Open in app ↗

Search Medium

✦ Member-only story

DATA COLLECTION

Four Principles of Data Collection

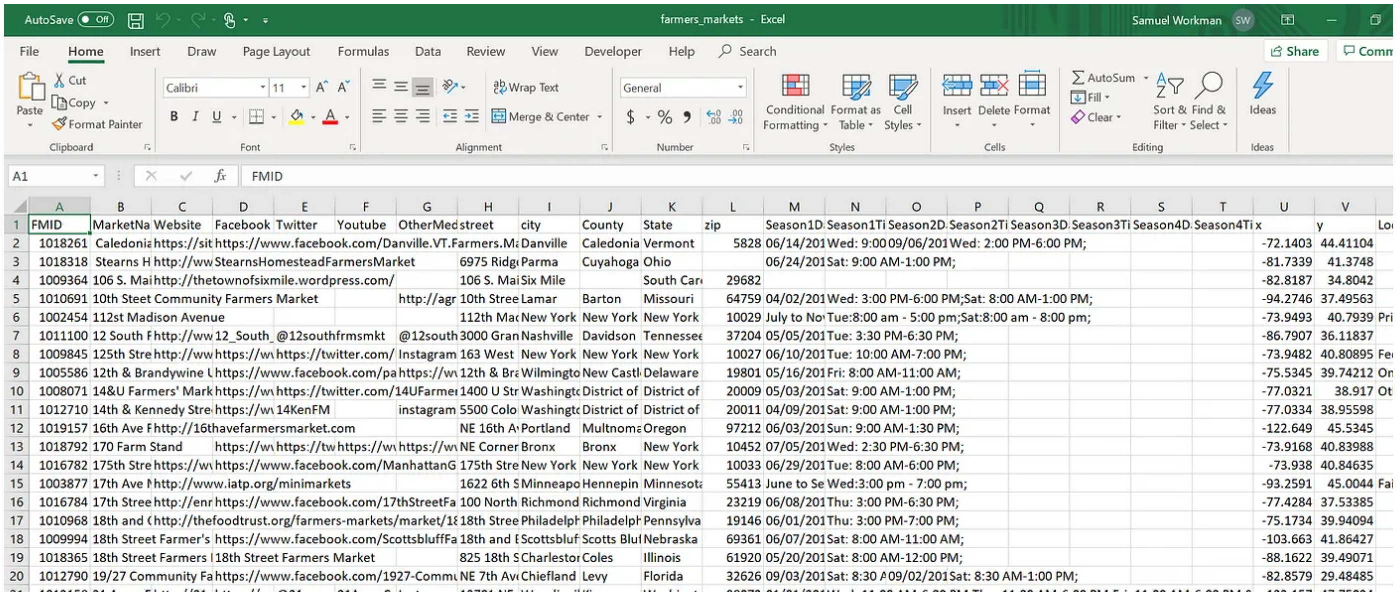
Simplify and streamline analysis by doing less to the data

Samuel Workman · Follow

Published in Towards Data Science

8 min read · May 12, 2020

Listen Share More



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	FMID	MarketNa	Website	Facebook	Twitter	Youtube	OtherMed	street	city	County	State	zip	Season1D	Season1Ti	Season2D	Season2Ti	Season3D	Season3Ti	Season4D	Season4Ti	x	y	Lo
2	1018261	Caledonia	https://sit	https://www.facebook.com/Danville.VT.Farmers.Ma					Danville	Caledonia	Vermont	5828	06/14/201	Wed: 9:00	09/06/201	Wed: 2:00 PM-6:00 PM;					-72.1403	44.41104	
3	1018318	Stearns H	http://ww	StearnsHomesteadFarmersMarket				6975 Ridg	Parma	Cuyahoga	Ohio		06/24/201	Sat: 9:00 AM-1:00 PM;							-81.7339	41.3748	
4	1009364	106 S. Mai	http://thetownofsixmile.wordpress.com/					106 S. Mai	Six Mile		South Car	29682									-82.8187	34.8042	
5	1010691	10th Steet Community Farmers Market					http://agr	10th Stree	Lamar	Barton	Missouri	64759	04/02/201	Wed: 3:00 PM-6:00 PM;Sat: 8:00 AM-1:00 PM;							-94.2746	37.49563	
6	1002454	112st Madison Avenue						112th Mac	New York	New York	New York	10029	July to No	Tue:8:00 am - 5:00 pm;Sat:8:00 am - 8:00 pm;							-73.9493	40.7939	Pri
7	1011100	12 South F	http://ww	12_South_	@12southfrmsmkt		@12south	3000 Gran	Nashville	Davidson	Tennessee	37204	05/05/201	Tue: 3:30 PM-6:30 PM;							-86.7907	36.11837	
8	1009845	125th Stre	http://ww	https://w	https://twitter.com/		Instagram	163 West	New York	New York	New York	10027	06/10/201	Tue: 10:00 AM-7:00 PM;							-73.9482	40.80895	Fe
9	1005586	12th & Brandywine U		https://www.facebook.com/pa			https://w	12th & Bra	Wilmingto	New Castl	Delaware	19801	05/16/201	Fri: 8:00 AM-11:00 AM;							-75.5345	39.74212	On
10	1008071	14&U Farmers' Mark		https://w	https://twitter.com/14UFarme			1400 U Str	Washingt	District of	District of	20009	05/03/201	Sat: 9:00 AM-1:00 PM;							-77.0321	38.917	Ot
11	1012710	14th & Kennedy Stre		https://w	14KenFM		instagram	5500 Colo	Washingt	District of	District of	20011	04/09/201	Sat: 9:00 AM-1:00 PM;							-77.0334	38.95598	
12	1019157	16th Ave F	http://16thavefarmersmarket.com					NE 16th A	Portland	Multnoma	Oregon	97212	06/03/201	Sun: 9:00 AM-1:30 PM;							-122.649	45.5345	
13	1018792	170 Farm Stand		https://w	https://tw	https://w	https://w	NE Corner	Bronx	Bronx	New York	10452	07/05/201	Wed: 2:30 PM-6:30 PM;							-73.9168	40.83988	
14	1016782	175th Stre	https://w	https://www.facebook.com/ManhattanG				175th Stre	New York	New York	New York	10033	06/29/201	Tue: 8:00 AM-6:00 PM;							-73.938	40.84635	
15	1003877	17th Ave N	http://www.iatp.org/minimarkets					1622 6th S	Minneapo	Hennepin	Minnesota	55413	June to Se	Wed:3:00 pm - 7:00 pm;							-93.2591	45.0044	Fai
16	1016784	17th Stree	http://enr	https://www.facebook.com/17thStreetFa				100 North	Richmond	Richmond	Virginia	23219	06/08/201	Thu: 3:00 PM-6:30 PM;							-77.4284	37.53385	
17	1010968	18th and C	http://thefoodtrust.org/farmers-markets/market/18					18th Stree	Philadelph	Philadelph	Pennsylva	19146	06/01/201	Thu: 3:00 PM-7:00 PM;							-75.1734	39.94094	
18	1009994	18th Street Farmer's		https://www.facebook.com/ScottsbluffFa				18th and E	Scottsbluf	Scotts Blu	Nebraska	69361	06/07/201	Sat: 8:00 AM-11:00 AM;							-103.663	41.86427	
19	1018365	18th Street Farmers M		18th Street Farmers Market				825 18th S	Charleston	Coles	Illinois	61920	05/20/201	Sat: 8:00 AM-12:00 PM;							-88.1622	39.49071	
20	1012790	19/27 Community Fa		https://www.facebook.com/1927-Commu				NE 7th Av	Chiefland	Levy	Florida	32626	09/03/201	Sat: 8:30 A	09/02/201	Sat: 8:30 AM-1:00 PM;					-82.8579	29.48485	

Source: Agricultural Marketing Service Local Food Directory

I've spent an academic lifetime compiling original data sets, both in my work and as part of broader research teams. In doing so, I have become quite opinionated about the role of data in science, its collection, and its organization. In this post, I detail four principles of data collection and organization useful to any analyst needing to collect data, whether by hand or machine. Often, students, junior scholars, business analysts, and public managers engage in data collection. Yet, it receives scant attention in most curricula. Despite the rise of data science, with its tools for scraping, parsing, and automating cleaning, organization, and learning processes, many of us still need to collect our data.

I study elite texts and policy documents quantitatively. My projects address the bureaucracy, policy implementation, and regulatory policy. My book uses data from almost a quarter-million regulatory proposals, and I am currently working on a National Science Foundation (NSF) funded project that examines a 100k public comments on regulations by paragraph.

I also teach undergraduate and graduate-level statistics at a large public university. I am often struck by students who understand the conceptual logic of statistics and how to build regression models but have little intuition for how to collect and organize data. Even as data science and statistics begin to exert more influence in everyday management, I find this intuition even more lacking in the private sector. The following set of fundamental principles serve as a guide to data collection. Fidelity to them will improve the quality of data and the impact it can have on decision-making — whether for research or public and private sector management.

Be Faithful to the Data-Generating Process

We most often discuss or characterize science, or learning, as a process revolving around theory and hypothesis-testing. This depiction is only partly accurate. At its bud, science is structured observation. Without precise recording and preservation of observations, developing and testing theory is not possible. Moreover, science is not reproducible.

Data are the result of social, biological, or physical processes. The world doesn't give us half-persons, percentages, or log-dollars; it gives us people, counts, and dollars. Fidelity to the data-generating process allows us to improve calculation and measurement in the process of iterative decision-making and the transitions between cycles of inductive and deductive reasoning.

Principle 1: Record data; don't calculate or transform it.

In the past, the data limitations of available spreadsheets or databases (e.g., MS Excel or MS Access) limited the number of characters one could store in a cell or the number of rows a given data set could contain. These limitations meant keeping a record of the original form of the data was challenging, and required lots of files and space. Thankfully, this is no longer the case for most spreadsheet applications. (SQL or SQLite databases quickly deal with remaining limitations.) Given these advancements, store data as-is during collection.

1A: If possible, store data as text or in text compatible format.

Data collection is not the only danger present in obtaining data faithful to the data-generating process. Programs, especially 'tooled-up' spreadsheets, often add hidden structure or content to data. For example, leading or trailing whitespace is particularly annoying, because the problem is not readily visible. Other issues involve storing text as numbers or vice versa and reading all character strings as factor variables. For this reason, it's advisable to store data you collect as text or text compatible files (e.g., .txt or .csv 'delimited' files).

1B: Back up data.

There is no need for a lot of words here. Store your data in multiple locations. And, if essential, one of those should be a physical hard-drive — all clouds have glitches from time to time. And, if following these principles, you should only need to store the original, text-based file.

Principle 2: Curate Data Organization

Once data are collected, the first thing to do is to wrap your mind around its structure. I use the word "curate" here because it implies a careful (re)organization of the data — something done with thought. Data is easily transitioned between "wide" or "long" format, so long as its structure and organization are consistent. (The difference between wide and long format data could be a separate post.)

2A: Observations appear in rows; variables appear in columns; values for observations on variables appear in the matrix of cells between them.

The essential concern of data organization is the distinction between observations and variables or indicators. Observations, subjects, or cases always appear in rows; variables appear in columns. (The first column contains the row [i.e., observation] labels.)

2B: Nesting structure should appear in columns, not rows.

Observations may contain groupings. Examples include time, geographic location, repeated tests, religions, etc. In business, common clusters are divisions, categories, or stores. These types of groupings should not be displayed in rows (with an indentation or any other means), but instead should appear as separate variables in columns. In a national survey, for example, respondents' state of residence may be recorded. 'State' should appear in a column somewhere to the right of the row labels in a column of its own — with column label "state."

Business calculations provide an excellent example of this violation. In my forays into business consulting, I often see periods displayed chronologically in columns. It makes sense when you are calculating singular values from these columns, but doing anything else with the data is nigh-on impossible, absent considerable reorganization, especially if rows contain nesting. Periods are grouping variables and should appear in their column.

2C: Beware complicated row, column, or value labels.

Row, column, or value labels with case sensitive characters, special characters, or whitespace cause problems in analytical software beyond the spreadsheet (they can be a problem within the spreadsheet as well). Use lower cases that fully denote the observation, variable, or label, unless data is used as-is. Avoid spaces. Use underscores rather than periods to indicate white space. Avoid special characters — "percent" or "pct" is better than "%."

Finally, realize that organization and labels are meaningful because others will use your data. We generally do not produce data for single-use cases. Clean data with simple organization fosters its use and a shared understanding of procedures and analysis.

Principle 3: Calculation and categorization are external to data collection.

Remember that measures are models of observations. Measures are not wholly faithful representations of data. By recording observations faithfully and in their original form, the map from observation to measure is visible, allowing for assessment, reproduction, updating, and adjustment.

I find beginning analysts (especially students) often misunderstand what data are, mistaking them for statistics calculated from data (e.g., means are not data, but parsimonious representations of data). The misunderstanding is often due to us older analysts not spending much time teaching what data are or conveying what is vital about data collection.

3A: All calculations should occur outside the data repository.

In business, the spreadsheet is the engine of analysis. Principle 3 is met by merely keeping an original, un-adulterated copy of the data in a separate sheet or file. For social scientists, all calculations and other statistical procedures should occur in statistical programs and not in the native data. In either realm, parsing calculations and analytical methods from data collection and preservation has logistical benefits. Updating an analysis means merely updating the data set (again in the native form) called by the procedure if scripts and functions are well-documented. Automating reporting and analysis is a big deal in both the public and private sectors. Carrying calculations, summaries, and analysis within the data structure gets in the way of efficient updating.

3B: Do not summarize data during collection.

Unless the need is pressing, do not summarize data during collection. Summaries necessarily imply the loss of information. The prohibition of summarizing includes the calculation of things like totals and averages within the data structure. Again, remember to be faithful to the data-generating process. All summaries are easily repeatable with well-documented analytical routines.

In the case of text-as-data, you should endeavor to carry the full text alongside the summary if a summary is required. Again, this summary should occur after collection.

Principle 4: If wrong, be wrong consistently.

This principle may sound odd at first, and it requires an understanding of the distinction between reliability and validity for classification or categorization. For purposes here, validity is simply whether an observation is put in the "right" category or classified "correctly." Reliability pertains to whether observations, cases, subjects, or items that are alike are categorized or classified in the same way. Both qualities are essential for data collection and measurement.

4A: Privilege reliability over validity in initial attempts to classify or categorize data.

I have employed or developed myriad categorization and classification schemes. These adventures have taught me that being consistently wrong in classification or categorization is better than being right on average, but with wide variance. Getting classification or categorization correct (imbued with validity) is an iterative process that is facilitated by high degrees of reliability. There are three contexts in which you find yourself adjusting assigned categories or classifications:

1. Classification is not correct, but reliability is high.
2. Classification is correct on average, but reliability is low.
3. Classification is not correct, and reliability is low.

If finding yourself in the first situation, adjustment is easy, because all incorrect classifications can be assigned a new classification en masse. Note, however, number two means all data must be re-coded observation by observation. In other words, you must repeat the entire endeavor for want of reliability. Finally, number three means you probably don't understand the concept well enough yet to be classifying observations.

4B: Classification or categorization should be mutually exclusive.

Ambiguous and double categorization is the enemy of reliability. Within a column, each observation should receive one and only one code. Taken together, 4A and 4B are the enemies of comparison. It is essential to remember that all research and analysis (whether public, private, or academic sectors), is an effort to compare things across time or space and make choices about what to do. If you feel the need to double categorize cases or observations, do so in a separate column. Though remember, if codes are reliable, the safety net of double coding isn't as necessary.

In Summation

Whether in business, the public sector, or the academy, answering questions and making decisions depends first on information garnered in observation — recording and preserving data. There is nothing complicated about employing these principles, and they make life much easier for the analyst. On a final note, it is always tempting to use existing data sets. It's great when data are adaptable to many questions. Use caution. All data exists for a purpose likely unrelated to your own. These principles cannot answer the conceptual problem of fit between your objective and existing data.

Originally published at https://www.samuelworkman.org on May 13, 2020.

Data Science Data Collection Logistics Data Organization Classification

Follow

Written by Samuel Workman

255 Followers · Writer for Towards Data Science

Professor, Data & Statistical Consultant, West Virginian, Author of The Dynamics of Bureaucracy in the U.S. Government https://amzn.to/3ilKSuh