ERIN WALDRON FROM DATA DOZEN ◍ MAY 7, 2019 AT KC DAMA DAY

LONG AND SKINNY: STRUCTURING DATA FOR VISUALIZATION

Prepping the data is 90% of the work.

And many analysts don't know

how data is structured behind the scenes.

THE DATA STRUCTURE I HAD

STUDENT ID	2014 FALL FUNDING TYPE	2014 FALL FUNDING AMOUNT	2015 SPRING FUNDING TYPE	2015 SPRING FUNDING AMOUNT
1	TEACHING ASSISTANTSHIP	$10,000	TEACHING ASSISTANTSHIP	$10,000
2	TEACHING ASSISTANTSHIP	$10,000	RESEARCH ASSISTANTSHIP	$10,500
3	UNIVERSITY FELLOWSHIP	$10,000	UNIVERSITY FELLOWSHIP	$10,000

THE DATA STRUCTURE I NEEDED
(BUT DIDN'T KNOW HOW TO ASK FOR)

STUDENT ID	SEMESTER	FUNDING TYPE	FUNDING AMOUNT
1	201405	TEACHING ASSISTANTSHIP	$10,000
1	201502	TEACHING ASSISTANTSHIP	$10,000
2	201405	TEACHING ASSISTANTSHIP	$10,000
2	201502	RESEARCH ASSISTANTSHIP	$10,500
3	201405	UNIVERSITY FELLOWSHIP	$10,000
3	201502	UNIVERSITY FELLOWSHIP	$10,000

This is exactly what we had requested.

OUR GOAL IS TO CREATE SHORTCUTS

PROBLEM	TELL TALE SIGN	SOLUTION
The 10 most common data structure problems analysts encounter before they can start analyzing.	The symptoms such data structures have, and how analysts attempt to troubleshoot on their own.	How we can restructure the data for analysis (even if the analysts can't articulate what exactly they need.)

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 data. 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.

BEFORE PIVOT

1) DATA IN THE HEADERS

STUDENT ID	2014 FALL FUNDING TYPE	2014 FALL FUNDING AMOUNT	2015 SPRING FUNDING TYPE	2015 SPRING FUNDING AMOUNT
1	TEACHING ASSISTANTSHIP	$10,000	TEACHING ASSISTANTSHIP	$10,000
2	TEACHING ASSISTANTSHIP	$10,000	RESEARCH ASSISTANTSHIP	$10,500
3	UNIVERSITY FELLOWSHIP	$10,000	UNIVERSITY FELLOWSHIP	$10,000

AFTER PIVOT

1) DATA IN THE HEADERS

STUDENT ID	SEMESTER	FUNDING TYPE	FUNDING AMOUNT
1	201405	TEACHING ASSISTANTSHIP	$10,000
1	201502	TEACHING ASSISTANTSHIP	$10,000
2	201405	TEACHING ASSISTANTSHIP	$10,000
2	201502	RESEARCH ASSISTANTSHIP	$10,500
3	201405	UNIVERSITY FELLOWSHIP	$10,000
3	201502	UNIVERSITY FELLOWSHIP	$10,000

BEFORE PIVOT

1) DATA IN THE HEADERS

CITY	2014	2015	2016	2017
KANSAS CITY	#####	#####	#####	#####
ST. LOUIS	#####	#####	#####	#####
DENVER	#####	#####	#####	#####

AFTER PIVOT

1) DATA IN THE HEADERS

CITY	YEAR	POPULATION
KANSAS CITY	2014	#####
KANSAS CITY	2015	#####
KANSAS CITY	2016	#####
KANSAS CITY	2017	#####
ST. LOUIS	2014	#####
ST. LOUIS	2015	#####

2) 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.

BEFORE UNION

2) 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

AFTER UNION

2) 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

3) CONCATENATED VARIABLES

PROBLEM

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

TELL TALE SIGN

Analysts can see the data the 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.

BEFORE SPLIT

3) CONCATENATED VARIABLES

SERVICE REQUEST ID	LOCATION	REQUEST TYPE
1	3248 NE Mohawk Ln 64116 (39.153838°, -94.557429°)	Water-No Water / Pressure-Low Pressure
2	6815 Wilson Ave 64125 (39.107564°, -94.501859°)	Water-Leak-At Curb / In Yard
3	10615 Blue Ridge Blvd 64134 (38.930798°, -94.507242°)	Streets / Roadways / Alleys-Pothole-District 2

AFTER SPLIT

3) CONCATENATED VARIABLES

SERVICE REQUEST ID	STREET ADDRESS	ZIP	LONGITUDE	LATITUDE	CATEGORY	TYPE	DETAIL
1	3248 NE Mohawk Ln	64116	39.153838°	-94.557429°	Water	No Water / Pressure	Low Pressure
2	6815 Wilson Ave	64125	39.107564°	-94.501859°	Water	Leak	At Curb / In Yard
3	10615 Blue Ridge Blvd	64134	38.930798°	-94.507242°	Streets / Roadways / Alleys	Pothole	District 2

4) SPLIT VARIABLES

PROBLEM	TELL TALE SIGN	SOLUTION
Single pieces of information have been broken apart into multiple columns.	Analysts keep running into issues with date/time or location data.	Concatenate, or combine, these disparate columns into a single column.

4) SPLIT VARIABLES

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

AFTER CONCATENATING

4) SPLIT VARIABLES

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

5) DATA IN MULTIPLE TABLES

PROBLEM	TELL TALE SIGN	SOLUTION
Different variables exist in different tables, and the analyst needs to use them at the same time.	Analysts start to manually copy data from the secondary table into new columns in their primary table.	Join the tables into a single data source leveraging a common variable as the crosswalk.

BEFORE JOIN

5) DATA IN MULTIPLE TABLES

SERVICE REQUESTS

SERVICE REQUEST ID	CATEGORY	DEPARTMENT
1	Water	Water Services
2	Water	Water Services
3	Streets / Roadways / Alleys	Public Works

DEPARTMENT DIRECTORY

DEPARTMENT	WEBSITE	PHONE NUMBER
Water Services	https://www.kcwater.us/	816-513-1313
Public Works	http://kcmo.gov/publicworks/	816-513-9970

AFTER JOIN

5) DATA IN MULTIPLE TABLES

SERVICE REQUEST ID	CATEGORY	DEPARTMENT	WEBSITE	PHONE NUMBER
1	Water	Water Services	https://www.kcwater.us/	816-513-1313
2	Water	Water Services	https://www.kcwater.us/	816-513-1313
3	Streets / Roadways / Alleys	Public Works	http://kcmo.gov/publicworks/	816-513-9970

6) SUBTOTALS

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.

BEFORE REMOVING

6) SUBTOTALS

NEIGHBORHOOD	POPULATION
NEIGHBORHOOD 1	#####
NEIGHBORHOOD 2	#####
JACKSON COUNTY TOTAL	#####
NEIGHBORHOOD 3	#####
NEIGHBORHOOD 4	#####
CLAY COUNTY TOTAL	#####

AFTER REMOVING

6) SUBTOTALS

COUNTY	NEIGHBORHOOD	POPULATION
JACKSON COUNTY	NEIGHBORHOOD 1	#####
JACKSON COUNTY	NEIGHBORHOOD 2	#####
CLAY COUNTY	NEIGHBORHOOD 3	#####
CLAY COUNTY	NEIGHBORHOOD 4	#####

7) 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.

7) OVERLY AGGREGATED DATA

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

AFTER DISAGGREGATING

7) OVERLY AGGREGATED DATA

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

8) UNDEFINED FIELDS

PROBLEM	TELL TALE SIGN	SOLUTION
The analyst can has variables in their dataset they can't explain.	Analysts are scared to use the data they have.	Add context. This might involve including metadata, documentation, or simply adding more variables which help explain the ones already included.

8) UNDEFINED FIELDS

SOSEM	GRADUATION RATE
201405	95%
201405	92%
201402	86%
201503	99%

AFTER CONTEXT

8) UNDEFINED FIELDS

SORT SEMESTER	WRITTEN SEMESTER	GRADUATION RATE	TOTAL APPLICATIONS FOR GRADUATION	TOTAL DEGREES CONFERRED
201405	Fall 2014	95%	102	97
201405	Fall 2014	92%	111	102
201402	Spring 2014	86%	143	123
201503	Summer 2015	99%	98	97

9) SURVEY DATA

PROBLEM

Participants have completed a survey and the analysts are unable to start working with the data effectively.

TELL TALE SIGN

There's just one table containing survey responses.

SOLUTION

Create three tables: one with the survey responses, one with the participant's demographic information, and one with the metadata from the survey instrument itself.

9) SURVEY DATA

ID	Q1	Q2	Q3	Q4
1	Yes	Very Likely	4	I wish that...
2	No	Not At All Likely	5	
3	Yes	Extremely Likely	2	The main problem is...

9) SURVEY DATA

RESPONSE TABLE

ID	QUESTION ID	RESPONSE
1	Q1	Yes
1	Q2	Very Likely
1	Q3	4

PARTICIPANT TABLE

ID	AGE	SEX
1	28	Male
2	48	Female
3	49	Male

SURVEY INSTRUMENT TABLE

QUESTION ID	CATEGORY	LONG FORM
Q1	A	Are you part of...
Q2	A	How likely...
Q3	A	On a scale

10) UNREPEATABLE DATA

PROBLEM

Team members are making assumptions and changes to their work as they go that are not documented.

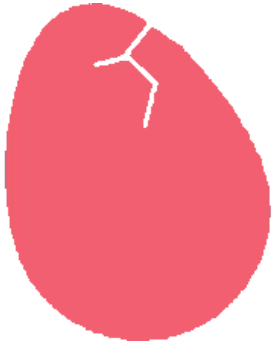
TELL TALE SIGN

Folks can't recreate their work. If someone asks how they arrived at a certain number or data point, analysts can't provide an answer.

SOLUTION

Communication! But that's a whole session in its own right. We'll wrap up today with The Communication Checklist.

Let's build.



DATA DOZEN

DATA VISUALIZATION EDUCATION