实用

数据清洗

19个必备技巧彻底清理您的脏数据
（并让您的老板满意）

李贝克

实用

数据清洗

19个必备技巧彻底清理您的脏数据

（并让您的老板满意）

李贝克

首席执行官
卡方创新

目录
内容

引言：不要惊慌！！！

1：数据收集

2：数据清洗

3：数据编码和分类

4：数据完整性

5：工作更智能，更高效

关于作者





引言

不要惊慌！！！

我们生活在一个越来越丰富的数据世界中-目前存在的数据量每18个月翻一番。

这是一个惊人的增长速度，我们只是在开始一个令人难以置信的旅程，创建可以自动处理这些难以想象的大量数据的令人敬畏的智能应用程序。

这个大数据运动正在发生在一个极端。

另一方面，全球各地有数百万人收集和处理小数据- 这些数据足够小，可以放在Excel电子表格中并存储在软盘上（还记得那些吗？）。

无论您是科学家还是企业家，在学术界还是商界，如果您正在收集数据来解答一些问题，那么您需要了解基本原理。

您可能会花费大量时间观察、测量、计数、分类和量化您所看到的内容，一旦收集到数据，您将需要对其进行分析。

但是，让我们不要过于急于求成...

在您能够得到任何答案之前，您需要做以下几点：

- 收集
- 记录和存储
- 清洗和分类

教科书往往不会过多涉及实际问题，因为，坦率地说，这可能会变得非常混乱，但这些步骤非常重要，如果您想充分利用您的数据，那么您真的需要知道如何正确执行它们。

所以让我们回到开始，看看我们能做些什么来让你有一个良好的开端...

以下是3条起步规则：

1. 不要惊慌！！！
2. 在开始收集数据之前，先思考一下数据
3. 立下个人誓言，要了解数据的基础知识

只要你知道，你可以自由地与任何人分享这本电子书，只要你不改变它或收费（无聊的细节在结尾处）。

准备好了吗?

好的，我们开始吧...

章节

1

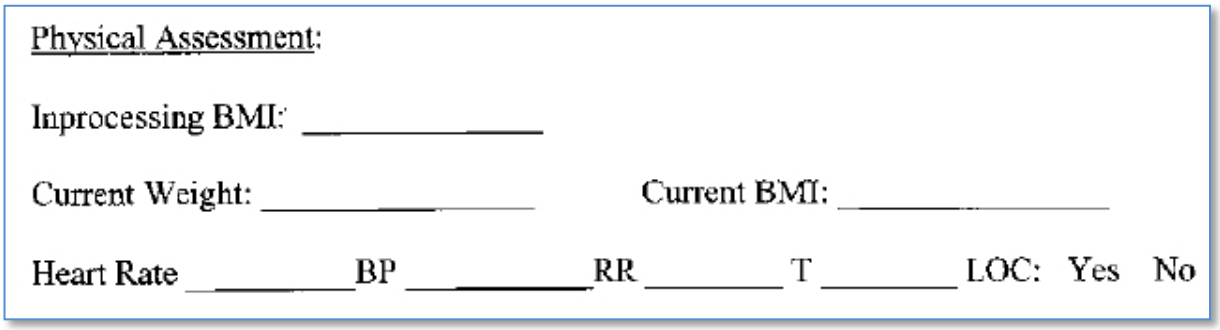
数据收集

技巧 #1

首先，将数据记录在纸上...

所以你有了你的假设（理论、想法或直觉）。 一旦你决定了需要收集哪些数据，你应该做的第一件事是设计一个纸质表格来存储所有的数据（假设至少有一部分数据将会手工记录）。

保持简单，打印出来，然后用笔和纸手动记录你的数据。 每个案例/患者/客户/试管等都需要一个表格。


Physical Assessment:

Inprocessing BMI: _____

Current Weight: _____ Current BMI: _____

Heart Rate _____BP _____RR _____ T _____ LOC: Yes No

提示 #2

...然后将其转移到电子媒介上

虽然我们生活在一个电子世界，但最终您需要一个系统，您（或其他人）可以从头到尾跟踪数据的路径，并且更重要的是，可以从末尾到开头追溯数据。

偶尔您可能会在数据上犯错误，因此设计一个方法让您可以通过回顾所有步骤来发现并纠正错误非常重要。

现在，您已经将数据记录在纸上，需要将其转移到电子系统中。 很可能会选择Microsoft Excel或Access。

一般来说，Excel更常见且更易于使用，并且具有一个额外的优势，即您可以在其中操作数据并进行一些简单的分析，而无需导出数据。

大多数数据存储在Excel中（作为一名医学统计学家，我在7年的时间里只有一次收到Access中的数据，其他所有时间都是Excel），所以我们从这里开始使用Excel...

技巧 #3

尽可能在一个工作表上输入您的数据

当数据分布在多个工作表中时，尝试对其进行排序可能会导致各种问题，所以尽量避免这种情况 - 将所有数据保留在一个工作表中。

Excel 2003限制可用工作表的行数和列数，而这些限制对于大多数数据集来说已经足够大了。

如果您需要更高的限制，可以使用Excel 2010或2013。

Excel 2003限制：

- 65,536行
- 256列

Excel 2010和2013限制：

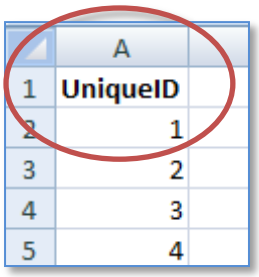
- 1,048,576行
- 16,384列

技巧 #4

使用唯一ID列

您可能需要多次按不同的列对数据进行排序，因此您需要一种恢复原始顺序的方法。

使用列**A**作为唯一标识符，插入连续的数字，从1开始。 这可能很简单，但非常有效。



	A
1	UniqueID
2	1
3	2
4	3
5	4

当你将唯一ID放入A列后，回到原始纸张上也写下唯一ID。

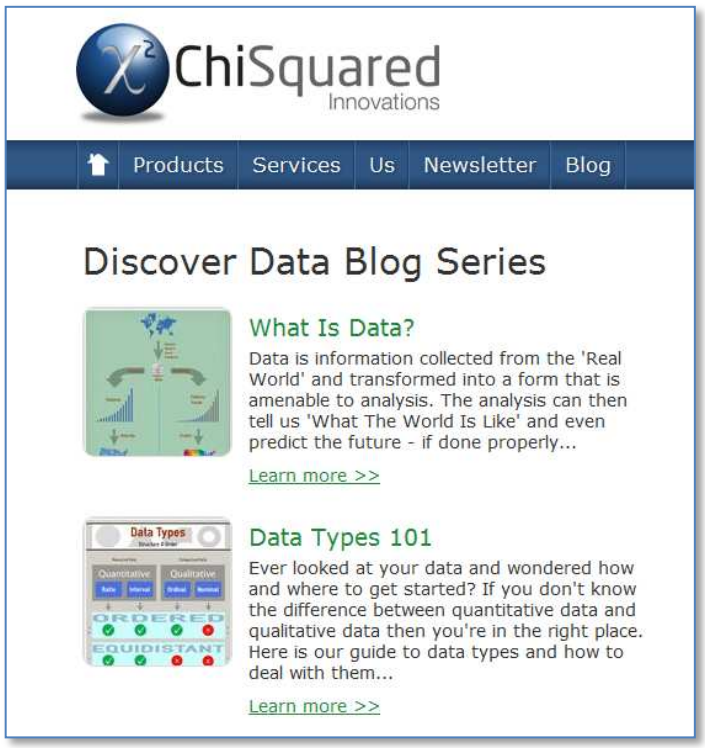
相信我，以后你会感谢我的这个建议...

技巧 #5

每个变量占据一列

每个变量都应该有...哦，等一下，什么是变量?

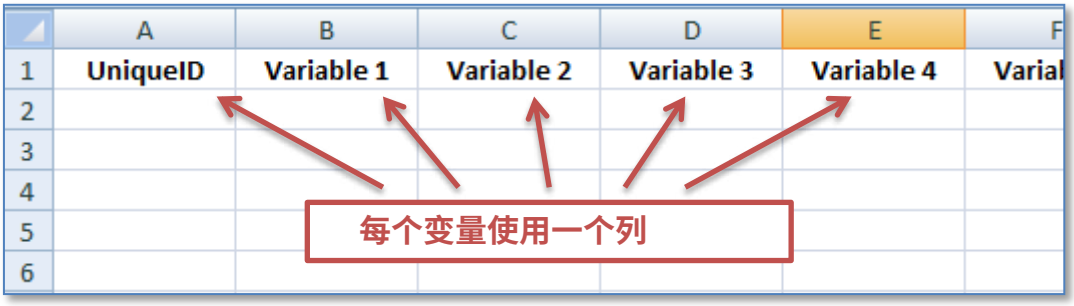
简单来说，这些是您研究中可能会改变或可以改变的事物。 简而言之，这些是您观察、测量、计数和收集的所有信息，如年龄、性别、距离、温度等。



您可以在我们的“发现数据博客系列”中找到更多关于数据、数据类型等的信息。

我们在哪里？啊对了...

每个变量都应该有自己的列，每个变量应该对应一条信息。



如果你要输入患者的年龄，只需输入他们的年龄，不要在同一列或单元格中输入他们的出生日期。

如果您想记录他们的年龄和出生日期，则使用两个单独的列。

如果您正在记录由2个或更多组成部分组成的复合变量，例如身体质量指数（由身高和体重组成），则应将它们记录在单独的列中。

您始终可以将它们合并为一个单一的变量。

技巧 #6

第1行是变量名称

最终，您将需要分析您的数据，并可能需要将其导出到统计程序中。

几乎所有商业统计程序的标准是第一行保留变量的名称，而其他所有行用于数据。

因此，请不要诱惑地将第2、3和4行以及第1行用于变量名称。

这可能会使Excel中的所有内容看起来整洁漂亮，但它只会给您带来更多的工作。

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

技巧 #7

每个单元格都应该有内容

空单元格告诉您什么？

- 等待更多信息？
- 数据未记录？
- 原始数据不正确？

一个空单元格只是一个巨大的问号，告诉你什么都没有。

更糟糕的是，不完整的数据集给审阅人员一个理由用一根比喻性的棍子敲打你的头部（相信我，他们会这样做的，我经历过很多次...）。

所以确保每个单元格都输入了一些内容。

使用“非法”数字作为代码来提供信息是相当常见的，所以对于一个变量的条目只能是正值（比如年龄或身高），我们可以使用诸如以下的代码：

	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		

如果负数没有用处，那么使用字母**a**、**b**、**c**等等。

如果您不习惯在严格不应该出现的单元格中输入内容（毕竟，在分析数据之前，您将不得不清理它们），那么可以使用Excel的批注功能。

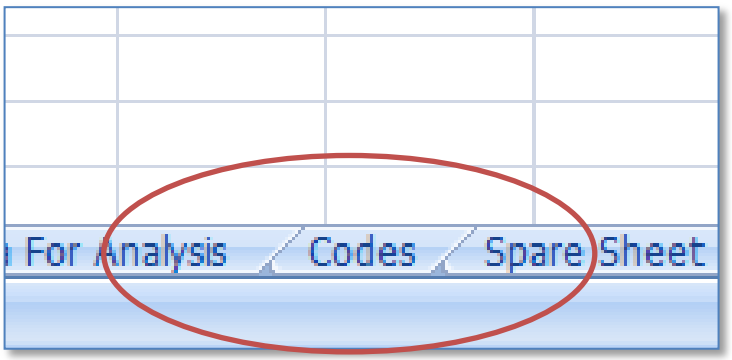
我倾向于节约使用，但这只是我的个人意见...

提示 #8

保持良好的笔记

当使用代码时，您需要保留笔记以告诉您代码的含义。

将代码和笔记保存在不同的电子表格中。



在这个话题上，重要的是：

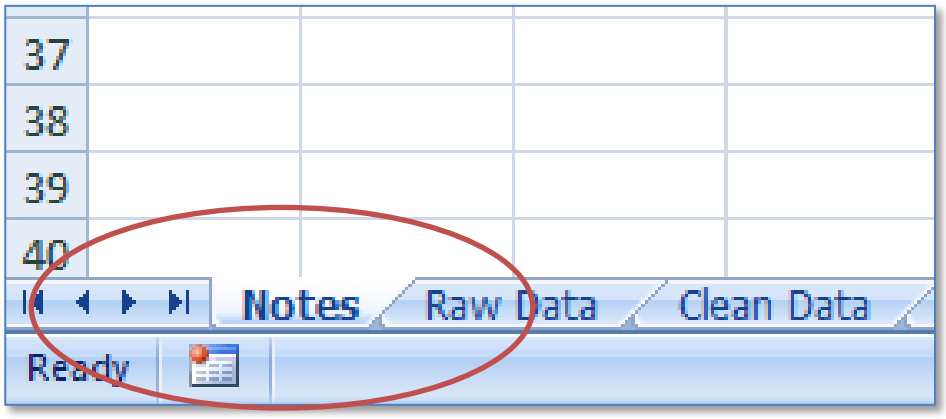
保持良好的笔记！！！

您可能不是唯一一个使用这个数据集的人，所以习惯于把东西写下来。

解释项目的全部内容，您试图回答的问题，为什么收集这些数据以及如何获得您寻找的答案。

解释您如何测量事物以及在什么条件下进行。

如果有多个人收集数据，请解释谁、什么、在哪里、何时、为什么以及如何进行。



这将是解释您的数据集的所有重要内容的文档，所以请写下来。

如果信息太多，无法舒适地放入Excel电子表格中，那么Microsoft Doc将非常适合 - 并将其保存在与数据集相同的文件夹中。

技巧 #9

保持一致

没有什么比得到一个需要花费两周时间清理的数据集更糟糕，因为数据输入不一致。

我的意思是，确保如果一个变量的输入应该是'正面'，那就输入'正面'，而不是其他变体：

B
Variable 1
Positive
Pos
POS
pos
positive
+ve
+

纠正拼写错误和打字错误已经很困难了，更不用说还要纠正故意输入不同的错误了。

限制能够输入数据的人数，以减少这些问题，并明确您的数据输入标准。

技巧 #10

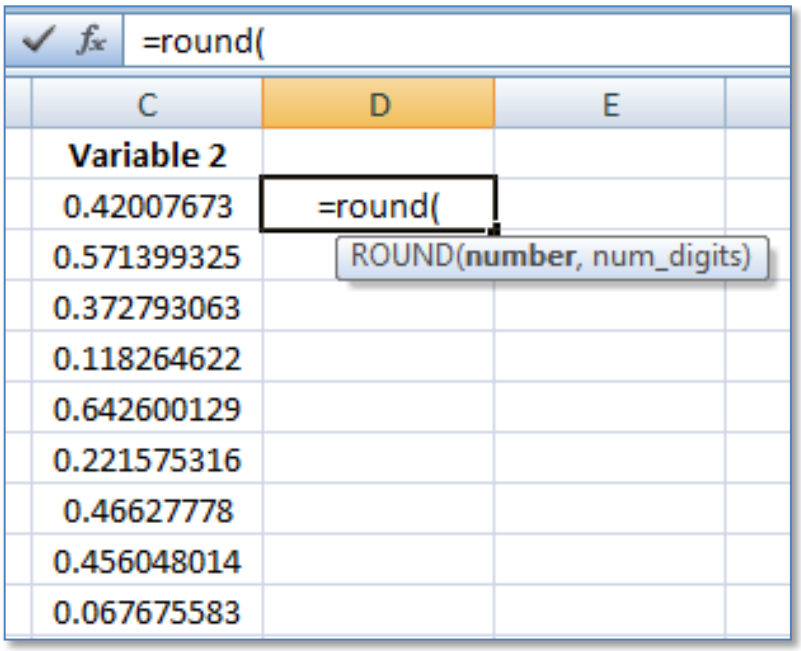
不要猜测

数据应尽可能准确地输入。

不要猜测、近似、四舍五入！！！

按照纸上的注册值输入数值。

使用Excel的函数来舍入数据，但不要在脑海中、纸上或计算器上进行计算，否则会出错，而这些错误可能很难，甚至不可能在后期发现。



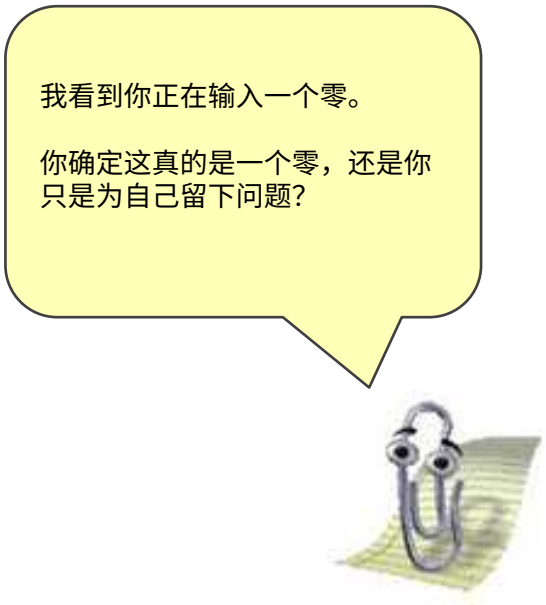
提示 #11

零是一个实数

除非测量、计数或计算的结果为零，否则不要在单元格中输入零。

我经常收到很多零的数据集，当我询问时，这些零意味着'我没有这方面的数据'。

问题是，如果你想计算一些东西，比如平均值，那么所有的零都将被用于计算，你将得到一个不准确的答案-或者一个完全错误的答案！



章节

2

数据清洗

如果你已经收集了所有的数据，并且非常小心，你可能会拥有一个完美的数据集。

干得好！

就个人而言，我从未见过一个完美的数据集-它是最稀有的生物。

很可能在开始分析之前，您需要清理您的数据。

再次提醒，教科书上很少提供实际建议，所以
让我们深入了解并制定一些基本规则，这将帮助您节省时间并让您的老板满意...

提示 #12

制作副本

您拥有一个“原始”数据集，基本上是您收集的所有纸质数据的电子副本。

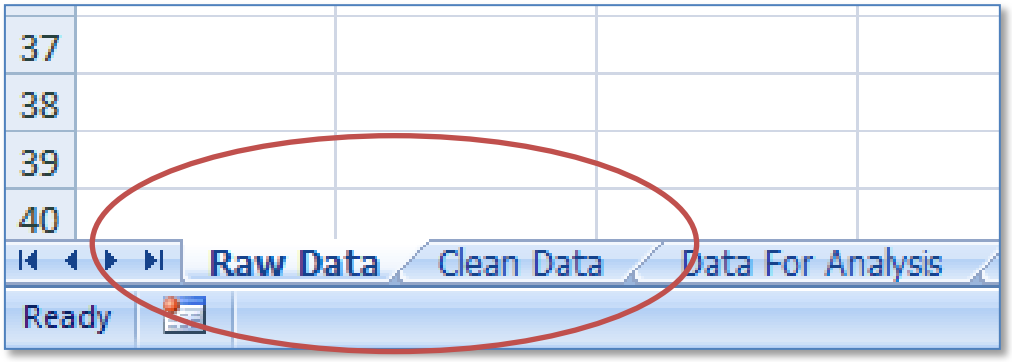
如果您在电子副本中出现输入错误，您可以随时查看原始纸质副本。

当您开始进行数据清洗时，您将更改**数据，并且您需要能够撤消任何清洗错误**您可能犯的错误，相信我 - 您会犯一些错误。

因此，创建一个数据集的副本工作表。

信不信由你，这是数据清洗中最重要的步骤之一。

将原始数据命名为“原始数据”，新数据命名为“清洗中进行中”，直到完成清洗，然后您可以将名称更改为“清洗数据”。



是的，还要确保两个工作表都有唯一标识列。

技巧 #13

在单独的工作表中清理您的数据

在清理单个数据列时，您将使用内置于Excel中的各种不同工具，如“查找和替换”功能。

当您使用“查找和替换”功能时，它只会在所选列上操作还是整个工作表上操作?

您确定吗?

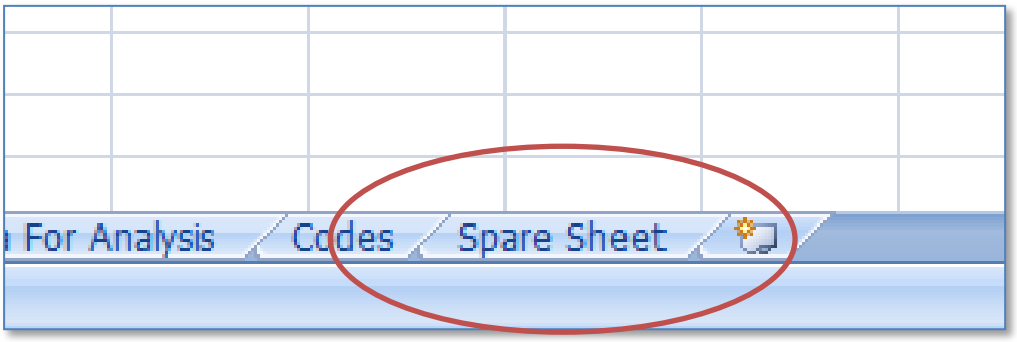
真的，真的确定吗?

所有内置函数的工作方式都相同吗?

如果回答错误，您会发现自己在整个数据集中引入了错误，而没有简单的撤消方法（在这里点击“撤消”无效）。

因此，当您想要清理单个数据列时，将该列复制到一个空白工作表中，并在那里进行操作。

完成后，您可以将清理后的列复制回去，替换之前的未清理列。



这可能需要您更多的时间，但是这是值得的-错误可能非常昂贵，并且在问题出现之前就解决它们是明智的。

哦，我真讨厌那个陈词滥调...

技巧 #14

向原始来源报告错误

如果同样的数据需要一次又一次地以完全相同的方式进行清理，那么清理数据就没有意义。

如果您正在使用共享数据集，比如部门数据库，请确保向原始来源报告任何错误。

这样，下次您需要从同一来源分析更多数据时，需要进行的清理工作就会少得多。

	A
1	My Variable 1
2	Correct
3	Correct
4	Correct
5	Correct
6	Corrrrecttt
7	Correct
8	Correct
9	Correct
10	Correct
11	Correct
12	Correct
13	Correct

技巧 #15

使用*Excel*函数来完成繁重的工作...

尽可能避免手动清理数据。

拼写错误、打字错误和不正确输入的最大来源之一就是手动输入，那么为什么要使用导致问题的同样方法呢?

Excel**有很多函数可以帮助数据清理，所以要好好利用它们。**

如果您有一个基于文本的列，请使用Excel的“删除重复项”功能。

结果将是该列中所有项目的列表。

然后，您可以使用“查找和替换”来纠正拼写错误的条目，包括纠正大小写错误的条目，例如“case”，“Case”或“CASE”。

提示＃16

...并使用*Excel*公式来完成更困难的工作

我无法告诉你我失去了多少周的生活 - 我再也找不回来了 - 试图找到错误的来源，结果发现是单元格中开头或结尾的空格。

你看不见它，但它仍然存在，当你开始进行分析时，它可能会造成严重破坏。

Excel忽略空格，因此它们非常难以检测，但其他分析和统计软件不会忽略它们，并将该条目视为不同的内容。

空格是我生活中的梦魇！！！

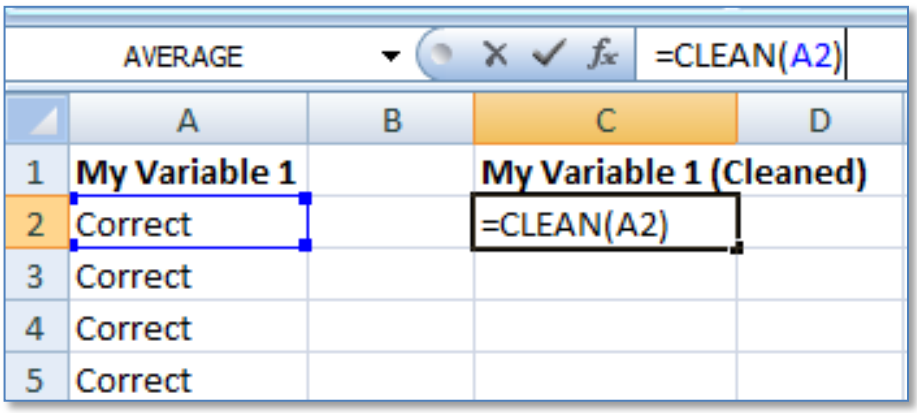
那么该怎么办呢?

Excel有几个不同的公式可以用于检测和修剪空格和其他不需要的字符，例如：

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

所以学习如何在Excel中进行简单编码并使用这些和其他公式。

我保证-这绝对是值得花时间的！



章节

3

数据编码和分类

所以，现在您拥有了一个完全干净的数据集，但在开始分析之前，您仍然需要做一些工作。

重要的是，您要注意您的代码意味着什么-毕竟，它们不是秘密，对吧？

假设您已经将变量的数据输入为1、2或3。

那是什么意思？

- 小、中或大？
- 猪、羊或山羊？

这很重要，因为您不应该期望记住您以何种方式、为何以及为何编码您的数据的所有细节。

技巧 #17

保留一个代码表

将您的代码保存在单独的工作表中，并将其命名为“代码”。 对于每一列，请记下您使用的代码以及它们的真实含义。

如果您使用了‘非法’条目，例如负数或字母，额外的代码，请记下它们的含义。

当您离开数据集几周后再回来时，您会很高兴您像这样组织好了数据。

这也会让您的老板、同事和友好的统计学家感到高兴，这从来不是一件坏事...

	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

技巧 #18

识别您的数据类型

当您进入分析阶段时，您需要了解数据类型-比率、区间、序数和名义-因此请花点时间决定哪些变量适用于每个类型，并在代码表中记录下来。

查看我们的发现数据博客系列获取更多信息...

当您有一个具有超过2个类别的变量时，请检查数据是否存在某种顺序或进展（序数），例如‘小’、‘中’或‘大’。

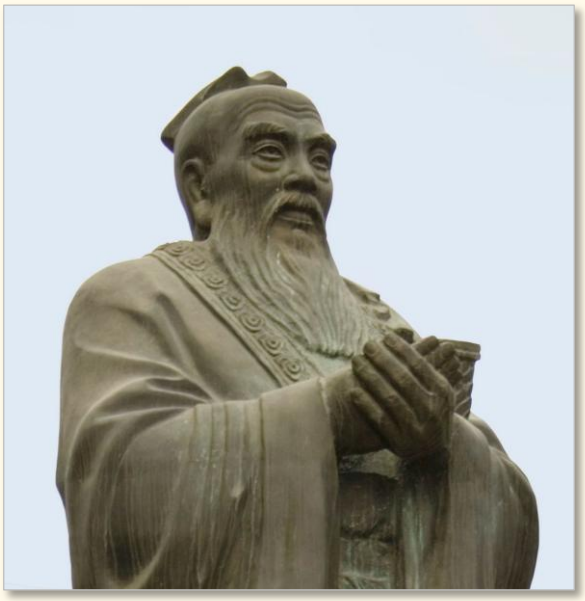
如果类别没有顺序但是具有描述性（名义性），比如'猪'、'羊'或者'山羊'，您需要为每个类别创建一个新的变量，就像这样：

	A	B	C	D	E
1	**Animal**		**Pig**	**Sheep**	**Goat**
2	Goat		No	No	Yes
3	Pig		Yes	No	No
4	Goat		No	No	Yes
5			-1	-1	-1
6	Pig		Yes	No	No
7	Goat		No	No	Yes
8	Pig		Yes	No	No
9	Pig		Yes	No	No
10	Sheep		No	Yes	No
11	Sheep		No	Yes	No

章节

4

数据完整性



一个犯了错误
却不纠正的人，就
是另外犯了一个错
误

孔子

仅仅因为您拥有一个完全干净、分类、编码和组织良好的数据集，并不意味着数据是正确的。

现实生活遵循规则，您的数据也必须如此！！！

我曾经发现我们医院正在治疗的是世界上最年长的人。

他已经超过300岁了，显然他度过了'美好的一生'。

在我分析的数据集中，他的出生日期（在18世纪某处）和住院日期（21世纪）之间的差距意味着他确实非常老了。

或者他的出生日期可能不太准确...

他的出生日期的错误在Excel的标准错误检查中无法检测到，因为它是一个完全合法的日期。

技巧 #19

检查您的数据是否合理

有时，将2个或多个数据组合在一起可以揭示出难以发现的错误，因此对每个变量进行一些简单的计算以检查数据是否符合合理的规则是明智的，例如：

- 计算最小值、最大值和平均值
- 为每个变量和每个类别保留计数
- 检查日期之间的差异

进行这些检查（在一个单独的工作表中！）可以帮助您找到异常值，例如年龄为负数或几百岁的人，并让您对数据有一个良好的了解。

对答案感觉不对劲?
那就再深入研究一下吧。

真的没有什么能代替亲自动手的经验!

	A	B	C	D
1		Gender	Age (y)	Height (m)
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

章节

5

工作更聪明，而不是更辛苦

额外提示

自动化您的数据清洗

即使您已经遵循了这里的所有提示，清理数据集仍然需要花费几天或几周的时间，如果数据集很小的话。
清理大型数据集可能需要几个月甚至更长时间。

如果您可以在几分钟而不是几周或几个月内自动清理数据，那将是非常棒的，不是吗？

我们认为是的，这正是我们所做的。

我们创建了一个完全自动化的数据清洗工具，它具有以下特点：

- ✓ **快速**
- ✓ **简单**
- ✓ **准确**

更好的是，它是智能的，因此它清理的数据越多，速度和准确性就越高。

而且您甚至可能可以免费使用它

- ✓ **节省时间和金钱**
- ✓ **消除压力**
- ✓ **更早完成您的研究**

所以，请查看我们的最新产品，然后与我们交谈。

我们很乐意听取您的意见！！！

下一步

订阅

嗯，我希望你喜欢这本电子书。

为什么不通过订阅我们的免费
新闻简报来学到更多呢：

小数点-CSI的嗡嗡声

你永远不知道，这可能不是你今天做的最糟糕的事情...

发现更多！！！

我们永远不会分享您的数据-永远！

版权

实用

数据清洗

李
贝克

李贝克是一位屡获殊荣的软件创作者，热衷于将数据转化为故事。
作为一名自豪的约克郡人，他现在居住在苏格兰东海岸的闪耀海岸线上。

作为一名物理学家、统计学家和程序员，他是60年代花力量迷幻时代的孩子，令人惊讶的是他变得如此正常！

他放弃了有前途的学术事业，选择做一些更有满足感的事情。作为Chi-Squared Innovations的首席执行官和联合创始人，他现在工作时间翻倍，薪水减半，压力增加十倍，但乐趣增加了100倍！

ChiSquared Innovations