Some Basic Techniques for Data Quality Evaluation Using SAS

George J. Hurley
Manager, Advanced Analytics

"Mathematics may be compared to a mill of exquisite workmanship, which grinds your stuff to any degree of fineness; but, nevertheless, what you get out depends on what you put in; and as the grandest mill in the world will not extract wheat flour from peas cods, so pages of formulae will not get a definite result out of loose data."

-Thomas Henry Huxley

"Experts often possess more data than judgment."

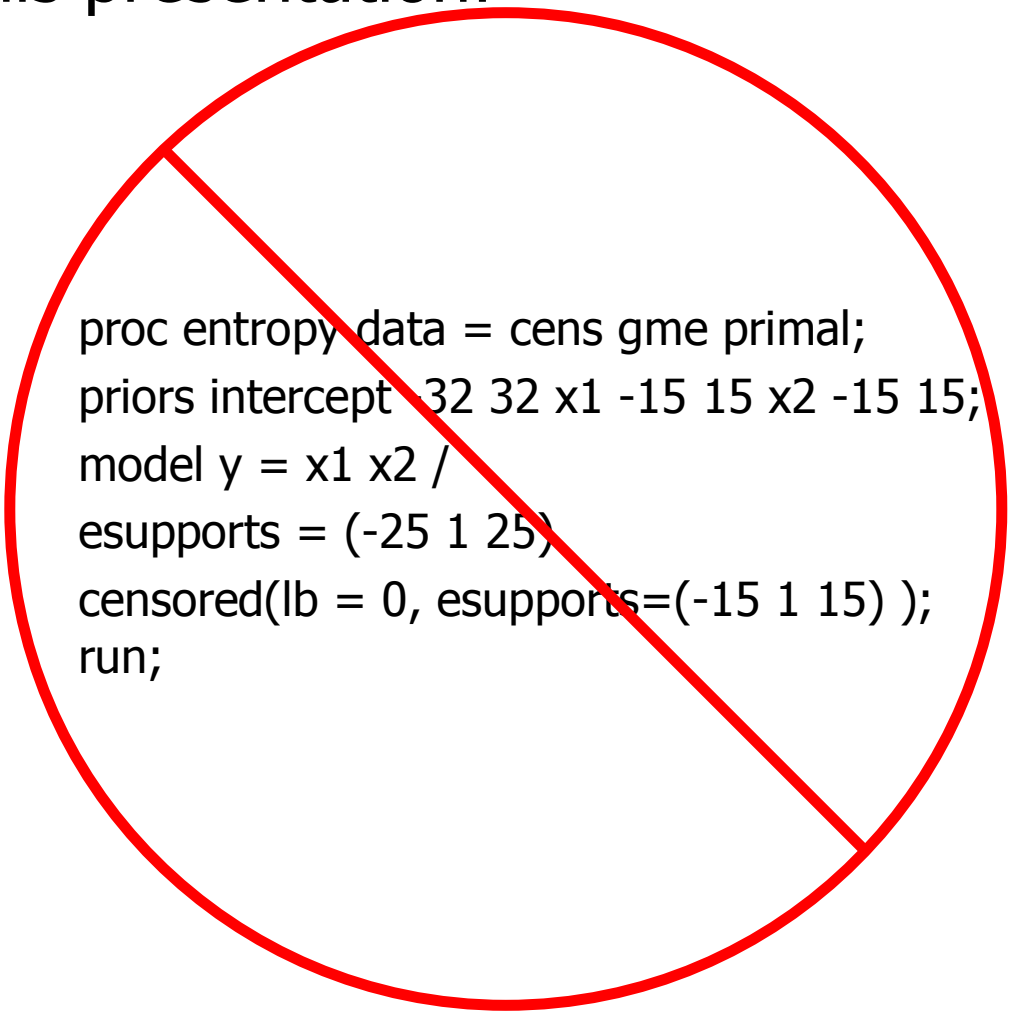
-Colin Powell

"If you torture data long enough, it will tell you anything you want !"

-Unknown Source

Disclaimer

There will be no equations, exotic options, or strange procs used in this presentation!



Agenda

- Previewing the Data – Finding the House
- Dataset Construction – The Physical Examination
- Outliers – Building Codes and Other Details
- What to do now – Do I buy?

Previewing the Data – Finding the House
Intro to Proc Contents

- Find Your House with Proc Contents
- Proc Contents data=xxx;
- Run;

Previewing the Data – Finding the House Example: Header

- You are given dataset containing all US Zip Codes.
  - 545 Observations
  - US Military Zip Codes
- This is not the right dataset.

Data Set Name	SASHELP.ZIPMIL	**Observations**	545
Member Type	DATA	**Variables**	18
Engine	V9	**Indexes**	1
Created	Monday, December 22, 2008 02:08:10 PM	**Observation Length**	512
Last Modified	Monday, December 22, 2008 02:08:10 PM	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	YES
Label	US Military Zipcodes-lat/long NA-assigned missing; Source: zipcodedownload.com April 2008, SAS Release 9.2		
Data Representation	WINDOWS_32		
Encoding	us-ascii ASCII (ANSI)		

Previewing the Data – Finding the House
Example: Engine/Host Dependent Info

- Typically for more advanced users.
- Page Size is the number of bytes of data read into the buffer per I/O operation
- Number of Data Pages is how many pages of this size it takes to store your data.

Engine/Host Dependent Information	
Data Set Page Size	16384
Number of Data Set Pages	19
First Data Page	1
Max Obs per Page	31
Obs in First Data Page	24
Index File Page Size	4096
Number of Index File Pages	5
Number of Data Set Repairs	0
Filename	C:\Program Files\SAS\SASFoundation\9.2\graph\sashelp\zipmil.sas7bdat
Release Created	9.0202M0
Host Created	XP_PRO

Previewing the Data – Finding the House
Example: Alphabetic List of Variables and Attributes

- Variable Names
- Variable Types
- Length
- Format
- Label

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Label
18	ALIAS_CITY	Char	300		Alternate names of city separated by ‖
12	AREACODE	Num	8		Single Area Code for ZIP Code.
13	AREACODES	Char	12		Multiple Area Codes for ZIP Code.
5	CITY	Char	35		Name of city/org
9	COUNTY	Num	8		FIPS county code.
10	COUNTYNM	Char	25		Name of county/parish.
16	DST	Char	1		ZIP Code obeys Daylight Savings: Y-Yes N-No
15	GMTOFFSET	Num	8		Diff (hrs) between GMT and time zone for ZIP Code

Note: This is partial output

Previewing the Data – Finding the House
Example: Indexes and Sort Information

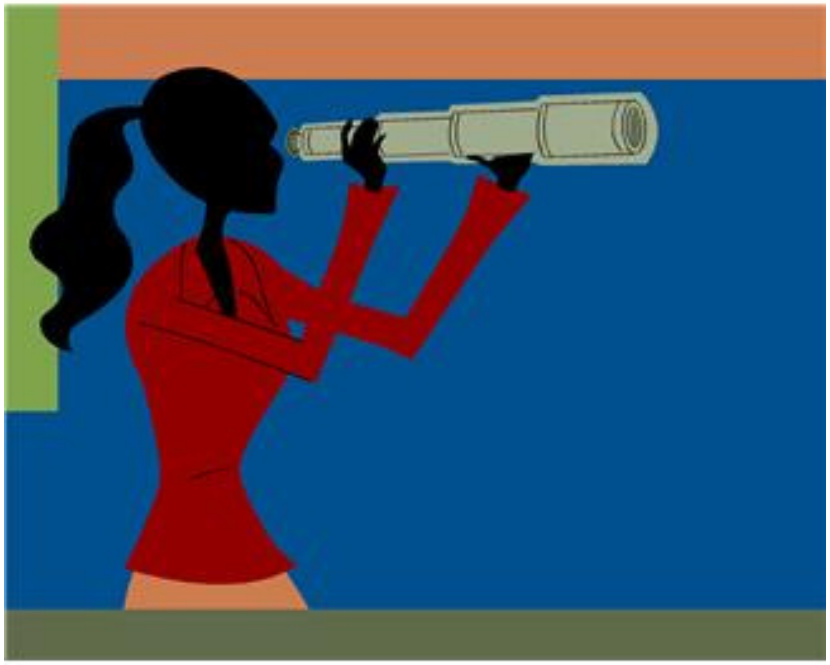
- Index
- Validated flag set to yes by proc sort
  - If no, most procs will check ordering, but index creation faster
  - Gets to no by using sortedby= dataset option
- Character Set affects ordering

Alphabetic List of Indexes and Attributes		
#	Index	# of Unique Values
1	ZIP	545

Sort Information	
Sortedby	ZIP ZIP_CLASS
Validated	YES
Character Set	ANSI

Previewing the Data – Finding the House Example Conclusions

- The data set we were given was not a dataset containing all US Zip Codes.
- Request the correct dataset.



You have to find the house to inspect it.

Dataset Construction – The Physical Examination
Intro to Proc Print

- How does the house look? Use Proc Print!
- Proc Print noobs data=xxx (firstobs= obs=);
- Var W X Y Z … ;
- Run;

Dataset Construction – The Physical Examination
Example: Proc Print Output

- You are given dataset containing all US Zip Codes
- Look at first n observations (5 here)
- Many things can be learned here
  - Look of the Data
  - Data Errors

ZIP	Y	X	ZIP_CLASS	CITY	STATE	STATECODE	STATENAME	COUNTY	COUNTYNM	MSA
00501	40.813078	-73.046388	U	Holtsville	36	NY	New York	103	Suffolk	5380
00544	40.813223	-73.049288	U	Holtsville	36	NY	New York	103	Suffolk	5380
00601	18.165950	-66.723627		Adjuntas	72	PR	Puerto Rico	1	Adjuntas	0
00602	18.383005	-67.186553		Aguada	72	PR	Puerto Rico	3	Aguada	60
00603	18.433236	-67.151954		Aguadilla	72	PR	Puerto Rico	5	Aguadilla	60

AREACODE	AREACODES	TIMEZONE	GMTOFFSET	DST	PONAME	ALIAS_CITY
787	787/939	Atlantic	-4	N	Aguada	
787	787/939	Atlantic	-4	N	Aguadilla	Ramey

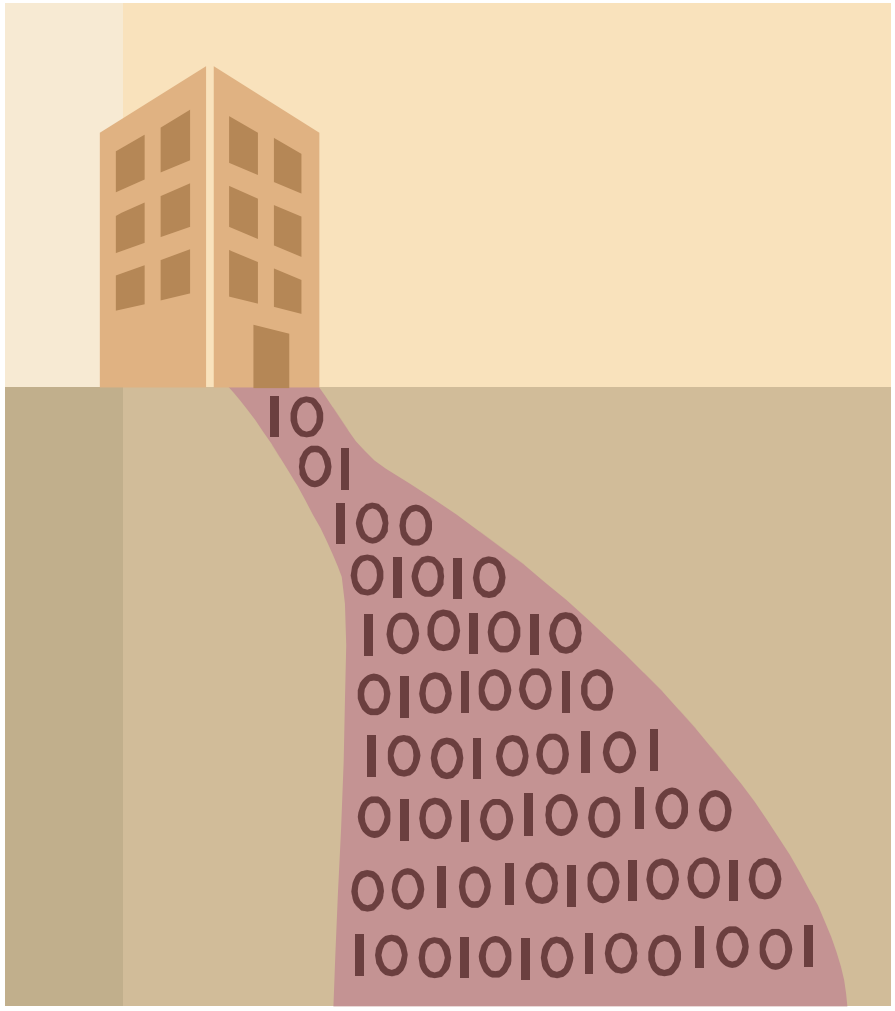
Look for the obvious problems first!

Outliers – Building Codes and Other Details
Intro to Proc Freq and Proc Univariate

- Use Proc Freq and Proc Univariate to See if the House is Up to Code.

- Proc Freq data=xxx;
- Tables v*w x y … ;
- Run;

- Proc Univariate data=yyy;
- Var a b c … ;
- Run;



Outliers – Building Codes and Other Details
Example: Proc Freq

- You have a dataset about cars and their details.
- Use Proc Freq to look at the categorical data.
  - Make*Model
  - Origin

Origin	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Asia	158	36.83	158	36.83
Europe	123	28.67	281	65.50
USA	148	34.50	429	100.00

Table of Type by Cylinders

Type	Cylinders							
Frequency Percent Row Pct Col Pct	3	4	5	6	8	10	12	Total
Hybrid	1 0.29 25.00 100.00	2 0.58 50.00 1.63	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.29 25.00 25.00	4 1.17
Sedan	0 0.00 0.00 0.00	96 27.99 36.64 78.05	6 1.75 2.29 85.71	120 34.99 45.80 79.47	38 11.08 14.50 67.86	0 0.00 0.00 0.00	2 0.58 0.76 50.00	262 76.38
Sports	0 0.00 0.00 0.00	11 3.21 23.40 8.94	0 0.00 0.00 0.00	20 5.83 42.55 13.25	14 4.08 29.79 25.00	1 0.29 2.13 100.00	1 0.29 2.13 25.00	47 13.70
Wagon	0 0.00 0.00 0.00	14 4.08 46.67 11.38	1 0.29 3.33 14.29	11 3.21 36.67 7.28	4 1.17 13.33 7.14	0 0.00 0.00 0.00	0 0.00 0.00 0.00	30 8.75
Total	1 0.29	123 35.86	7 2.04	151 44.02	56 16.33	1 0.29	4 1.17	343 100.00

Frequency Missing = 2

Outliers – Building Codes and Other Details
Example: Proc Freq Follow-Up

- You can follow up on this with a Proc Print
- Proc Print noobs data=cars2;
- Where cylinders in (., 12) or missing(type);
- Run;

Make	Model	Type	Origin	DriveTrain	MSRP	Invoice	EngineSize	Cylinders	Horsepower	MPG_City	MPG_Highway	Weight	Wheelbase	Length
Mazda	RX-8 4dr automatic	Sports	Asia	Rear	$25,700	$23,794	1.3	.	197	18	25	3053	106	174
Mazda	RX-8 4dr manual	Sports	Asia	Rear	$27,200	$25,179	1.3	.	238	18	24	3029	106	174
Mercedes-Benz	CL600 2dr	Sedan	Europe	Rear	$128,420	$119,600	5.5	12	493	13	19	4473	114	196
Mercedes-Benz	SL600 convertible 2dr	Sports	Europe	Rear	$126,670	$117,854	5.5	12	493	13	19	4429	101	179
Volkswagen	Phaeton W12 4dr	Sedan	Europe	Front	$75,000	$69,130	6	12	420	12	19	5399	118	204
		Hybrid	USA		.	.	.	12	.	.	.	10000	.	.

Outliers – Building Codes and Other Details
Example: Proc Univariate Tables

- Use Proc Univariate to look at the numeric data.
  - Weight

Moments			
N	345	Sum Weights	345
Mean	3399.07246	Sum Observations	1172680
Std Deviation	672.387017	Variance	452104.3
Skewness	2.77005502	Kurtosis	26.0441366
Uncorrected SS	4141548176	Corrected SS	155523879
Coeff Variation	19.7814852	Std Error Mean	36.2001001

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	93.89677	Pr > |t|	<.0001
Sign	M	172.5	Pr >= |M|	<.0001
Signed Rank	S	29842.5	Pr >= |S|	<.0001

Basic Statistical Measures			
Location		Variability	
Mean	3399.072	Std Deviation	672.38702
Median	3416.000	Variance	452104
Mode	3175.000	Range	8150
		Interquartile Range	755.00000

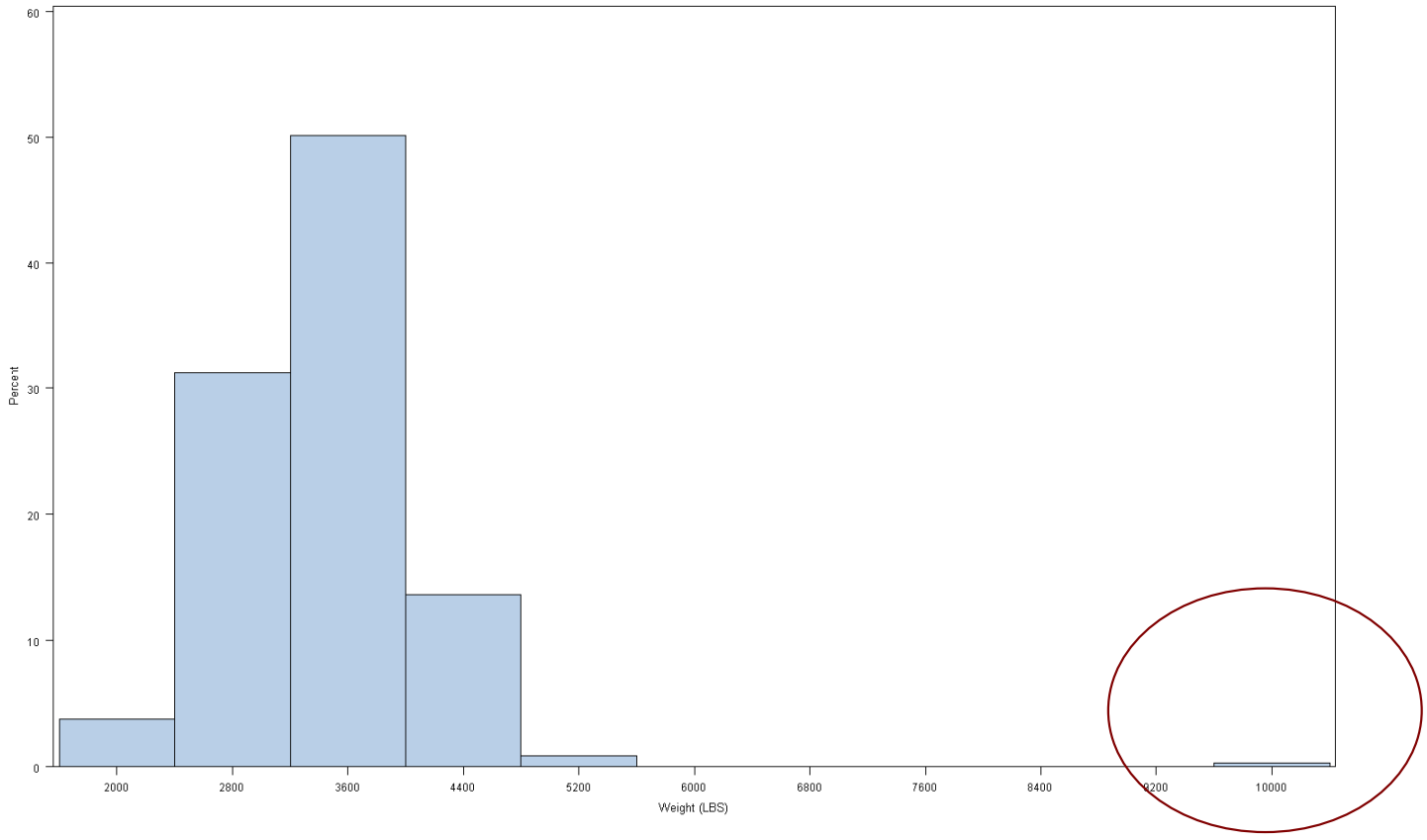
Note: The mode displayed is the smallest of 3 modes with a count of 4.

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	10000
99%	4802
95%	4365
90%	4065
75% Q3	3778
50% Median	3416
25% Q1	3023
10%	2626
5%	2458
1%	2085
0% Min	1850

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1850	116	4675	88
2035	304	4802	168
2055	306	5194	329
2085	305	5399	330
2195	318	10000	345

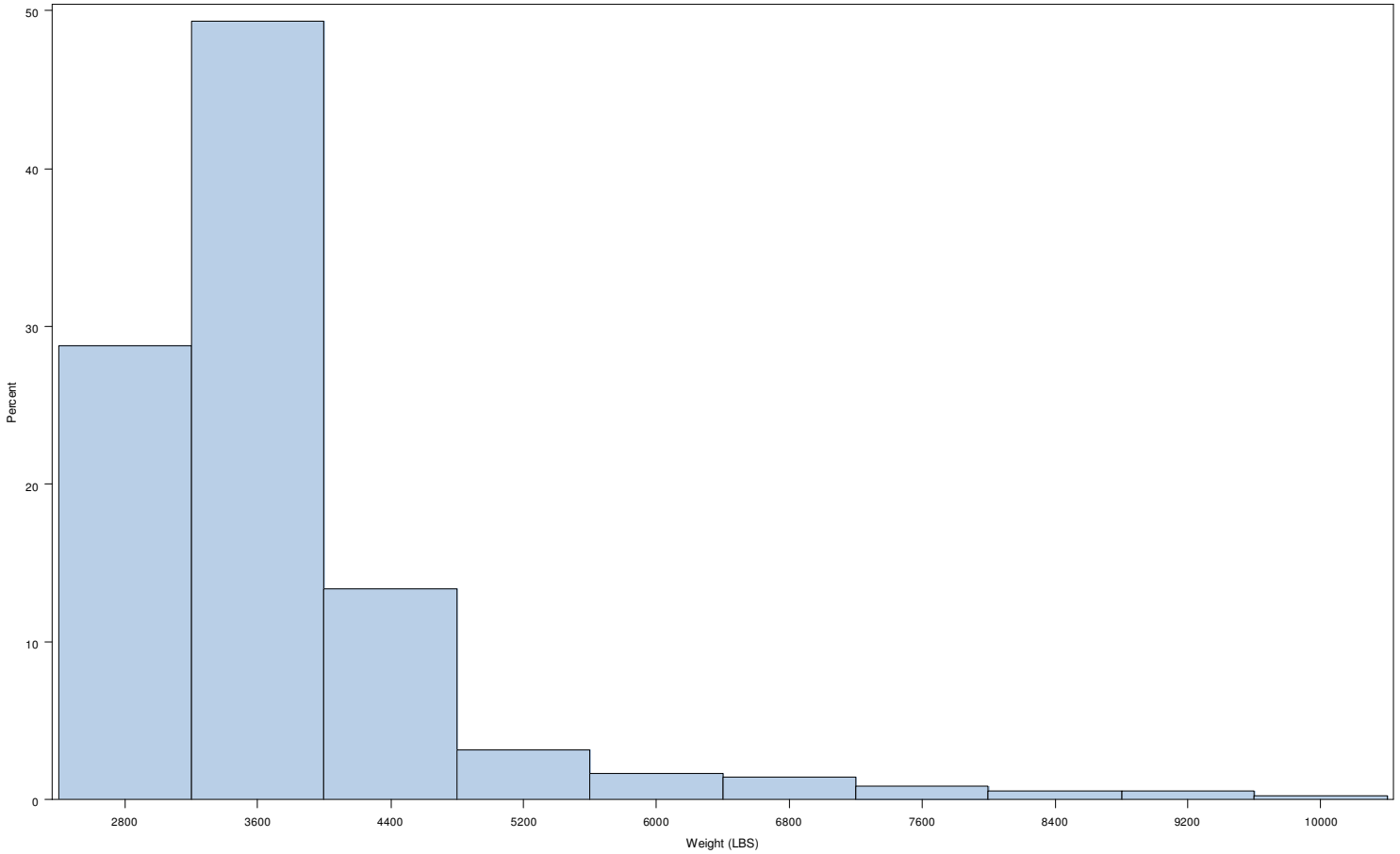
Outliers – Building Codes and Other Details
Example: Proc Univariate Histogram Statement

- Proc Univariate ;
- Histogram weight;
- Run;



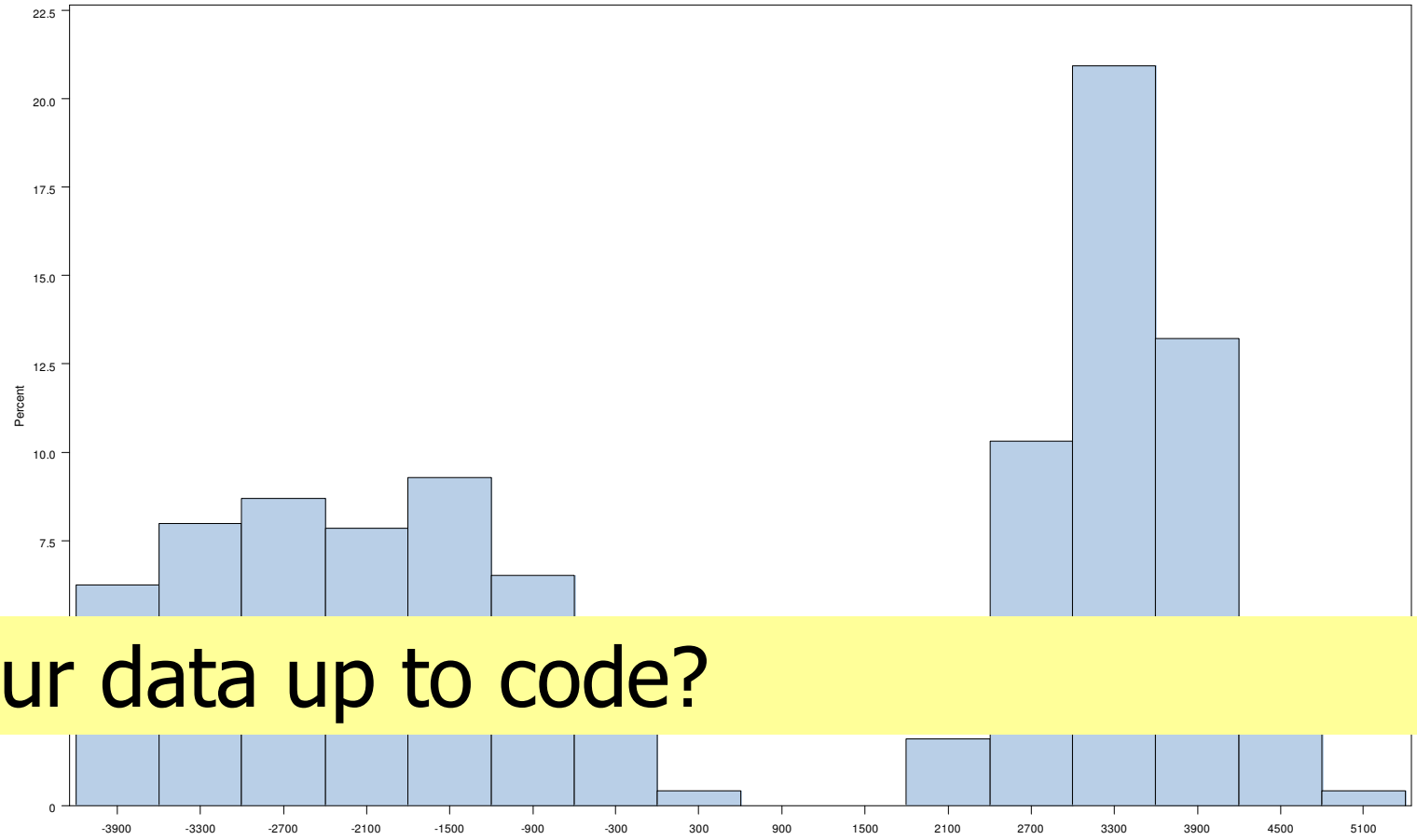
Outliers – Building Codes and Other Details
Example: Proc Univariate Histogram Skewness

- Severe skewness might require alternate analytic methods



Outliers – Building Codes and Other Details
Example: Proc Univariate Histogram Bimodality

- Bimodality will require much thought.
  - Data from two distributions?



Is your data up to code?

What to do now – Do I buy?

- Is the house there?
  - Variables that are Expected are Present
  - Variables of type expected
  - Correct Number of Observations
- Is the roof missing?
  - Data should be populated
  - Strange observations
- Does the data pass code?
  - Outliers
  - Skewness
  - Bimodality

Once remediated, it is time to buy!

Summary

- You have to find the house to inspect it.
- Look for the obvious problems first!
- Is your data up to code?
- Once remediated, it is time to buy!