DM STAT-1 CONSULTING **BRUCE RATNER, PhD**
574 Flanders Drive North Woodmere, NY 11581 br@dmstat1.com
516.791.3544 ~ fax 516.791.5075 www.dmstat1.com

Opening the Dataset: A Twelve-Step Program for Dataholics

Bruce Ratner, PhD

My name is Bruce Ratner, and I am a dataholic. I am also an artist [1] and poet [2] in the world of statistical data. I always await getting my hands on a new dataset to crack open and paint the untapped, unsoiled numbers into swirling equations, and pencil data gatherings into beautiful verse.

I see numbers as the prime coat for a bedazzled visual percept. Is not the nums-pic in Figure 1, below, grand? [3] I also see mathematical devices as the elements in poetic expressions, which allow truths to lay bare. A poet's rendition of love in Figure 2, below, gives a thinkable pause. [4] For certain, the irresistible equations are poetry of numerical letters. The most powerful and famous is E = m*(c^2). [5] The fairest of them all is e^(i*pi) + 1 = 0. [6] The above citations of the trilogy of art, poetry, and data, which makes an intensely imaginative interpretation of beauty, explain why I am a dataholic.

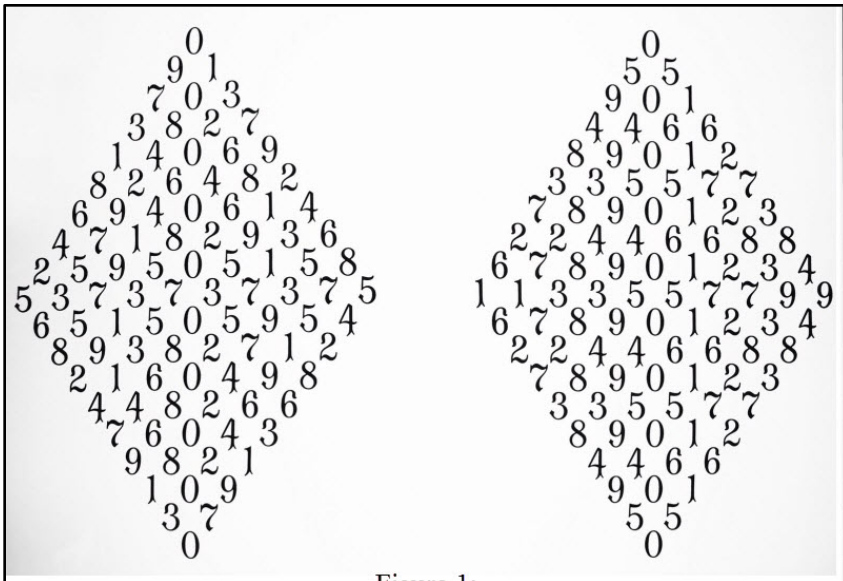

Figure 1:
Parallel Intervals

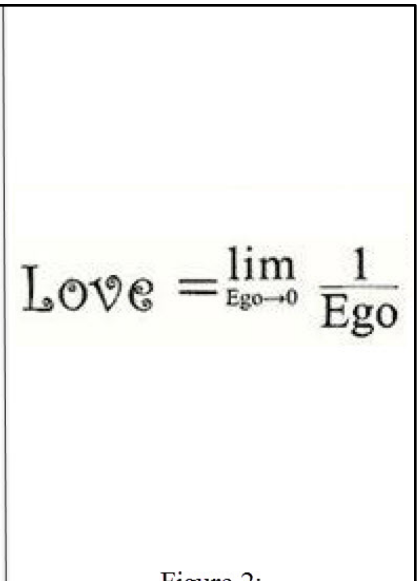

Figure 2:
Love Equation

The purpose of this cur-sorry article is to provide a staircase of twelve steps to ascend upon cracking open a dataset regardless of any application the datawork may entail.

Before painting the numbers by the numbers, penciling dataiku verses, and formulating equation poems, I brush my tabular canvas with four essentials markings for the just out dataset. The markings, first encountered on stairstepping to the rim of the dataset, are:

Step/Marking 1. Determine sample size, an indicator of data depth.

Step/Marking 2. Count the number of numeric and character variables, an indicator of data breadth.

Step/Marking 3. Air the listing of all variables in a format. This permits copying and pasting of variables into a computer program editor. A copy-pasteable list forwards all statistical tasks.

Step/Marking 4. Calculate the percentage of missing data for each numeric variable. This provides an indicator of havoc on the assemblage of the variables due to the missingness. Character variables really never have missing values: We can get something from nothing.

The following eight steps complete my twelve-step program for dataholics, at least for cracking open a freash dataset.

Step 5. Follow the contour of each variable. This offers a map of the variable's meaning through patterns of peaks, valleys, gatherings, and partings across all or part of the variable's plain.

Step 6. Start a searching wind for the unexpected of each variable: Improbable values, say, a boy named Sue; impossible values, say, age is 120 years; and, undefined values due to irresponsibilities like X/0.

Step 7. Probe the underbelly of the pristine cover of the dataset. This uncovers the meanings of misinformative values, such as, NA, the blank, the number 0, the letters o and O, the varied string of 9s, the dash, the dot, and many QWERTY expletives. Decoding the misinformation always yields unscrambled data wisdom.

Step 8. Know the nature of numeric variables. I.e., declare the formats of the numerics as decimal, integer or date.

Step 9. Check the reach of numeric variables. This task seeks values "far from" or "outside" the fences of the data. [7]

Step 10. Check the angles of logic within the dataset. This allows for weighing contradictory values with conflict resolution rules.

Step 11. Stomp on the lurking typos. These lazy and sneaky characters earn their keep by ambushing the integrity of data.

Step 12. Find and be rid of noise within thy dataset. Noise, the idiosyncrasies of the data, the nooks and crannies, the particulars, are not part of the sought-after essence of the data. Ergo, the data particulars are lonely, not-really-belonging-to-pieces of information that happen to be both in the population from which the data were drawn and in the data themselves. Paradoxically, as the analysis/model includes more and more of the prickly particulars, the analysis/model build becomes better and better. Yet, the analysis/model validation becomes worse and worse.

Noise must be eliminated from the data by 1) identifying the idiosyncrasies, and 2) deleting the records that define the idiosyncrasies of the data. Once the data are rid of noise, the analysis/model reliably represents the sought-after essence of the data.

Brush Marking

I request the reader to allow use of my poetic license. I illustrate the reveal of the four data markings by using not only a minikin dataset, but also identifying the variable list itself. At the onset of a big-data project, the sample size is perhaps the only knowable; the variable list is often not known, if so it is rarely copy-pasteable; and for sure, the percentages of missing data are never in showy splendor.

Markings 1 and 2: Determine sample size, and count the number of numeric and character variables.

Consider the otherwise unknown dataset IN in Appendix-A. I run the SAS program in Appendix-B to obtain the sample size and number of variables by numeric-character type. See Table 1, below.

Table 1:
Sample Size and Number of Variables by Type

TYPE	SAMPLE_ SIZE	NUMBER_of_ VARIABLES
Numeric	5	4
Character	5	1

Marking 3: Air the listing of all variables in a copy-pasteable format.

I run the SAS program in Appendix-C to obtain the copy-pasteable variable list, which is found in the Log window. The yellow and pink highlighted text shows the variable list for the copy-paste. See Figure 3, below.

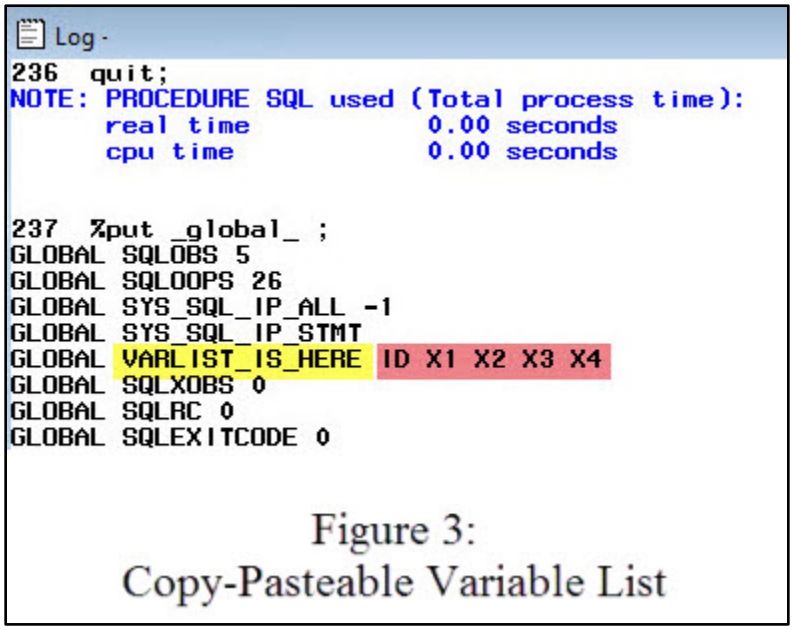

Figure 3:
Copy-Pasteable Variable List

Marking 4: Calculate the percentage of missing data for each numeric variable.

I run the SAS program in Appendix-D to obtain the percentages of missing data for variables X1, X2, X3, and X4. See Table 3, below.

Table 3:
Percent Missing per Variable

PCT_MISSING_X1	PCT_MISSING_X2	PCT_MISSING_X3	PCT_MISSING_X4
20.0%	40.0%	60.0%	80.0%

Conclusion

My name is Bruce Ratner, and I am a functioning dataholic. I do not feel guilty about my dataholism. But, I do lie about my data habits. I need to write dataiku verses, and paint swirling equations to relax. As I did well in kindergarten – "he shares and plays well with others" - I hope my twelve-step program helps others, like me, who love to data.

References

1. Ratner, B., "Shakespearian Modelogue," *Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling, Analysis of Big Data*, 2012.
2. Ratner, B., "The Statistical Golden Rule: Measuring the Art and Science of Statistical Practice", 2013.
3. Mathematicalpoetry.blogspot.com
4. Hispirits.com
5. Einstein
6. Euler
7. Tukey, EDA

Appendices

```
/*** Appendix-A ***/

data IN;
input ID $1. X1 X2 X3 X4;
cards;
1 1 2 3 4
2 1 2 3 .
3 1 2 . .
4 1 . . .
5 . . . .
;
run;

/*** Appendix-B ***/

proc contents data=IN noprint
 out=out1(keep=libname memname nobs type);
run;

proc format;
 value typefmt 1='Numeric' 2='Character';
run;

proc summary data=out1 nway ;
class libname memname type;
id nobs;
output out=out2
(drop=_type_  LIBNAME    MEMNAME
rename=(_freq_=NUMBER_of_VARIABLES NOBS=SAMPLE_SIZE)
);
format type typefmt.;
run;

proc print data=out2 noobs;
run;

/*** Appendix-C ***/

proc contents data=IN
out = vars (keep = name type)
noprint;
run;
proc sql noprint;
select name into :varlist_is_here separated by ' '
from vars;
quit;
%put _global_ ;
```

```
/*** Appendix-D ***/

proc summary data=in;
var x1 x2 x3 x4;
output out=out3(drop=_type_ rename=(_freq_=sam_size)) nmiss=n_miss1-n_miss4;
run;

data out4;
set  out3;
array nmiss  n_miss1-n_miss4;
array pct_miss PCT_MISSING_X1-PCT_MISSING_X4;
do over nmiss;
pct_miss= nmiss/sam_size;
end;
keep PCT_MISSING_X1-PCT_MISSING_X4;

proc print data=out4;
format PCT_MISSING_X1-PCT_MISSING_X4 PERCENT8.1;
run;
```