WEIGHT variables.

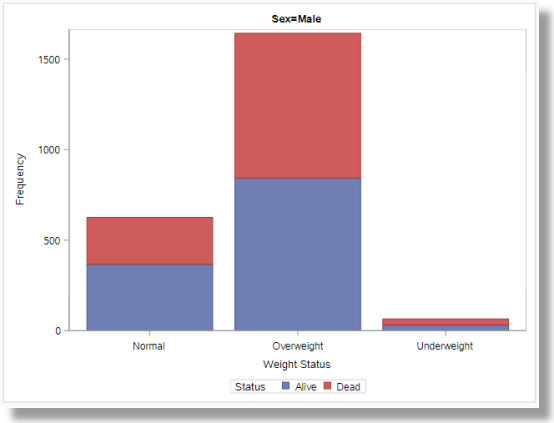
```
proc sgplot data=sashelp.heart ;
  histogram weight / group=status ;
run ;
```



In the next data visualization example, a PROC SORT is specified to sort the SASHELPHEART data set in ascending order by the SEX variable. Then, a PROC SGPLOT with a HISTOGRAM statement is specified to display a vertical bar chart for the STATUS and WEIGHT variables.

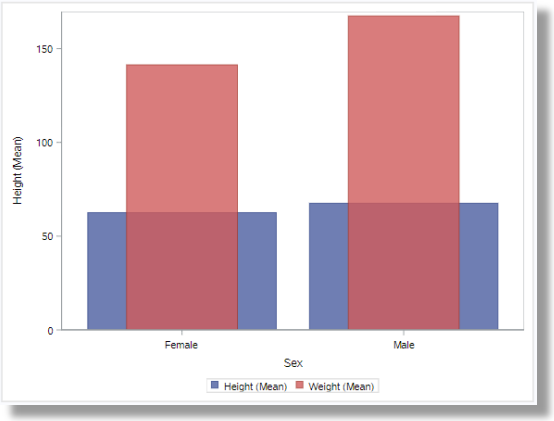
```
proc sort data=sashelp.heart out=heart_sorted ;
  by sex ;
run ;

proc sgplot data=heart_sorted ;
  vbar weight_status / group=status  ;
  by sex ;
run ;
```





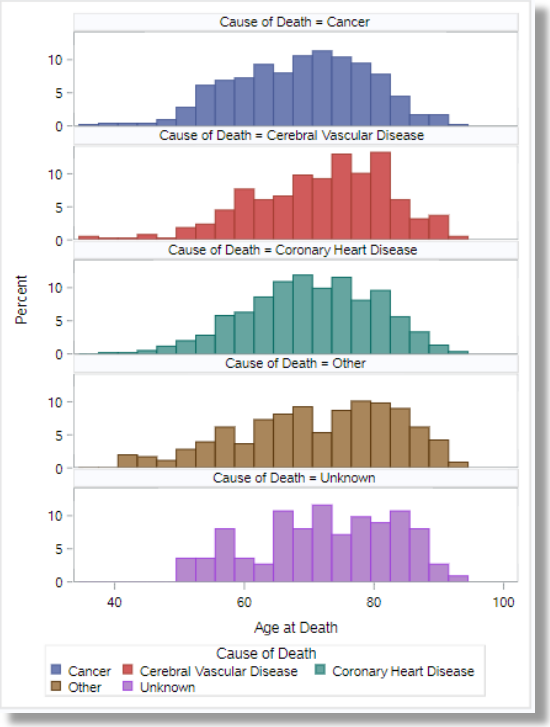
In the next data visualization example, a PROC SGPLOT with two VBAR statements is specified to display an overlay of vertical bar charts for the SEX, HEIGHT, and WEIGHT variables.

```
proc sgplot data=sashelp.heart ;
  vbar sex / response=height stat=mean ;
  vbar sex / response=weight stat=mean
             barwidth=0.5 transparency=0.2 ;
run ;
```



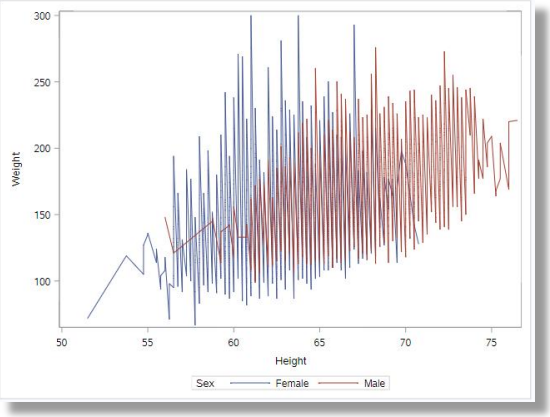
In the next data visualization example, a PROC SGPANEL with the PANELBY and HISTOGRAM statements is specified to display five distinct groups associated with the DEATHCAUSE (Cause of Death) and AGEATDEATH (Age at Death) variables.

```
proc sgpanel data=sashelp.heart ;
  panelby deathcause / rows=5 ;
  histogram ageatdeath / group=deathcause ;
  where deathcause NE '' ;
run ;
```



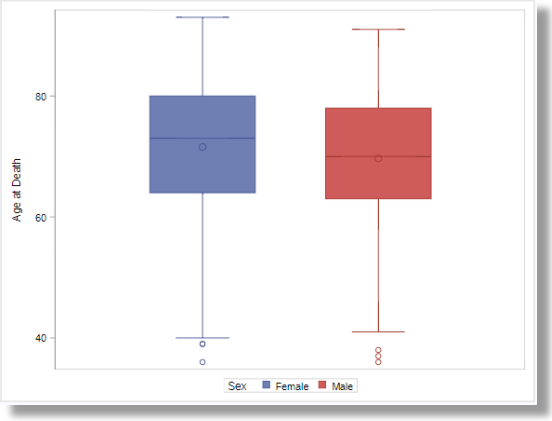
In the next data visualization example, a PROC SORT is first specified to sort the SEX, HEIGHT, and WEIGHT variables in ascending order. Then, a PROC SGPLOT with a SERIES statement is specified to display a plot of the HEIGHT and WEIGHT variables grouped by the SEX variable.

```
proc sort data=sashelp.heart
           out=heart_sorted ;
  by sex height weight ;
run ;

proc sgplot data=heart_sorted ;
  series x=height y=weight / group=sex ;
run ;
```



In the final data visualization example, a PROC SORT is specified to sort the SEX and AGEATDEATH (Age at Death) variables in ascending order. Then, a PROC SGPLOT with a VBOX statement is specified to display a box plot of the AGEATDEATH variable grouped by the SEX variable.

```
proc sort data=sashelp.heart
           out=heart_sorted ;
  by sex ageatdeath ;
run ;

proc sgplot data=heart_sorted ;
  vbox ageatdeath / group=sex ;
run ;
```



Conclusion

Considerable resources are expended by organizations in gathering, cleansing, transforming, and modeling data in the production of the data analysis results. Unfortunately, the data analysis process doesn't always include a narrative, or story, to help convey information about the discovery, evaluation, and understanding of the data analysis results. Not only does this omission fail to bring context, insight and interpretation of the data analysis results to stakeholders, it neglects to bring meaning, relevance and interest to the "key" points of the data analysis results. This paper describes and illustrates the importance, considerations and steps needed to develop a compelling narrative, along with the necessary visual analytics, to communicate a convincing point-of-view to help persuade others to understand the complexities associated with the data analysis results.

References

Bessler, LeRoy PhD (2012). "*Get the Best Out Of SAS® ODS Graphics and the SG (Statistical Graphics) Procedures: Communication-Effective Charts, Things That SAS/GRAPH® Cannot Do As Well, and Macro Tools To Save Time and Avoid Errors*," Proceedings of the 2012 MidWest SAS Users Group (MWSUG) Conference.

Lafler, Kirk Paul (2017). *"Removing Duplicates Using SAS,"* Proceedings of the 2017 SAS Global Forum (SGF) Conference.

Lafler, Kirk Paul (2015). *"*Basic SAS® PROCedures for Quick Results*,"* Proceedings of the 2015 MidWest SAS Users Group (MWSUG) Conference.

Acknowledgments

The author thanks Rachel Straney and John Cohen, Building Blocks Section Chairs, for accepting my abstract and paper; Jason Brinkley, SESUG 2017 Academic Chair; Meenal Sinha, SESUG 2017 Operations Chair; the SouthEast SAS Users Group (SESUG) Executive Board; and SAS Institute for organizing and supporting a great conference!

Trademark Citations

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

About the Author

Kirk Paul Lafler is an entrepreneur, consultant and founder of Software Intelligence Corporation, and has been using SAS since 1979. Kirk is a SAS application developer, programmer, Certified Professional, provider of IT consulting services, advisor and professor at UC San Diego Extension, educator to SAS users around the world, mentor, and emeritus sasCommunity.org Advisory Board member. As the author of six books including Google® Search Complete (Odyssey Press. 2014) and PROC SQL: Beyond the Basics Using SAS, Second Edition (SAS Press. 2013); Kirk has written hundreds of papers and articles; been an Invited speaker and trainer at hundreds of SAS International, regional, special-interest, local, and in-house user group conferences and meetings; and is the recipient of 25 "Best" contributed paper, hands-on workshop (HOW), and poster awards.

Comments and suggestions can be sent to:

Kirk Paul Lafler
SAS® Consultant, Application Developer, Programmer, Data Analyst, Educator and Author
Software Intelligence Corporation
E-mail: KirkLafler@cs.com
LinkedIn: http://www.linkedin.com/in/KirkPaulLafler
Twitter: @sasNerd