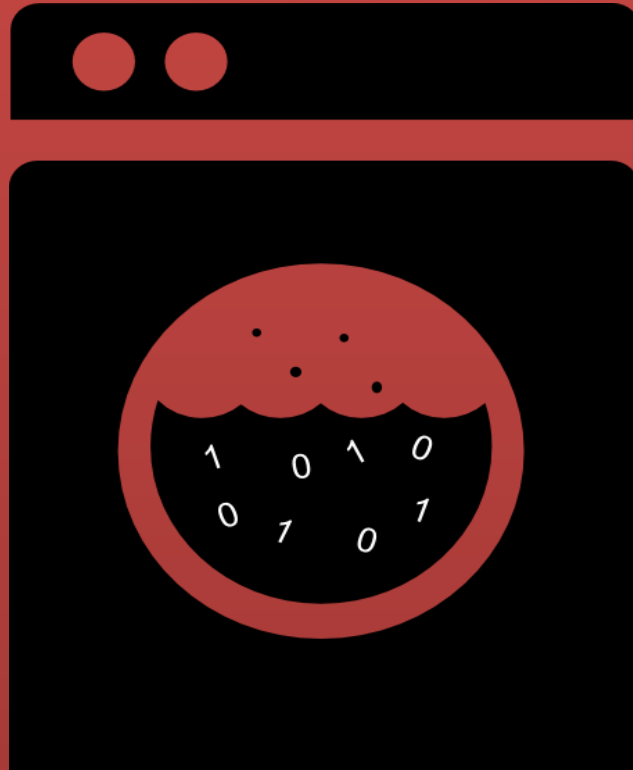*PRACTICAL*

DATA CLEANING

19 Essential Tips to Scrub Your Dirty Data
(and keep your boss happy)

LEE BAKER

2nd Edition

PRACTICAL

DATA

CLEANING

19 Essential Tips to Scrub your Dirty Data

(and keep your boss happy)

LEE BAKER

CEO
Chi-Squared Innovations

TABLE OF CONTENTS

Preface to 2nd Edition

The success of the 1st edition of Practical Data Cleaning has been overwhelming and has come as a bit of a shock – there have been thousands of downloads in the first couple of months since launch. We had no idea it would be so popular!

Firstly, we wish to thank all of you that took the time to download and read it. We hope you found it useful and got something of real value out of it.

Secondly, we wish to thank all of you that have interrupted your busy schedules to get in touch with us. Your words of support have touched us and inspired us to continue writing. Please keep sending me your feedback – I love reading it whether it's good or bad!

Finally, a quick note on the 2nd edition. Many of you wanted to know more about the different types of data that I sort of hinted at in Tip #18, so I've gone into more detail and enhanced the info. I think it's better, but – as ever – I'll let you be the judge.
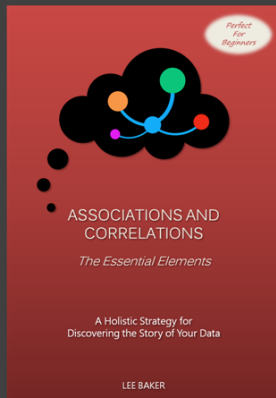
Oh, yes – and I've added a fresh new cover. I hope you like it.

Come and tell me, I'd love to hear from you…

ebookfeedback@chi2innovations.com

INTRODUCTION

Don't Panic !!!

We live in an increasingly rich world of data – the amount of data that currently exists doubles every 18 months.

That's a phenomenal rate of growth and we're just at the beginning of an incredible journey creating awesome intelligent applications that can handle these unimaginable amounts of data automatically.

This **Big Data** movement is happening at one end of the scale.

At the other, there are millions of people around the globe collecting and working with **Small Data** – data that is small enough to fit in an Excel spreadsheet and store on a floppy disc (remember those?).

It doesn't matter whether you're a scientist or an entrepreneur, in academia or in business, if you're collecting data to try to answer some questions then **you need to understand the fundamentals**.

You'll likely spend a lot of time observing, measuring, counting, classifying and quantifying what you see, and once you've collected your data you're going to have to analyse it.

But let's not get too far ahead of ourselves…

Before you can get any answers you're going to have to:

- • Collect
- • Record & Store
- • Clean & Classify

The textbooks tend not to dwell on the practical issues too much because, well, to be honest, it can get quite messy, but these are vitally important steps and you really do need to know how to do them properly if you're going to **get the most out of your data**.

So let's rewind to the beginning and see what we can do to get you off to a good start...

Here are 3 rules to start off with:

1. Don't Panic !!!
2. Start thinking about the data *before* you start collecting it
3. Make a personal vow to understand the basics of data

Just so's you know, **you are free to share this eBook** with anyone – as long as you don't change it or charge for it (the boring details are at the end).

Ready?
OK, let's go...

CHAPTER

1

Data Collection

Tip #1

Record Data on Paper First...

So you've got your hypothesis (theory, idea or hunch). Once you've decided what data you need to collect, the first thing you should do is **design a paper-based form to store all your data** (assuming that at least some of your data is going to be recorded by hand).

Keep it simple, print it out, then manually record your data with pen and paper. One form per case/patient/customer/test-tube, etc..

Physical Assessment:

Inprocessing BMI: _____

Current Weight: _____   Current BMI: _____

Heart Rate _____ BP _____ RR _____ T _____ LOC: Yes   No

Tip #2

...Then Transfer it to an Electronic Medium

We may be living in an electronic world, but ultimately you need a system where you (or anyone else) can **follow the data trail from beginning to end** and – more crucially – **from end to beginning**.

From time to time you WILL make a mistake with the data, so it is vitally important that you design a method that will let you spot and **rectify the mistake by going back through all the steps** until you find the error.

So now you have your data recorded on paper you need to transfer it into an electronic system. More than likely this will be either Microsoft Excel or Access.

In general, Excel is more common and easier to use, and has the added advantage that you can manipulate the data and do some simple analyses right there without having to export your data.

Most data is stored in Excel (in 7 years as a medical statistician I was only once given data in Access – all the other times it was in Excel), so we'll go with that from here on in...

Tip #3

Enter Your Data on a Single Worksheet Whenever Possible

Trying to sort your data when it is spread across multiple worksheets can lead to all sorts of problems, so try to avoid it whenever you can - **keep all your data on a single worksheet**.

Excel 2003 limits the number of usable worksheet rows and columns, and these limits are large enough for most datasets.

If you need higher limits you can use Excel 2010 or 2013.

Excel 2003 limits:
- 65,536 rows
- 256 columns

Excel 2010 and 2013 limits:
- 1,048,576 rows
- 16,384 columns

Tip #4

Use a Unique ID Column

You'll likely have to sort your data many times and by different columns, so you're going to need a way of restoring the original order.

Use column A as a unique identifier to insert consecutive numbers starting from 1. It may be simple, but it's very effective.

	A
1	**UniqueID**
2	1
3	2
4	3
5	4

When you've put your Unique IDs into column A, go back to your original paper sheets and write the Unique ID there as well.

Trust me – you'll thank me for this tip later…

Tip #5

One Column per Variable

Each variable should have… oh, hold on a minute, what's a variable?

Well, simply put, these are the things that **can change** or **can be changed** as part of your study. In short, these are all the pieces of information that you are observing, measuring, counting and collecting, like age, gender, distance, temperature, etc..

You can find more information on data, data types and more in our **Discover Data Blog Series**.

Where were we? Ah yes…

Each variable should have its own column, and each variable should correspond to just **one piece of information**.

	A	B	C	D	E	F
1	UniqueID	Variable 1	Variable 2	Variable 3	Variable 4	Varial
2						
3						
4						
5			Use one column per variable			
6						

If you're entering the age of a patient, then just enter their age, don't enter their date of birth in the same column or cell.

If you want to record their age and DOB, then use 2 separate columns.

If you're recording a composite variable made up of 2 or more constituent parts, like Body Mass Index – made up of Height and Weight – then record them in separate columns.

You can always combine them into a single variable later.

Tip #6

Row 1 is the Variable Name

Eventually you'll need to analyse your data and you may need to export it to a statistical program.

The standard for pretty much all commercial stats programs is that **the first row is reserved for the name of the variable** and all other rows for the data.

So don't be tempted to use rows 2, 3 and 4 as well as row 1 for the variable name.

It might keep everything looking nice and tidy in Excel, but it will only create more work for you later.

	A	B	C	D	E
1	UniqueID	Gender	Age	Height	Weight
2					
3					
4					

Tip #7

Every Cell Should Have Something In It

What do empty cells tell you?

- waiting for more information?
- data not recorded?
- original data incorrect?

An empty cell is just a great big question mark and tells you nothing.

Worse still, incomplete datasets give reviewers a reason to whack you about the head with a metaphorical stick (and believe me they will – I've been there many times…).

So make sure that something is entered in every cell.

It is quite common to **use 'illegal' numbers as codes** to give you information, so where the entries for a variable can only be positive values (like age or height), we can use codes such as:

	A	B
1	**My Variable Code**	**What It Really Means**
2	-1	Data not recorded
3	-2	Waiting for lab
4	-3	Dave screwed it up, the idiot…
5		

If negative numbers aren't useful, then **use letters a, b, c**, etc..

If you're not comfortable entering something in cells that strictly shouldn't be there (after all, you are going to have to clean them up later before you can analyse your data), then use Excel's Comment feature.

I tend to use this sparingly, but that's just me…

Tip #8

Keep Great Notes

When using codes you'll need to keep notes to tell you what the codes mean.

Keep the codes and notes in a different spreadsheet.
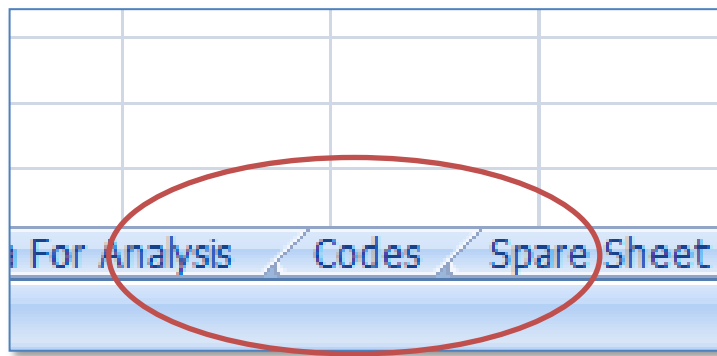
While we're on the subject, it's really important to:

KEEP GREAT NOTES !!!

You're likely not the only person that will ever work with this dataset, so **get used to writing stuff down**.

Explain what the project is all about, the questions you're trying to answer, why you're collecting this data and how you're going to get the answers you're looking for.

Explain how you measured things and under what conditions.

If more than one person is collecting data, then explain who, what, where, when, why and how.
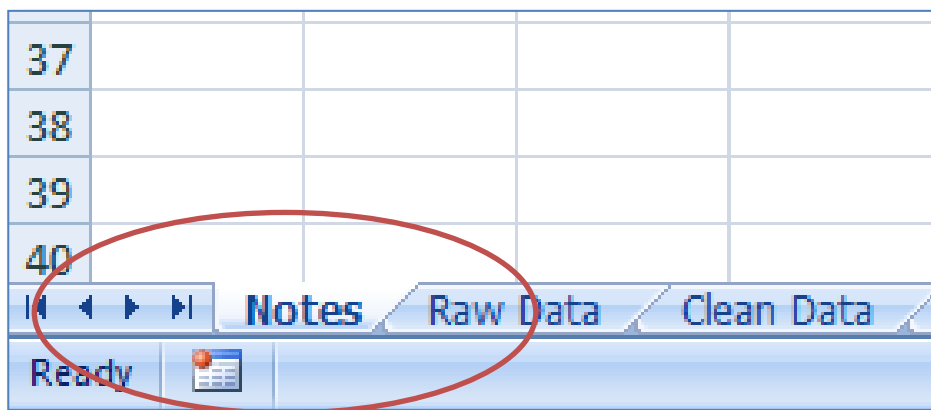
This will be the document that explains all the important stuff about your dataset, so **write it down**.

If there's too much information to comfortably put into an Excel spreadsheet, then a Microsoft Doc will be just fine – and keep it in the same folder as the dataset.

Tip #9

Be Consistent

There's nothing worse than getting a dataset that takes a fortnight to clean because data entry has not been consistent.

B
Variable 1
Positive
Pos
POS
pos
positive
+ve
+

By that I mean make sure that if the entry for a variable should be 'Positive', then make it 'Positive' and not some other variation:

It's hard enough correcting speeling missteakes and typos without also having to correct things that were deliberately entered differently.

Restrict the number of people that can enter data to cut down on these issues, and make it clear what your **data entry standards** are.

Tip #10

Don't Guess

Data should be entered as accurately as possible.

Don't guess, approximate, round up or down !!!

Enter the value exactly as registered on paper.

Use Excel's functions to round your data, but don't do calculations in your head, on paper or in a calculator – you'll make mistakes which can be difficult, if not impossible, to spot later.

	✓ *fx*	=round(	
	C	**D**	**E**
	Variable 2		
	0.42007673	=round(	
	0.571399325	ROUND(**number**, num_digits)	
	0.372793063		
	0.118264622		
	0.642600129		
	0.221575316		
	0.46627778		
	0.456048014		
	0.067675583		

Tip #11

Zero is a Real Number

Don't enter the number Zero into a cell unless what has been measured, counted or calculated results in the answer Zero.

I've often received datasets with lots of zeros and when I asked, the zeros meant 'I don't have data for this'.

The problem is that if you want to calculate something, like the mean, then all the zeros will be used in the calculation and you will get an **inaccurate answer** – or one that is **just plain wrong**!

I see you're entering a zero.

Are you sure this is really a zero or are you just storing problems for yourself later?

2

Data Cleaning

If you've collected all your own data and you've been *very* careful you might just have a perfect dataset.

Well done!

Personally I've never seen a perfect dataset – it is the rarest of creatures.

Most likely you will have to clean your data before you can start to analyse it.

Yet again the textbooks will give you little practical advice here, so let's dive in and set a few ground-rules that will help you save time and keep your boss happy…

Tip #12

Make a Copy

You've got a 'raw' dataset that is essentially an electronic copy of all the paper-based data you have collected.

If you have made an entry error in the electronic copy you can always check back to the original paper copy.

When you move on to the data cleaning you're going to be changing the data and **you need to be able to undo any cleaning mistakes** you've made, and trust me – you're going to make a few.

So **create a duplicate worksheet** of your dataset.

Believe it or not, this is one of the most important steps in data cleaning.

Call the original one 'Raw Data' and the new one 'Cleaning In Progress' until you've finished cleaning, then you can change the name to 'Clean Data'.
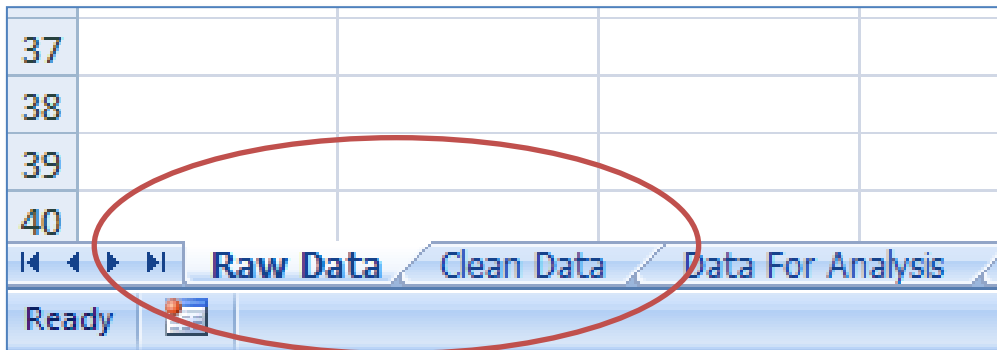
Oh yes – and make sure both worksheets have got the Unique ID column.

Tip #13

Clean Your Data in a Separate Worksheet

When cleaning an individual column of data you'll use a variety of different tools built into Excel, like 'Find And Replace'.

When you use 'Find And Replace' will it operate only on the selected column or on the whole worksheet?

Are you sure?

Really, really sure?

Do all of the in-built functions work in the same way?

Get the answer wrong and you'll find that you've just introduced errors across your entire dataset with **no easy way to undo them** (hitting 'Undo' doesn't work here).

So when you want to clean a single column of data, **copy that column into a spare worksheet** and work on it there.

When you're done you can copy back, replacing the previous uncleaned column.
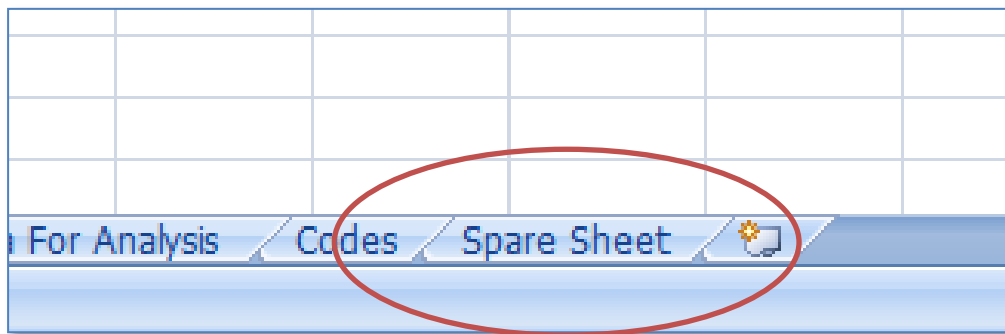
It may take you a little more time, but it's worth it – **mistakes can be very costly** and it pays to head them off at the pass.

Oh, I do hate that cliché…

Tip #14

Report Errors Back to the Original Source

It makes no sense cleaning your data if the same data has to be cleaned in exactly the same way time and time again.

If you're using a shared dataset, such as a departmental database, make sure you **report back to the original source** any errors that you've found.

Then, next time you have to analyse some more data from the same source you'll have a lot less cleaning to do.

Tip #15

Use Excel Functions to do the Hard Work...

Whenever possible, try not to clean data manually.

One of the biggest sources of spelling errors, typos and incorrect entries comes from manual entry, so why use the same method that got you into trouble in the first place?

Excel has a shed-load of **functions that can help with data cleaning**, so use them.

If you have a text-based column, use Excel's 'Remove Duplicates' function.

The result will be a list of all the items that you have in that column.

You can then use 'Find and Replace' to correct misspelled entries, including correcting entries with the wrong case, like 'case', 'Case' or 'CASE'.

Tip #16

...And Use Excel Formulae to do the Even Harder Work

I cannot tell you how many weeks of my life I have lost – that I will never get back – trying to find the source of error that turn out to be a space at the beginning or end of the data in a cell.

You can't see it, but it's still there and it can wreak havoc when you start to do analyses.

Excel ignores spaces, so they can be incredibly difficult to detect, but other analysis and stats packages don't ignore them and they treat the entry as something different.

Spaces are the bane of my life!!!

So what to do?

Excel has a few different formulae that can be used to detect and trim spaces and other unwanted characters, like:

- TRIM()
- CLEAN()
- SUBSTITUTE()

so **learn how to do simple coding in Excel** and use these – and other – formulae.

I promise – it will definitely be time well spent!

	A	B	C	D
1	**My Variable 1**		**My Variable 1 (Cleaned)**	
2	Correct		=CLEAN(A2)	
3	Correct			
4	Correct			
5	Correct			

3

Data Codification & Classification

So you now have a perfectly clean dataset, but you still have some work to do before you start analysing it.

It's important that you **note what your codes mean** – after all, they're not a secret are they?

Say you've entered the data for a variable as 1, 2 or 3.

What does that mean?

- Small, Medium or Large?
- Pig, Sheep or Goat?

It matters because you shouldn't be expected to remember all the details of how, what and why you coded your data that way.

Tip #17

Keep a Code Sheet

Keep your codes in a separate worksheet and name it 'Codes'. For each column **make a note of what codes you've used** and what they really mean.

If you've used additional codes using 'illegal' entries such as negative numbers or letters, make a note of what they mean too.

When you come back to the dataset after a couple of weeks away from it, you'll be glad you got organised like this.

You'll also make your boss, colleagues and local friendly statistician happy too, and **that's never a bad thing**…

	A	B	C	D	E	F	G
1	Variable	0	1	2	3	-1	-2
2	Gender	N/A	Male	Female		Not Recorded	Incorrect
3	Menopause	N/A	Pre-	Peri-	Post-	Not Recorded	Incorrect
4	Cancer	No	Yes			Not Recorded	Incorrect
5	Estrogen Receptor	Negative	Positive			Not Recorded	Incorrect
6	Tumour Grade	N/A	Grade 1	Grade 2	Grade 3	Not Recorded	Incorrect

Tip #18

Identify Your Data Types

When you get to the analysis stage you'll need to know your data types – Ratio, Interval, Ordinal and Nominal – so take a little time to decide which of these are appropriate for each variable, and note this down in your code sheet.

Not sure what these are? OK, then let's take a little step back.

There are two types of data:
- **Quantitative**
- **Qualitative**

Data is quantitative when it is measured with a ruler, jug, weighing scales, stop-watch, thermometer and so on.

It is qualitative when it is observed and placed into categories, such as gender (male, female), health (healthy, sick), opinion (agree, neutral, disagree).

Quantitative and qualitative data can be sub-divided into four further classes of data:

- Quantitative (measured)
    - Ratio
    - Interval
- Qualitative (categorised)
    - Ordinal
    - Nominal

The difference between them can be established by asking just three questions:

Ordered

Can some sort of progress be detected between adjacent data points or categories or can the data be ordered meaningfully?

Equidistant

Is the distance between adjacent data points or categories consistent?

Meaningful Zero
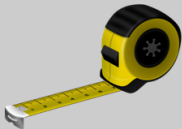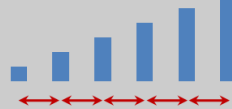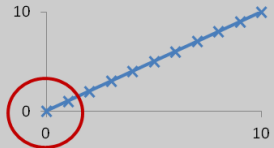
Does the scale of measurement include a unique, non-arbitrary zero value?

Nominal Data has the following properties:

NOMINAL DATA

| Measured | Ordered | Equidistant | Meaningful Zero |

Chi-Squared Innovations   www.chi2innovations.com

You can differentiate between nominal categories based only on their names, hence the title 'nominal' (from the Latin *nomen*, meaning 'name').
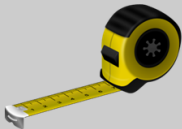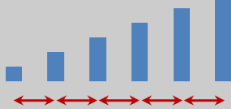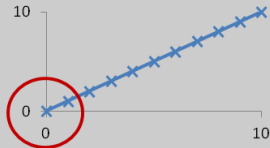
If the categories are descriptive (Nominal), like 'Pig', 'Sheep' or 'Goat', it can be useful to separate each category into its own column, such as Pig [Yes; No], Sheep [Yes; No], and Goat [Yes; No].

Ordinal Data has the following properties:

ORDINAL DATA

Measured	Ordered	Equidistant	Meaningful Zero
✗	✓	✗	✗

Chi-Squared Innovations   www.chi2innovations.com

With ordinal data, the categories can be ordered (1st, 2nd, 3rd, etc. – hence the name 'ordinal'), but there is no consistency in the relative distances between adjacent categories.

Mathematically, you can make simple comparisons between the categories, such as more (or less) healthy, agree more or less, etc., and since there is an order to the data, we can rank them and compute the median to find the central value.
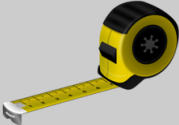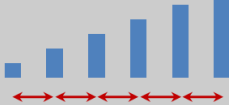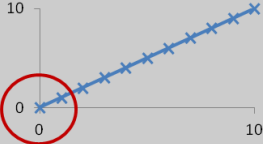
Interval Data has the following properties:

INTERVAL DATA

Measured	Ordered	Equidistant	Meaningful Zero
✓	✓	✓	✗

Chi-Squared Innovations   www.chi2innovations.com

Interval data can be continuous (have an infinite number of steps) or discrete (organised into categories), and the degree of difference between items is meaningful (their intervals are equal), but not their ratio.
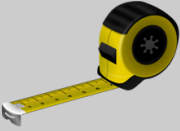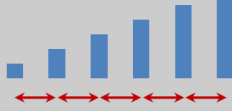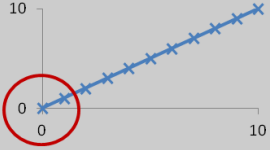
Mathematically you can compare the degrees of the data (equality/inequality, more/less) and add/subtract the values, such as '6pm is 3 hours later than 3pm'. However, you cannot multiply or divide the numbers, so you can't say '6pm is twice as late as 3pm'.

Ratio Data has the following properties:

As with interval data, ratio data can be continuous or discrete, and differs from interval data in that there is a non-arbitrary zero-point to the data.

Ratio data are the best to deal with because all possibilities are on the table. You can find the central point of the data by using any of the mode, median or mean and use all of the most powerful statistical methods to analyse the data.

4

Data Integrity
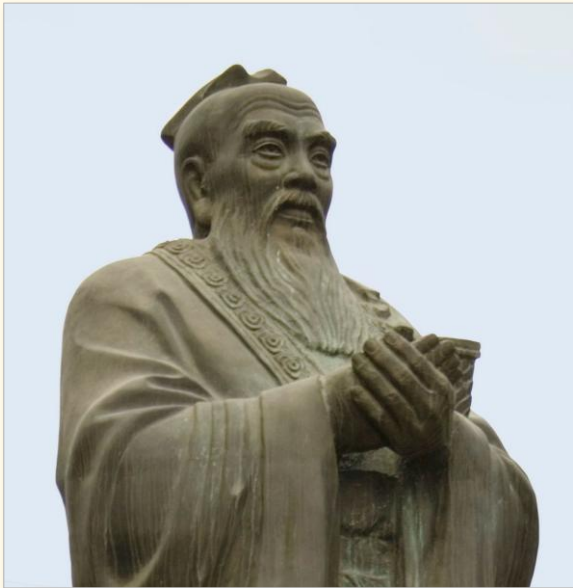
A man who has
committed a mistake
and does not correct
it is committing
another mistake

Confusius

Just because you've got a perfectly clean, classified, codified and organised dataset, it doesn't mean that the data are correct.

Real life follows rules, and your data must too !!!

I once discovered that we had the oldest man in the world currently being treated in the hospital.

At well over 300 years old he'd clearly had 'a good innings'.

In the dataset I was analysing, the difference between his date of birth (somewhere in the 18th century) and date of hospital admission (21st century) meant that he was very old indeed.

Or perhaps his DOB wasn't quite right...

The error in his DOB couldn't be detected by standard error-checking in Excel because it was a perfectly legitimate date.

Tip #19

Check That Your Data is Sensible

Sometimes, putting together 2 or more pieces of data can reveal errors that otherwise can be difficult to find, so it is sensible to **do a few simple calculations** on each variable to check that the data conform to sensible rules, such as:

- Calculate the minimum, maximum and mean
- Keep a count for each variable and each category
- Check differences between dates

Making these checks (in a separate worksheet!) lets you find outliers, such as people who have a negative age or are several hundred years old, and gives you **a good feel for your data**.

Something doesn't feel right about the answers?

Then dive back in and take a look.

There really is no substitute for getting your hands dirty!

	A	B	C	D
		Gender	Age (y)	Height (m)
1				
2	Count	2105	2002	2212
3	Minimum	0	-312.3	1.31
4	Mean	0	53.2	1.73
5	Maximum	0	93.6	19.53
6	Negatives	27	32	0
7	Zeros	15	0	12

CHAPTER

5

Work Smarter, Not Harder

Bonus Tip

Automate Your Data Cleaning

Even if you've followed all of the tips here, it will still take you days or weeks to clean your dataset – and that's if it's small.

Cleaning large datasets can take months or longer.

Wouldn't it be great if you could **clean your data automatically** in minutes rather than weeks or months?

We think so, which is why **this is exactly what we've done**.

We've created a fully automated data cleaning tool – **DataKleenr** – that is:

- ✓ **Fast**
- ✓ **Simple**
- ✓ **Accurate**

Better still, it is *intelligent,* so the more data it cleans the faster and more accurate it becomes.

And you might even be able to **use it for FREE**

- ✓ **Save time AND money**
- ✓ **Eliminate stress**
- ✓ **Complete your research sooner**

So **check out DataKleenr**, then **come and talk to us**.

We'd love to hear from you !!!

SUBSCRIBE

Well, I hope you enjoyed this ebook.

Why not learn more by subscribing to our free newsletter:

Decimal Points – The CSI Buzz

You never know, it might not be the worst thing you do today...

Discover More !!!

We will never share your data – EVER !

LOGO

COPYRIGHT

The copyright in this work belongs to the author, who is solely responsible for the content.

Please direct content feedback or permissions questions to the author.

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs License.

You are given the unlimited right to print this manifesto and to distribute it electronically (via email, your website, or any other means).

You can print out pages and put them in your favourite coffee shop's windows or your doctor's waiting room.

You can transcribe the author's words onto the sidewalk, or you can hand out copies to everyone you meet.

You may not alter this manifesto in any way, though, and you may not charge for it.

PRACTICAL
DATA CLEANING

Lee
Baker

Lee Baker is an award-winning software creator with a passion for turning data into stories.
A proud Yorkshireman, he now lives by the sparkling shores of the East Coast of Scotland.

Physicist, statistician and programmer, child of the flower-power psychedelic '60s, it's amazing he turned out so normal!

Turning his back on a promising academic career to do something more satisfying, as the CEO and co-founder of Chi-Squared Innovations he now works double the hours for half the pay and 10 times the stress - but 100 times the fun!

X² ChiSquared Innovations