

Data Literacy 101: Understanding Data and the Extraction of Insights

Kirk Paul Lafler, sasNerd

ABSTRACT

Data is ubiquitous and growing at extraordinary rates, so a solid foundation with data essentials is needed. Topics include the fundamentals of data literacy, how to derive insights from data, and how data can help with decision-making tasks. Attendees learn about the types of data - nominal, ordinal, interval, and ratio; how to assess the quality of data; explore data using visualization techniques; and use selected statistical methods to describe data.

Keywords: SAS, data, data literacy, understand data, structured data, unstructured data, nominal, ordinal, interval, ratio, data visualization, analytics, insights

INTRODUCTION

Data can tell an essential story about the challenges facing an organization and, often, the best way to handle these challenges – at least if you take the time to listen to your data. But data alone isn't sufficient. To gain the greatest advantage with your data, you must learn how to use it effectively. This paper will guide you through the steps to help strengthen your understanding of your data, by enhancing your data literacy skills. What does this mean? You will learn valuable techniques to effectively measure changes in your data, understand data patterns, and extract insights from your data.

So, let's get started.

DATA LITERACY 101

Data literacy refers to the ability to take a data file, ask questions about the data, and come to conclusions about the data. A data literate person can read, clean, transform, analyze, and engage with data in a way that provides meaning to others. A data literate person may also demonstrate competency in using tools such as SAS, Python, R, and other software, tools, and technologies.

The good thing to recognize about data literacy is that you don't need to be a data scientist, statistician, or other professional to possess data skills. And, because data is everywhere, employees, managers, and other stakeholders can build data literacy skills by embracing and promoting a data culture within and throughout an organization.

THE IMPORTANCE OF DATA LITERACY

A basic understanding of data is important for everyone in an organization to have. The benefits of being able to read and interpret the different types of data, data files and their structures, datasets, metadata, and other data content include:

- ✓ The ability to ask more intelligent questions and receive more accurate answers about an organization's operations and processes.
- ✓ The ability to enhance productivity by having a better understanding about your data.
- ✓ The ability to make better decisions for yourself, your team, and your organization.
- ✓ The ability to gain a competitive advantage over your competitors by using available data sources.

DATA CHALLENGES

Many data challenges face organizations in the 21st century. A list of the most notable ones include:

- ✓ Inability to access standardized and clean data.
- ✓ Inability to handle big data issues related to Volume, Velocity, and Variety.
- ✓ Organization and/or staff resistance to adopting a data-driven culture.
- ✓ Inability to define a clear set of objectives about what is needed for data analysis.
- ✓ Not knowing the “best” data visualization technique to use when trying to understand your data.
- ✓ Inability to process the desired data for analysis and reporting purposes.
- ✓ Identifying Key Performance Indicators (KPIs) for decision-makers.
- ✓ Inability to select the right analytics tool.
- ✓ Inability to advance from data to analytics to results.

One of the greatest challenges you’ll have when working with data is an inability to access standardized, clean, and structured data. Consequently, you’ll be unable to advance from data to analytics to results. When this happens, the result is often the loss of valuable information about your data and an inability to make the type of decisions that are based on evidence and facts. Instead, decisions are made on nothing more than generalized perceptions.

Another data challenge is having too much or vast amounts of data, often resulting in an inability to process and/or store large data sizes. This can also result in finding and fixing data quality issues. Another issue with big data is it’s forcing organizations to evaluate and select big data technologies. From Hadoop to Apache Spark, traditional databases to NoSQL databases – the choices are many. Without using the right tools, the result can quickly escalate with out-of-control expenses and unscalable processing of large quantities of data.

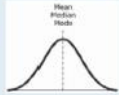
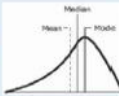
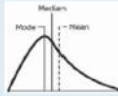
QUALITATIVE VERSUS QUANTITATIVE DATA

Data Type	Description
Nominal Data (Qualitative - Categorical)	Nominal data is the 1 st level of measurement and is expressed as alphabetical characters or as numbers, but nominal data cannot be ordered, measured, or used in arithmetic operations. This type of data is used for naming or labeling purposes, or for classifying data into mutually exclusive categories. Examples include gender, marital status, hair color, ethnicity, country, “yes/no”.
Ordinal Data (Qualitative - Categorical)	Ordinal data is the 2 nd level of measurement and has a set order to it with the data possessing an order or ranking capability. But, like nominal data, ordinal data cannot be measured, quantified, or have arithmetic operations performed. Examples include letter grades, education level, position in a queue, frequency, Likert-type questions.
Discrete (Quantitative)	Discrete data is countable and involves whole numbers (integers), can be quantified, have arithmetic operations performed, and possess values of a finite nature. Examples include sports scores, population numbers, the number of products in inventory, grains of sand.
Continuous (Quantitative – Interval and Ratio)	Continuous data is measurable, quantifiable, and can possess an infinite number of values. This type of data is further grouped into two categories: Interval data and Ratio data. Interval data is the 3 rd level of measurement and represents a numerical scale where the order of the data is known along with the difference between the data values, but a “true zero” or a fixed beginning cannot be calculated. Examples include population, rainfall, temperature, elevation, IQ score, SAT score, year. Ratio data is the 4 th level of measurement and displays order, the difference between data values, and information about “true zero”. Examples include age, height, weight, income, sales, market share, unemployment rate, product defect rate, distance, area.

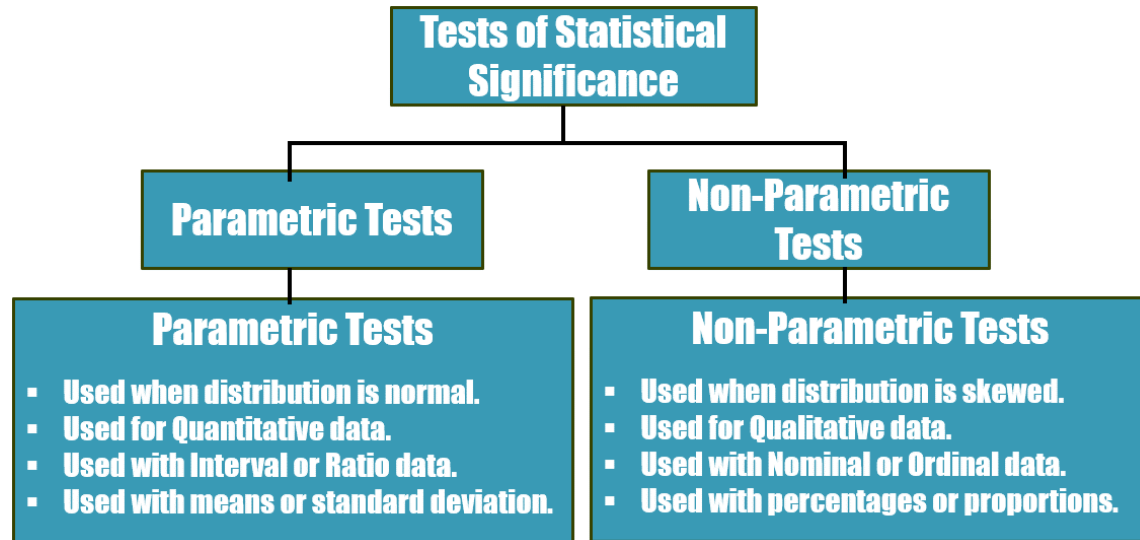
APPLICATION OF DATA SCALES TO STATISTICS

Data Scale	Mathematical Operations	Measures of Central Tendency	Measures of Variability
Nominal	Equality (=, EQ, NE)	Mode	None
Ordinal	Equality (=, EQ, NE) Comparison (>,<)	Mode Median	Range Interquartile Range
Interval	Equality (=, EQ, NE) Comparison (>,<) Addition (+), Subtraction (-)	Mode Median Arithmetic Mean	Range Interquartile Range Standard Deviation Variance
Ratio	Equality (=, EQ, NE) Comparison (>,<) Addition (+), Subtraction (-) Multiplication (x), Division (/)	Mode Median Arithmetic Mean Geometric Mean	Range Interquartile Range Standard Deviation Variance Relative Standard Deviation

QUANTITATIVE DATA ATTRIBUTES

Attribute	Description
Shape	The shape or symmetry of data in a dataset refers to how the points are distributed throughout the set. A popular way to view the shape of data is with a histogram. The data arrangement or shape displayed by a histogram are frequently referred to as: ▪ Symmetrical (Uniform)  ▪ Bimodal  ▪ Left-skewed  ▪ Right-skewed 
Center	The center in a dataset describes the location of a typical data point value. Three measures of center are the arithmetic mean (average), median, and mode.
Spread	The spread in a dataset describes the variation of the data. Typical measures of spread are the range, quartiles, standard deviation, and variance.
Outliers	An outlier in a dataset describes a data value that is much smaller or larger (or lies outside) most of the other values in a random sample produced from a population in a set of data. Two popular ways to view outliers is visually with a box plot and with statistical tests such as z scores to determine extreme data points.

TESTS OF STATISTICAL SIGNIFICANCE



LEVELS OF MEASUREMENT IN STATISTICS

Descriptive Statistic	Nominal	Ordinal	Interval	Ratio
Mean			Yes ❶	Yes ❶
Median		Yes	Yes ❷	Yes ❷
Mode	Yes	Yes	Yes	Yes
Range			Yes	Yes
Standard Deviation			Yes	Yes
Variance			Yes	Yes
Coefficient of Variation				Yes
Categorizes and Labels Variables	Yes	Yes	Yes	Yes
Ranks Categories in Order		Yes	Yes	Yes
Possesses Known, Equal Intervals			Yes	Yes
Possesses a “True Zero”				Yes
Frequency Distribution	Yes	Yes	Yes	Yes

Legend: ❶ Not skewed
❷ Skewed

STATISTICAL TESTS FOR ANALYZING DATA

Statistical Test	Nominal	Ordinal	Interval	Ratio
Chi-square Goodness of Fit Test ❶	Yes			
Chi-square Test of Independence ❶	Yes			
Mood's Median Test ❷		Yes		
Mann-Whitney U-test ❷		Yes		
Wilcoxon Matches Pairs Signed Rank Test ❷		Yes		
Kruskal-Wallis H Test ❷		Yes		
Spearman's rho (Rank Correlation Coeff.) ❷		Yes		
T-test ❶			Yes	Yes
ANOVA ❶			Yes	Yes
Pearson's r ❶			Yes	Yes
Simple Linear Regression ❶			Yes	Yes

Legend: ❶ Parametric – estimating the parameters of a probability distribution.

❷ Non-Parametric – does not make any assumptions of the characteristics of the sample or whether the observed data is quantitative or qualitative.

DATA FILE TYPES AND DESCRIPTIONS

File Type	Description
SAS Data Set (SAS7BDAT)	A proprietary SAS (SAS7BDAT) data format owned by SAS Institute Inc. that organizes and stores data values in the form of rows (observations) and columns (variables). A SAS dataset is processed by the SAS software as a temporary or permanent table of data in a SAS library (i.e., WORK, SASUSER, and User-assigned).
Tab-separated Text (TSV)	A text data format known as, a tab-separated values (TSV) data file, is created and used by spreadsheet programs and other software. A TSV data file contains rows of data values organized into one or more fields (or columns) where each data value is separated (or delimited) with a tab character.
Comma-separated Values (CSV)	A popular text data format that contains rows of data values organized into one or more fields (or columns) where each field is separated (or delimited) with a comma.
Excel (XLSX)	A proprietary data format owned by Microsoft Corporation and used to format, organize, store, and compute data in Excel spreadsheet software.
JavaScript Object Notation (JSON)	An open standard data format that is used to transmit web application data.

DATA FILE TYPES AND DESCRIPTIONS, continued

File Type	Description
HyperText Markup Language (HTML)	HTML is widely used to access data from websites. There are a variety of ways to access or scrape data from HTML including a Yahoo-developed tool called YQL.
Relational Data Base Management Systems (RDBMS)	There are numerous RDBMS systems including Oracle, MySQL, Teradata, SQL-Server, DB2, Sybase. Users typically interact with RDBMS technology using SQL to access data.
GIF, JPG, TIFF, and PNG	These are popular image file formats used on websites and file / photo viewers.
W3C-Recommended (RDF)	Uses URLs as identifiers.
eXtensible Markup Language (XML)	Widely used as transport files to maintain the structure in data.

STRATEGIES FOR IMPROVING DATA LITERACY SKILLS

The good news is that you don't necessarily need to be a data scientist to use and interpret data effectively. To be successful, you'll need to develop and hone a series of data literacy skills. A few suggestions to help improve data literacy skills include:

- ✓ Acquire basic software skills → SAS, SAS Studio, Python, R.
- ✓ Practice data cleaning techniques using "open" online data files.
- ✓ Practice data transformation techniques such as sorting, concatenation, merges / joins, and transposes.
- ✓ Study examples of the various techniques → data access, visualization, etc.
- ✓ Seek out and practice "best" practice techniques and guidance.
- ✓ Read published papers on LexJansen.com and SAS Communities.
- ✓ Attend and/or participate at conferences → SESUG, MWSUG, WUSS, PharmaSUG, Local SAS User Groups, ASA, Analytics Groups.
- ✓ Enroll in beginner, intermediate, and advanced online courses (you may even find some that are free).
- ✓ And most importantly, Have Fun and Enjoy the journey!

EXTRACTING INSIGHTS FROM DATA

Once you've developed your skills and understanding about data, you'll be able to apply your newfound knowledge to collect, clean, transform, analyze, and interpret your results. With practice, you'll learn how to improve your decision-making skills, communicate your findings, and produce positive outcomes to help your organization be more successful with all that data.

Adhering to the following steps will help make your journey easier. Practice, and soon you'll become a pro.

Task Step	Not Complete	In Progress	Complete
Form a Team - Experienced data user - Data Manager			
Establish Goals and Metrics - Identify questions you need answers to - Identify the metrics that will be used to measure your goal results			
Identify, Inventory, Gather Data - Collect Quantitative Data - Collect Qualitative Data			
Assess and Discuss Data Needs - Review datasets - Identify observations - Revise questions, as necessary - Assess whether additional data is needed			
Analyze Data - Clean data, as necessary - Transform data, as necessary - Perform an exploratory data analysis (EDA) - Identify data patterns - Identify data inconsistencies - Identify changes over time - Compare data to other data - Discover root causes			
Plan and Evaluate - Align data points to determine outcomes - Monitor and evaluate - Refine (or optimize), as necessary			

CONCLUSION

Acquiring data literacy skills is essential for anyone who wants to gain the greatest advantage with data. This paper provides information about data, and details on how you can improve your data literacy skills. By following a simple step-by-step process, you will acquire valuable techniques to effectively measure changes in your data, understand data patterns, and extract insights from your data.

ACKNOWLEDGMENTS

The author thanks the SESUG 2024 Conference Committee, particularly the Statistics, Modeling, and Analytics Section Chairs, Abbas Tavakoli and Mostafa Zahed, for accepting my paper; the SESUG 2024 Academic Chair, Jim Blum, and the Operation Chair, Lucia Alexander, for organizing and supporting a great “in-person” conference event; SAS Institute Inc. for providing SAS users with wonderful software; and SAS users everywhere for being the nicest people anywhere!

TRADEMARKS

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brands and product names are trademarks of their respective companies.

AUTHOR CONTACT INFORMATION

Kirk Paul Lafler is a consultant, developer, programmer, educator, and data scientist; and teaches SAS Programming and Data Management in the Statistics Department at San Diego State University. Kirk also provides project-based consulting and programming services to client organizations in a variety of industries including healthcare, life sciences, and business; and teaches “virtual” and “live” SAS, SQL, Python, Database Management Systems (DBMS) technologies (e.g., Oracle, SQL-Server, Teradata, MySQL, MongoDB, PostgreSQL, AWS), Excel, R, cloud-based technologies, and other software and tools. Currently, Kirk serves as the Western Users of SAS Software (WUSS) Executive Committee (EC) Open-Source Advocate and Coordinator and is actively involved with several proprietary and open-source software user groups and conference committees. Kirk is the author of several books including the popular [PROC SQL: Beyond the Basics Using SAS, Third Edition \(SAS Press, 2019\)](#). He is also an Invited speaker, educator, keynote, and leader; and is the recipient of 29 “Best” contributed paper, hands-on workshop (HOW), and poster awards.

Comments and suggestions are encouraged and can be sent to:

Kirk Paul Lafler, sasNerd

Consultant, Developer, Programmer, Data Scientist, Educator, and Author

Specializing in SAS® / Python / SQL / Database Management Systems / Excel / R / AWS / Cloud-based Technologies

E-mail: KirkLafler@cs.com

LinkedIn: <https://www.linkedin.com/in/KirkPaulLafler/>

Twitter: @sasNerd