## Introduction & Purpose

Data Literacy is the ability to think critically about data, and transform data into products that will help other people learn.[[1]](#endnote-1) As professionals working to make a positive impact on society, data literacy is essential, but is something we’re rarely taught.

This Data Literacy 101 guide will help you and your team build your data literacy muscles. The guide is organized in the following sections:

* **Data Terms**: Common data terms that analysts all-too-often assume their audiences understand. These definitions have been simplified; your organization may wish to use different definitions that better align with terminology in your issue area.
* **Reviewing Data Products**: How to be a savvy consumer of data presented to you.
* **Conducting Data Analyses**: A common process for analyzing data.
* **Resources & Endnotes:** Sources for this handout and further reading.

## Data Terms[[2]](#endnote-2)

**Record**: One instance of data that’s collected. For example, one record for a student reading assessment represents one student’s reading performance at one point in time.

**Variable**: One question or field collected in a dataset. For example, gender is a common variable contained in a client intake dataset.

**Dataset**: A data set is a technical term for a collection of data. Typically, a data set refers to the content of a single table (think of one worksheet in Excel), with multiple records and multiple variables.

**Database**: An electronic system that stores many interconnected datasets.

**Data Types**: Each variable in a dataset has a type, and different data types lend themselves to different kinds of analysis. Common types of data you may encounter include:

* Open Text
* Geographic (i.e., Addresses or Latitude and Longitude)
* Categorical or Picklist
* Numeric
* Date and Time
* Images
* Unique identifiers

**Qualitative Data:** Information that describes or categorizes something - the color of the sky, the smell of perfume, music genres, or coffee bean flavors are examples of qualitative data. Longer text, like case notes, interview transcripts, or product reviews, are also considered qualitative data.

**Quantitative Data:** Data represented by numbers. For example, monthly revenue, distance of a race and time of the winner, calories in a meal, temperature, salary, etc..

**Metadata**: Information about a dataset. All files typically have metadata such as “last updated date” and “last updated user”. Some datasets obtained by third-party sources contains robust metadata assembled in a data dictionary, which explains the data type for each field, what each field represents and how the data were collected.

**Data Visualization:** The process and product of visually representing data using charts, drawings, infographics, or other means. Visualizing data often helps analysts and audiences interpret data more easily than tables or raw numbers.

**File Types**: The format of the electronic file that contains data you’re interested in. For example, .csv, .xls, .pdf., .doc, or .txt. Some analysis software can accommodate only certain file types (e.g., Excel can open .xls and .csv files but not .pdf or .doc files).

**Data Cleaning**: The process of reviewing and editing or omitting raw data received to ensure data used in an analysis is as accurate as possible. Messy data can be caused by system failures, human error, or differences in data collection practice between individuals and over time. Aspects of data cleaning often include:

* Editing existing but messy data
* Removing duplicate records
* Identifying missing values on existing records
* Identifying records missing entirely

**Data Pipeline**: The pathway that a single data point or record follows as it is manipulated, analyzed, and made ready for presentation. Many elements of a data pipeline can be automated with the right technology and skills. Common steps in a data pipeline include: Find data, access / download data, verify data is in expected format, clean and transform data, analyze data, and present results.

**Normalization**: A process to make data points comparable in the same way. For example, crime rates are often presented as incidence per 1,000 residents so rates are comparable across different-sized cities and towns.

**Outlier**: An outlier is a piece of data that is distant from the remaining set of data.

**Measures of Central Tendency**: Different ways of measuring the typical value of a variable. While averages are the most common measures, median and modes may be more appropriate depending on the situation. For example, distributions of income often have very high and very low value outliers that skew the average, so median is often more appropriate.

* Average: Sum of all measurements for a variable divided by the number of observations of that variable.
* Median: Middle value that separates the higher half from the lower half of values for a variable.
* Mode: Most frequent value of a variable.

**Standard Deviation**: A common measure of the spread of data values.

**Distribution**: A list of all values for a variable and the frequency those values occur in a dataset. A common distribution is the normal distribution or bell curve, where most data points cluster around the average value.

**Histogram**: A column chart that visualizes the distribution for a variable. Histograms are useful for understanding the distribution of data.

**Statistical Significance**: The likelihood that a relationship between variables is reliable and not due to random chance.

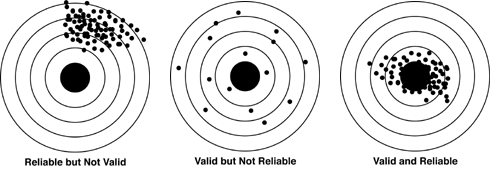
**P Value:** The probability that a result reviewed is due to chance or regularly occurring variation. Depending on the field, a relationship will be considered statically significant if the P Value is less than 0.05 – that is, there is less than a 5% chance the results are a fluke. If a result is claimed to be statistically significant, it is typically accompanied by a P Value and the Standard Deviation of the data.

**Correlation vs. Causation:** An important distinction when interpreting analysis results. Most statistical tests looking at data can estimate correlation, but an experiment with some kind of control group is typically needed to statistically prove out a causal relationship.

* Correlation: A relationship exists between two or more variables
* Causation: A *causal* relationship exists between two or more variables
* Example: Smoking is correlated with alcoholism, but doesn’t cause alcoholism. However, smoking causes an increase in the risk of developing lung cancer.

Reliability vs. Validity: An important distinction when understanding the quality of an assessment or measurement tool.

* Reliability: The extent to which an assessment or measurement will provide a consistent result if the underlying condition remains the same.
* Validity: The extent to which an assessment accurately measures the underlying condition.



## Reviewing Data Products

To become a savvy consumer of information, you should ask yourself three questions when revieiwing any data product: Where does it come from, what can you learn from it, and what can you do with it.[[3]](#endnote-3)

### Where does it come from?

First, you should understand the quality of the data source the data product is built upon. Some of the questions to consider when understanding the sources of a data product include:

* What are the strengths and weaknesses of this source?
* What real life behavior do the data represent?
* What information is emphasized?
* What relationships between different data elements are shown?
* What level of granularity is shown?
* What is the scope of the data (e.g., time, program, or geography)

### What can I learn from it?

As a data consumer, your next job is to identify what a dataset or data product means for you. In other words, analyze the data. In doing this, it’s helpful to look for the unexpected and have comparison points to put the results in context. A good data product author helps his or her reader by choosing appropriate data visualizations that include comparison points and highlight insights. However, in most cases you as a data consumer need to do some of the work of finding meaning in data. Common approaches I’ve found valuable include:

* Comparing to Expectations
* Looking for Patterns or Relationships
* Looking for Outliers
* Reviewing Trends Over Time
* Avoiding Common Pitfalls

**Comparing to Expectations**

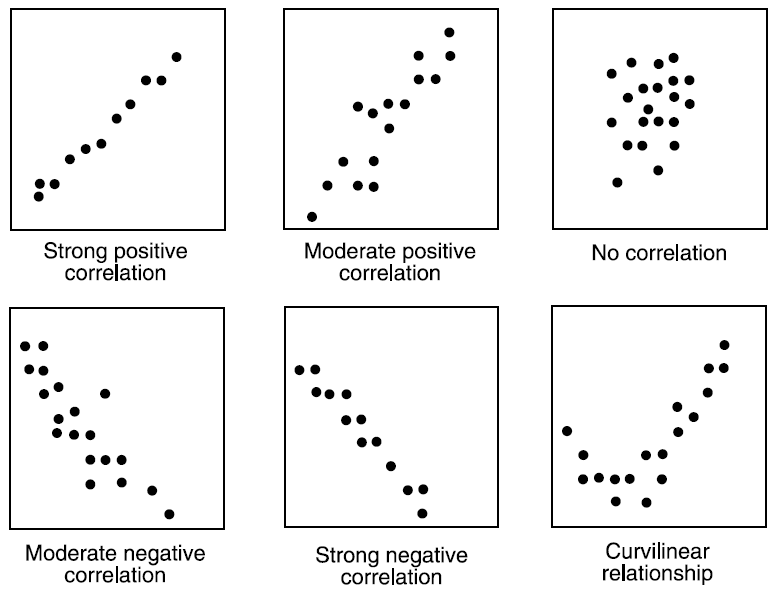
When looking at data, a helpful starting place is to have some expectations about what the data will reveal. You can then compare the actual data to your expectations. Where there any surprises? If so, were your perceptions off? Or perhaps you weren’t looking at the right data?

To develop expectations, try to complete the following sentence templates:

* Of the groups represented in this data, I expect \_\_\_\_\_\_ will have the most positive results, and \_\_\_\_\_ will have the most negative results.
* I think we will see a relationship between \_\_\_\_\_\_\_ and \_\_\_\_\_\_\_\_ in this data.
* When comparing \_\_\_\_\_\_ to last year, I expect to see \_\_\_\_\_\_.
* My hypothesis is that we will see \_\_\_\_\_\_\_\_\_ change over time.
* For \_\_\_\_\_\_\_\_\_\_\_, I’d be really happy if our result was \_\_\_\_\_\_\_\_, and really concerned if our result was \_\_\_\_\_\_.

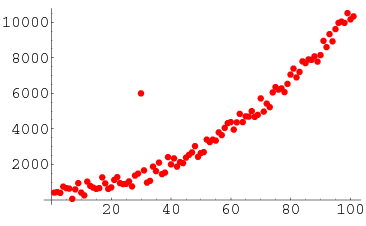
**Looking for Patterns or Relationships**

Many graphs and tables show relationships or correlations between two or more variables. When reviewing data, try to identify where a clear correlation exists. In addition, identify where you would expect a relationship to exist but one isn’t apparent. A few examples of relationships you might observe:



**Looking for Outliers**

Outliers represent a value **significantly above or below expectations**, either compared to a benchmark, a group average, or a trend over time. Look for outliers when comparing groups represented in a graph. When you spot an outlier, you may want to ask for more detailed information about that data point to understand what’s going on. An example of an outlier:



Rest of data ->

<- Outlier

**Reviewing Trends over Time**

**Trends** represent a pattern of change **over time**. Thus, not all data products will show a trend – only those where data is organized by time. Trends can be generally positive or negative, or can reveal regular patterns of variation (also known as seasonality). A few examples of trends:

**Avoiding Common Pitfalls**

Data, especially when aggregated, can be deceiving in several ways. Here are some common pitfalls to be aware of before making conclusions and taking action based on data:

* **Bad Sample:** Often a dataset contains only a sample of the individuals in a population – however it’s important to consider whether your sample is representative. For example, imagine you’ve sent a survey to your organization’s current clients. If you receive a response from 5% of your clients, can you generalize the results? Or are those the 5% who are most interested and willing to talk with you? While there are calculations you can run to determine what an ideal sample size is, you should also hold a discussion with your team to explore whether the sample of responses you have feels representative enough to make decisions.
* **Averages**: We so often default to summarizing data with averages. However, these averages may hide meaningful minimum / maximum values or outliers. The truth is there that no person is average in every respect, so designing a program or product for the “average person” means you’re designing for nobody at all.[[4]](#endnote-4) In addition, averages are quite problematic when a data set has a few very high or low values – medians may be better measures of central tendancy in these cases.
* **False Seasonality**: Trends over one or two months may reverse themselves in future months – on a longer time scale no trend actually exists. Be wary of claims of seasonality based on only one or two seasonal cycles of data.
* **Outliers & Data Quality**: Outliers may be due to data quality challenges as opposed to actual results. This is why reviewing a few outliers in depth can be a helpful exercise to understand your data.
* **No Trend**: You may find no clear trends, patterns, outliers, or implications may exist, and that’s okay! It’s natural for us to try to find meaningful patterns – but the reality of complex situation is that many factors are not accounted for that may influence results.

### What Can I Do With It?

Your last and most important task as a data consumer is deciding how to take action based on the data you’ve reviewed.

It’s important to note here that getting input **from multiple perspectives** will almost always help you reach better decisions on actions to take from data.[[5]](#endnote-5) Too often, interpreting and reporting results is delegated to the most technically-minded person in a nonprofit. Leaving data interpretation to only one person leaves important perspectives out of the process. I’ve found that an effective approach to reviewing and taking action on data involves both individual preparation and group discussion.

The previous steps in this guide outline the individual preparation steps: reviewing a data product to understand where it comes from, what can you learn from it, and what can you do with it.

Next, you should bring the results to a group. Consider the following questions:

* Check: Do we think the data are accurate? Why or why not?
* Reflect: What does the data tell us? What’s the story behind this data, and what else should we keep in mind as we’re interpreting this?
* Brainstorm: What should we collectively do considering these results?
* Commit: Of all of our potential actions, which can we build into activities we already have planned? What are the few are most important actions for us to commit to?

## Data Analysis

In general, you can approach a data analysis project using the following six-step process.

**1. Define the Question:** First, you need to define one or a small number of guiding questions for your analysis. These can come from your organization’s learning agenda if you have one. Be sure to prioritize questions, and focus on those where you anticipate being able to take action based on the result.

**2. Data Assessment**: Next, review the datasets for data quality. Identify any variables with unexpected or unreasonable values, and gut-check totals to look for missing records.

**3. Transform Dataset**: You will likely have to reshape the data files so they are in a format that will enable efficient and meaningful analysis. Often, this involves building a table where one row represents each client, and separate columns hold results from assessments conducted at different periods of time. Additional columns may be created to simplify demographic categories, or to calculate subtotals adding together multiple columns.

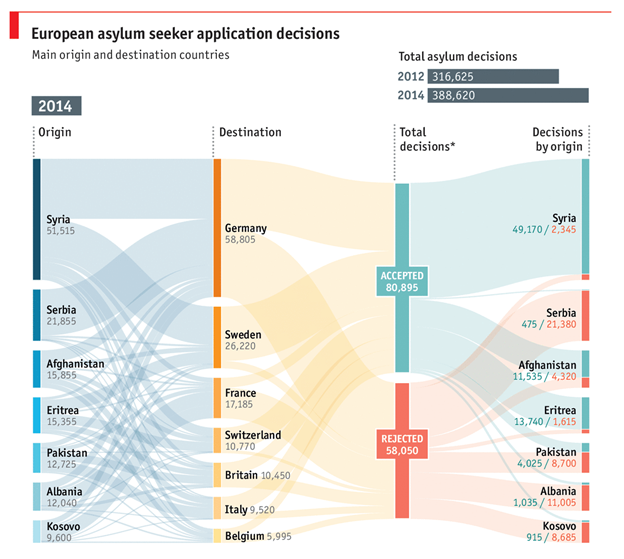
**4. Visualize**: Once the data is in the proper shape, build graphs and/or summary tables to gain an initial understanding of the results. Histograms are often helpful at this step to view the distributions of different variables. This step may also involve experimenting with various sub-sets of your data if time allows (i.e., outcomes for high-risk clients vs. low-risk clients).

**5. Analyze, Interpret**: Continue building summary tables and graphs to explore the learning questions you set out in the beginning of the process. If variables appear to have a relationship, consider applying statistical tests to suggest the level of confidence you can have in the results. While statistical tests are not an end-all-be-all, they can be useful by suggesting whether or not positive changes observed may be due to chance.

**6. Draft & Iterate:** Draft initial results in an appropriate data product (report, presentation, etc.) and layer on initial thoughts around the implications of the results. Then, lead a wider team through a group process to reflect on those results. This often brings to light new ways of looking at data and new learning questions, which you can incorporate in the final version of the analysis.

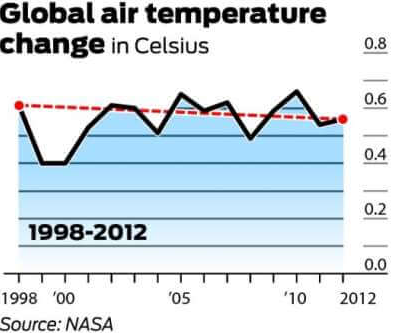
## Data Product Example 1:

This is an example of an effective data product, which you can discuss with your team.

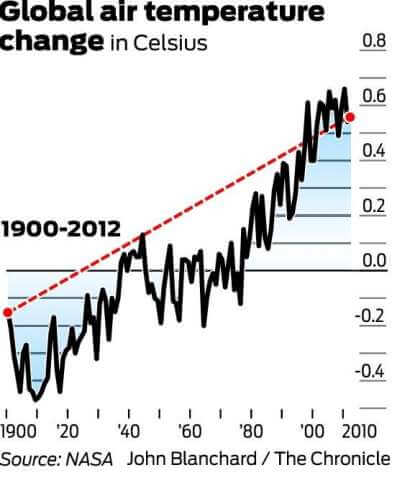


## Data Product Example 2:

MISLEADING GRAPH:



TRUTHFUL GRAPH:



## Resources & End Notes

**Further Reading**

Databasic.io – A suite of web-based tools that make it easy for beginners to work with data in fun ways. <https://databasic.io/en/>.

Data Fluency: Empowering your Organization with Effective Data Communication – A comprehensive book on effective data communications. Especially applicable for organizations using dashboarding tools like Tableau, Cliq, or PowerBI. <https://www.juiceanalytics.com/data-fluency/>

Data Smart: Using Data Science to Transform Information into Insight – One of the most reader-friendly books to introduce data science concepts. All concepts are presented in Excel, and the final chapter introduces R as a more efficient analysis tool. <https://www.wiley.com/en-us/Data+Smart%3A+Using+Data+Science+to+Transform+Information+into+Insight-p-9781118661468>.

You Don’t Need a Data Scientist, You Need a Data Culture - An effective blog exploring what a nonprofit data culture looks like and barriers to avoid. <https://digitalimpact.org/you-dont-need-a-data-scientist-you-need-a-data-culture/>.

The Intimidation Gap: How to Build Data Literacy in Nonprofits - This article outlines several simple tips to improve your nonprofit’s data practice. <http://www.peelleadershipcentre.org/intimidation-gap-build-data-literacy-nonprofits/>.

What Type of Data Should My Nonprofit or Foundation Collect? - This article outlines concepts that may help you “right-size” your organization’s data collection activities. Hint: Start with data you will use. <https://blog.techsoup.org/posts/what-type-of-data-should-my-nonprofit-or-foundation-collect>.

5 Steps your Nonprofit Needs to Take When Performing a Data Audit – This article is a helpful overview of steps to perform when conducting a data audit, and has informed my approach over the last few years. <https://actiongraphicsnj.com/blog/performing-a-data-audit/>.

Urban Institute Evidence Toolkit: Learning Agendas – This whitepaper provides an easy-to-consume summary of what a learning agenda is and how one may be used. <https://www.urban.org/sites/default/files/publication/97406/evidence_toolkit_learning_agendas_2.pdf>

How to Prioritize Decisions as a Team with a Questions and Assumptions Activity – This entry on Zapier’s blog is a wonderful outline of how to engage your team in thinking critically about the questions and assumptions underlying any strategy. <https://zapier.com/blog/team-decisions-design-thinking/>

Data Viz Project – This online catalogue shows many different types of data visualizations, and serves as a great visual encyclopedia of data presentation options. <https://datavizproject.com/>.

**End Notes**

1. See: Jonathan Gray, Liliana Bounegru, Lucy Chambers, “The Data Journalism Handbook,” The Open Knowledge Foundation, accessed December, 2018. https://datajournalismhandbook.org. See also, “WTFcsv Activity Guide,” Databasic.io, accessed December 2018. <https://databasic.io/en/wtfcsv/wtfcsv-activity-guide.pdf>. See Also [↑](#endnote-ref-1)
2. Geckoboard, “Data Science Terms Explained,” Accessed in December 2018. <https://www.geckoboard.com/learn/data-literacy/data-science-glossary/>.

   [↑](#endnote-ref-2)
3. Zach Gemignani and Chris Gemignani, “Data Fluency: Empowering Your Organization with Effective Data Communication,” Wiley, 2014. <https://onlinelibrary.wiley.com/doi/book/10.1002/9781119182368>.   
    [↑](#endnote-ref-3)
4. Todd Rose, “The Myth of Average,” Presentation at TEDx Sonoma County, June 2013. <https://www.youtube.com/watch?v=4eBmyttcfU4>. [↑](#endnote-ref-4)
5. Tina de los Santos, “The Intimidation Gap: How to Build Data Literacy in Nonrofits,” The Peel Leadership Centre, July 2017. <http://www.peelleadershipcentre.org/intimidation-gap-build-data-literacy-nonprofits/>. [↑](#endnote-ref-5)