

ERIN WALDRON FROM DATA DOZEN • FEB 25, 2021 FOR THE COWEN INSTITUTE

DATA PREPARATION

HOW TO STRUCTURE YOUR DATA SET FOR VISUALIZATION TOOLS



LISA GITELMAN

Raw data is an oxymoron.



DATA PREPARATION

*The most important skill with the **least** amount of training.*



OUR UNSPOKEN UNDERLYING ASSUMPTION

*The data is **worth** cleaning.*



OUR UNSPOKEN UNDERLYING ASSUMPTION

*There's a difference between messy and **useless** data.*



OUR UNSPOKEN UNDERLYING ASSUMPTION

To be worth cleaning, a data set needs to have a logical:

METHODOLOGY

RELIABLE AND
CONSISTENT



QUANTITY

ENOUGH DATA TO BE
MEANINGFUL



RELEVANCY

APPLICABLE TO OUR
RESEARCH QUESTION



MESSY DATA

There are two ways in which data can be messy:

CONTENT

WHAT'S WRITTEN
INSIDE THE CELLS



STRUCTURE

HOW THOSE CELLS
ARE LAID OUT



MESSY DATA

There are two ways in which data is messy:

CONTENT

WHAT'S WRITTEN
INSIDE THE CELLS



STRUCTURE

HOW THOSE CELLS
ARE LAID OUT



OUR GOAL

We want to be able to compare and aggregate.

CONTENT

WHAT'S WRITTEN
INSIDE THE CELLS



- ▶ We're inconsistent.
St. Louis, Saint Louis, STL
- ▶ We make errors.
Misspellings, wrong field, added an extra 0
- ▶ We don't have all the data.
Partial or empty cells
- ▶ The world is ambiguous.
World Bank vs United Nations country list
- ▶ And things change.
And this makes consistency hard.

MESSY DATA

There are two ways in which data is messy:

CONTENT

WHAT'S WRITTEN
INSIDE THE CELLS



STRUCTURE

HOW THOSE CELLS
ARE LAID OUT



MESSY DATA

We want the computer to be able to see what we see.

“TIDY DATA” BY HADLEY WICKHAM

- ▶ One column per variable.
Fields, measures, dimensions, metrics
- ▶ One row per transaction.
Records, events, items
- ▶ One table per granularity.
Long, skinny, repetitive data

STRUCTURE

HOW THOSE CELLS
ARE LAID OUT



TIDY DATA

*“Happy families are all alike; every unhappy family is **unhappy** in its own way” -TOLSTOY*



TIDY DATA

*“Like families, tidy data sets are all alike, but every messy data set is **messy** in its own way.” -HADLEY WICKHAM*





OUR GOAL IS TO CREATE SHORTCUTS

PROBLEM

The 10 most common data structure problems analysts encounter before we can start analyzing.

TELL TALE SIGN

The symptoms such data structures have, and how we try to troubleshoot or compensate on our own.

SOLUTION

How we can restructure the data for analysis (or so that the computer can see what we see on the spreadsheet.)



1) DATA IN THE HEADERS

PROBLEM

There is actually pieces of the data in the column headers instead of variable names in the column headers.

TELL TALE SIGN

The number of columns doesn't equal the number of variables in the data set. Analysts start creating new columns by manually copy and pasting in Excel.

SOLUTION

Pivot the columns with data in the header. To start, simply identify and list the actual variables in the data set. The number of columns will shrink to match the number of variables and the number of rows will grow.



1) DATA IN THE HEADERS

DEPARTMENT	2014	2015	2016	2017
ANTHROPOLOGY	#####	#####	#####	#####
ART HISTORY	#####	#####	#####	#####
CHEMISTRY	#####	#####	#####	#####



1) DATA IN THE HEADERS

DEPARTMENT	YEAR	STUDENTS ENROLLED
ANTHROPOLOGY	2014	#####
ANTHROPOLOGY	2015	#####
ANTHROPOLOGY	2016	#####
ANTHROPOLOGY	2017	#####
ART HISTORY	2014	#####
ART HISTORY	2015	#####



2) VARIABLE NAMES IN THE ROWS

PROBLEM

Analysts can see all of the variables and all of the individual events, transaction, etc. But the variables are stacked in the rows.

TELL TALE SIGN

None of the variables are available when the data set is loaded into Tableau.

SOLUTION

Rotate the entire table. The whole thing needs to go 90 degrees to the right. The rows need to become the columns and the columns need to become the rows.



2) VARIABLE NAMES IN THE ROWS

VARIABLES	STUDENT A	STUDENT B	STUDENT C	STUDENT D
GRADE POINT AVERAGE	#####	#####	#####	#####
TIME TO DEGREE	#####	#####	#####	#####
TOTAL CREDITS EARNED	#####	#####	#####	#####



2) VARIABLE NAMES IN THE ROWS

STUDENT ID	GRADE POINT AVERAGE	TIME TO DEGREE	TOTAL CREDITS EARNED
STUDENT A	#####	#####	#####
STUDENT B	#####	#####	#####
STUDENT C	#####	#####	#####



3) NEW RECORDS

PROBLEM

Analysts receive new records for a data set with which they've already worked. This might be a new year of data or a new department, but the columns (the data structure) are identical to previous tables.

TELL TALE SIGN

Analysts start replicating all of the work they've done before with previous data sets.

SOLUTION

Union the data. We'll stack the tables on top of each other. Then we'll add a new column that documents which table the record came from in case we need to be able to tell the two apart later.



3) NEW RECORDS

ANTHROPOLOGY

YEAR	APPLICANTS	ACCEPTED
2014	89	20
2015	95	21
2016	102	19

ART HISTORY

YEAR	APPLICANTS	ACCEPTED
2014	20	3
2015	25	4
2016	19	2



3) NEW RECORDS

DEPARTMENT	YEAR	APPLICANTS	ACCEPTED
Anthropology	2014	89	20
Anthropology	2015	95	21
Anthropology	2016	102	19
Art History	2014	20	3
Art History	2015	25	4
Art History	2016	19	2



4) NEW COLUMNS

PROBLEM

Different variables exist in different tables, and the analyst needs to use them at the same time.

TELL TALE SIGN

Analysts start to manually copy data from the secondary table into new columns in their primary table.

SOLUTION

Join the tables into a single data source leveraging a common variable as the crosswalk.



4) NEW COLUMNS

DEPARTMENT ENROLLMENT

DEPARTMENT	ENROLLMENT
1	####
2	####
3	####

DEPARTMENT DIRECTORY

DEPARTMENT	NAME	PHONE NUMBER
1	ANTHROPOLOGY	816-555-5555
2	ART HISTORY	816-555-5556
3	CHEMISTRY	816-555-5557



4) NEW COLUMNS

DEPARTMENT	ENROLLMENT	NAME	PHONE NUMBER	DEPARTMENT
1	####	ANTHROPOLOGY	816-555-5555	1
2	####	ART HISTORY	816-555-5556	2
3	####	CHEMISTRY	816-555-5557	3



5) MULTIPLE VARIABLES IN ONE COLUMN

PROBLEM

Multiple pieces of information have been combined, or concatenated, into a single column.

TELL TALE SIGN

Analysts can see the data they need right in front of them, but can't figure out how to leverage just one piece of it.

SOLUTION

Split the single columns into multiple columns, creating one column for each piece of data.



5) MULTIPLE VARIABLES IN ONE COLUMN

CLASS NUMBER	LOCATION
1	3248 NE Mohawk Ln 64116 (39.153838°, -94.557429°)
2	6815 Wilson Ave 64125 (39.107564°, -94.501859°)
3	10615 Blue Ridge Blvd 64134 (38.930798°, -94.507242°)



5) MULTIPLE VARIABLES IN ONE COLUMN

CLASS NUMBER	STREET ADDRESS	ZIP	LONGITUDE	LATITUDE
1	3248 NE Mohawk Ln	64116	39.153838°	-94.557429°
2	6815 Wilson Ave	64125	39.107564°	-94.501859°
3	10615 Blue Ridge Blvd	64134	38.930798°	-94.507242°



6) SAME VARIABLE IN TWO COLUMNS

PROBLEM

Single pieces of information have been broken apart into multiple columns.

TELL TALE SIGN

Analysts keep running into issues with date/time or location data.

SOLUTION

Concatenate, or combine, these disparate columns into a single column.



6) SAME VARIABLE IN TWO COLUMNS

APPLICATION ID	DATE	TIME
1	05/06/2019	09:56 AM
2	04/21/2019	10:13 PM
3	02/18/2019	07:20 AM



6) SAME VARIABLE IN TWO COLUMNS

APPLICATION ID	DATE/TIME
1	05/06/2019 09:56 AM
2	04/21/2019 10:13 PM
3	02/18/2019 07:20 AM



7) IMPLIED DATA

PROBLEM

There are empty cells or extra headers that human eyes easily connect to the data.

TELL TALE SIGN

All of this data is lost when the data set is connected to Tableau.

SOLUTION

Fill in all of the implied data. This will flatten the data set if there are any additional headers, sheet titles, color coding, etc., into a single table.



7) IMPLIED DATA

DEPARTMENT	STUDENT ID
Anthropology	1
	2
	3
Art History	4
	5
	6



7) IMPLIED DATA

DEPARTMENT	STUDENT ID
Anthropology	1
Anthropology	2
Anthropology	3
Art History	4
Art History	5
Art History	6



8) DIFFERENT GRANULARITIES

PROBLEM

Subtotals and grand totals have been included as rows along with the granular data.

TELL TALE SIGN

Analysts get totals for the entire data set that double or triple count the actual individual records.

SOLUTION

Remove the subtotals and grand totals. Add columns for any implied information that might be hidden in the subtotal labels.



8) DIFFERENT GRANULARITIES

DEPARTMENT	STUDENTS ENROLLED
ART HISTORY	#####
ENGLISH	#####
HUMANITIES DISCIPLINE TOTAL	#####
POLITICAL SCIENCE	#####
PSYCHOLOGY	#####
SOCIAL SCIENCES DISCIPLINE TOTAL	#####



8) DIFFERENT GRANULARITIES

DISCIPLINE	DEPARTMENT	STUDENTS ENROLLED
HUMANITIES	ART HISTORY	#####
HUMANITIES	ENGLISH	#####
SOCIAL SCIENCE	POLITICAL SCIENCE	#####
SOCIAL SCIENCE	PSYCHOLOGY	#####



9) OVERLY AGGREGATED DATA

PROBLEM

All of the data provided to the analyst has been aggregated. Often individual records contain multiple columns that can't be leveraged together.

TELL TALE SIGN

Analysts can only replicate the exact information that can be seen in the original table. The data feels very one dimensional.

SOLUTION

Create a data set with more granularity. This is usually closer to a direct export of the data instead an aggregated output.



9) OVERLY AGGREGATED DATA

PROGRAM	WHITE	AFRICAN AMERICAN	HISPANIC	ASIAN	NATIVE AMERICAN	MULTI-RACIAL / OTHER	FEMALE	MALE
A	%	%	%	%	%	%	%	%
B	%	%	%	%	%	%	%	%
C	%	%	%	%	%	%	%	%



9) OVERLY AGGREGATED DATA

PROGRAM	STUDENT ID	RACE	SEX
A	1	Native American	Female
A	2	White	Male
A	3	White	Female
A	4	African American	Femal



10) A VARIABLE OF VARIABLES

PROBLEM

The analyst can see all of the variables, but has trouble teasing them apart in their analysis.

TELL TALE SIGN

We're constantly using the filter card to narrow down to just one variable in Tableau.

SOLUTION

Cast these variables into their own individual columns. Now they will be able to be used independently.



10) A VARIABLE OF VARIABLES

PROGRAM	MEASURE	TIME TO DEGREE
Anthropology	Minimum	4 years
Anthropology	Maximum	7.3 years
Art History	Minimum	3.5 years
Art History	Maximum	9 years



10) A VARIABLE OF VARIABLES

PROGRAM	MIN TIME TO DEGREE	MAX TIME TO DEGREE
Anthropology	4 years	7.3 years
Art History	3.5 years	9 years

DATA DOZEN

a visual analytics studio

Copyright © 2021 Data Dozen, LLC All Rights Reserved