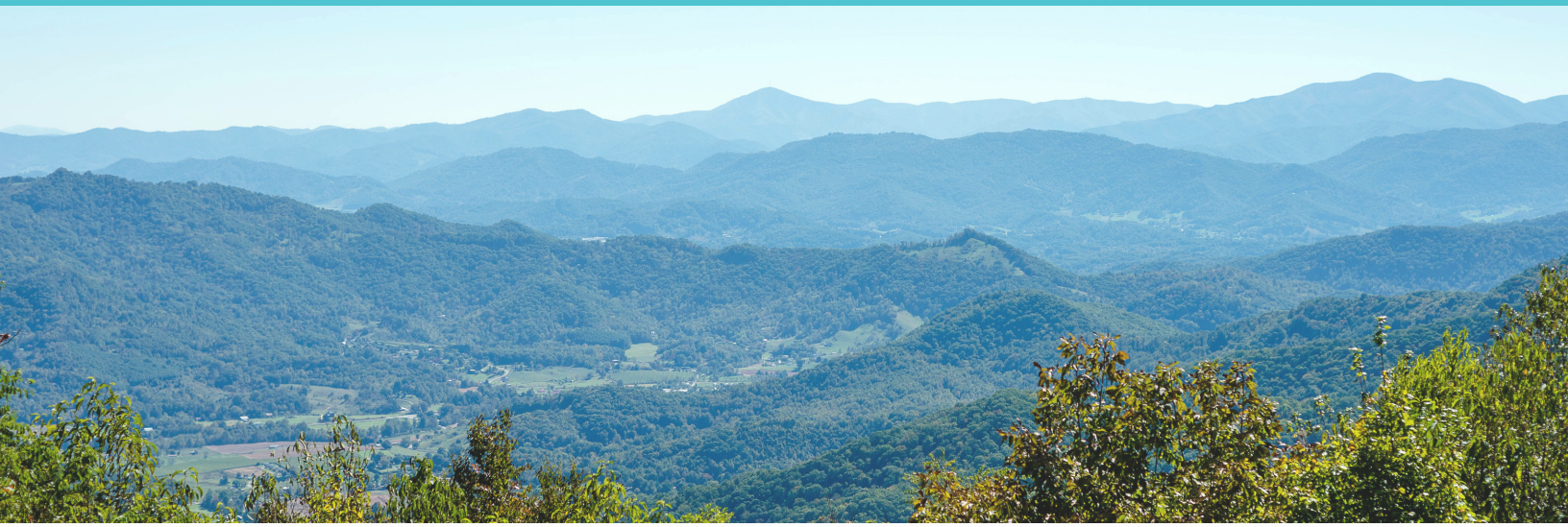
fahe.org
(859) 986-2321
319 Oak St, Berea, KY 40403
development@fahe.org



Data Consumer Pathway

Author: Camila Moreno
AmeriCorps Project, August 2022

Table of Contents

Introduction – Strengthening capacity through data literacy

"harnessing the power of data science in the service of humanity." *– DataKind*

In the wake of the ***Digital Revolution*** leaders across all sectors increasingly recognize the need for 'digital skills' and 'data literacy' as the rise of big data continues to transform how industries operate to create value or change. Notwithstanding, digital and data literacy training are not widespread, leading to considerable gaps in both access to digital skills and data knowledge. As the data landscape continues to evolve, individual awareness and understanding of data practices and their implications are central to closing the data literacy gap.

High quality data for underserved areas is notoriously hard to find and use. Consistent, accessible data is required to dismantle the narratives that do not reflect the complexity of individuals and communities' which nonprofits serve, and wholly illustrate the challenges facing underserved societies. To create lasting economic growth, we must learn to *harness the power of data* to contextualize the uniqueness of such disinvested communities and enhance how we tell stories with data.

The goal of this toolkit is to highlight data literacy gaps within the organization to increase capacity, not only within Fahe as an organization, but also extending support to members in managing and utilizing data to its greatest potential. The toolkit is intended to empower all people with varying levels of data knowledge and skill to become more comfortable with data to help drive lasting solutions for housing. In addition to filling a need, this work offers opportunities to engage and collaborate at various levels of organizational efforts more efficiently.

Data Literacy Toolkit

<u>What is it?</u>

The rise in big data has prompted higher demand for data literacy as a sought-after skillset, yet training and education efforts lag behind technological advancements. According to Forbes, over 90% of the world's data was amassed over the last few years alone, yet only about 1% of that data has been utilized in any way. A significant number of data projects never make is past the preliminary stages simply because of data silos and limited expertise. Data silos prevent organizations from being able to easily access, analyze and aptly communicate with the data.

Traditional data literacy education has focused on statistical and technical know-how, but a growing number of organizations recognize the benefits of investing in basic data skills and analysis training for all employees, regardless of their specific roles. A data

literacy toolkit is a resource developed to cultivate accessibility to data and its applications to optimize productivity and individual capabilities.

Data literacy, in the simplest of terms, refers to a broad set of knowledge and skills around the use of data for solving-real world problems. However, after data collection and processing, we must also consider the means in which findings are communicated inclusively, and effectively. How can we improve the ways in which we not only think about data, but also how we can share information in a reliable, collective, and secure manner?

Why use a toolkit?

Ability = Actual skill, either mental or physical; native or acquired.

Capacity = Potential to develop a skill, usually mental; native, as opposed to acquired.

Data driven work is pivotal to the advancement of policy change. By updating existing practices for data management, collection, processing and communication, we can more efficiently track changes, and build evidence of what works to help drive data-backed solutions for housing. Addressing data literacy gaps within the organization boosts the ***ability*** of nonprofits to not only measure outcomes, but also the ***capacity*** to identify the most efficient pathways for deploying resources where it's needed.

Even so, raw data without context is futile. Data is personal and more often than not, how individuals interpret information is determined by the degree to which they can connect with the message. Thus, data literacy encompasses the ability to simultaneously:

a. Work with data, including collection, management and analysis, and
b. Communicate data, which includes transparency of reporting, context and visualization & presentation.

That being said, it's important to note that data literacy is different for everyone, hence the skills individuals acquire will vary depending on their specific role or purpose within the organization. For the purposes of this toolkit, three different personas have been developed to help guide each individual through their skills-building and learning. Each pathway is tailored to the specific persona the individual may identify best with, offering resources and tools that align most with your needs and level of knowledge and skills.

Data Literacy Personas

Building a successful data culture throughout an organization begins with a clear and purposeful education pathway. Each role within in an organization is unique and thus, so are the data skills required to fulfill day-to-day responsibilities. Not everyone will need a high level of data fluency, therefore education pathways should offer flexibility in learning and reflect actual individual needs.

The following section describes three data literacy personas intended to provide a tailored pathway towards specific data literacy outcomes. Data literacy personas were developed based on various roles that enable a data-driven culture. Each persona pathway offers individual's resources that best align with their role and needs.

Data Consumer: Data consumers interface with data in their day-to-day to make informed decisions. Individuals who identify as a data consumer may not pursue additional training managing and interpreting data, but may need support cultivating skills to have an informed conversation with a data professional. This persona seeks to stay knowledgeable and learn more about the "why" behind the outcome.

Data Scientist: Data producers are the individuals who believe in the power of data and hope to cultivate a more structured, data-driven culture within their organization. Data producers offer support to data consumers and decision makers by assisting the navigation of access to data, using data and communicating with data. These individuals understand data management practices, and are familiar with existing data analysis programs, but may need additional support building trust in data to a non-technical audience.

Decisionmakers: Decision makers are leaders that frame the narratives which shape the attitudes and approaches to data collection, analysis, and communication within the organization. These individuals have a responsibility to their organization to support employees in not only cultivating a data-driven culture, but also ensuring that data goals align with existing capacity.

Data consumer pathway for data literacy

Pathway outcome: Boost numeracy skillset by learning to think critically about and interpret data products / discern reliable – unreliable sources / understand common data interpretation basics.

Introduction: Data interpretation is the process of extracting information from data that has been already been collected, analyzed and packaged for presentation. How we interpret [consume] data relies upon, not only how we connect with the message, but also a deeper understanding of ***numeracy***, e.g. data types and their applications.

Fundamentally, developing data literacy strongly coincides with building trust in data. As responsible consumers of data it's important to critically reflect on not only the impetus behind the collection and defining of data, but also how framing and presentation shapes narratives.

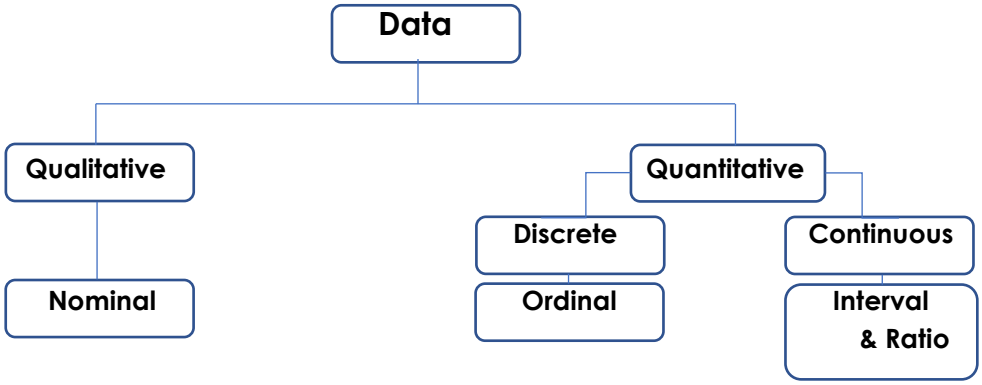
After working through the Data Consumer pathway, you should be able to:

- *Understand what is data, how & why its collected and the importance of context.*
- *Discern the differences between data types and their applications.*
- *Critically assess data quality and understand what is being communicated.*

What is data? Data is a collection of facts, observations and information that is gathered for analysis to inform decisions. It is the foundation for the information we gain about the world. Data can be either ***quantitative*** (numerical) or ***qualitative*** (categorical), but both are commonly used together to provide comprehensive insights into a sample population.

Understanding data types*:** Data types are broken down by the ***scale of measurement. Data types vary, and depending on the selected data's scale of measurement, determines how it is measured or evaluated. Similarly, understanding the differences between data types, and their ***attributes***, directly impacts your ability to appropriately interpret the meaning of the data and derive meaningful conclusions.

Data types are organized in one of two ways: ***discrete*** or ***continuous***. Discrete variables are distinct and finite. These variables can only take on certain values and are *counted*, not measured. Continuous variables are everything else. These values represent relative measures rather than actual counts.



How are data types used? Data types are a way we categorize different types of ***variables***. With each data type there is a statistical method and visualization tool best suited for analysis. In other words, nominal data cannot be analyzed or interpreted the same way as interval data because their ***measurement properties*** are fundamentally different. Tables 1 and 2 illustrate what operations are valid for each data type.

Why do data types matter: Distinctions between data types are not always clear and is reliant on the variable being measured, which consequently affects not only the tools and methods used for analysis, but also how they are presented and communicated.

Data analysis is primarily divided into two broad categories: Descriptive and Inferential statistics. ***Descriptive statistics*** describe the data, while ***inferential statistics*** involves more complex calculations that allow the user to make inferences, or predictions based off the sample dataset. The majority of data products we encounter on a day-to-day basis have undergone a combination of descriptive and inferential statistics, but what is important to understand as a data consumer is how those numbers shape insights and critically assess the story behind the data. The following section deconstructs each data type and their applications. **Scale of measurement (NOIR)**

Table 1

	Qualitative		Quantitative	
Measurement Properties	**Nominal**	**Ordinal**	**Interval**	**Ratio**
Categories/Identity	YES	YES	YES	YES
Meaningful Rank	NO	YES	YES	YES
Known, equal intervals.	NO	NO	YES	YES
Absolute or true zero.	NO	NO	NO	YES

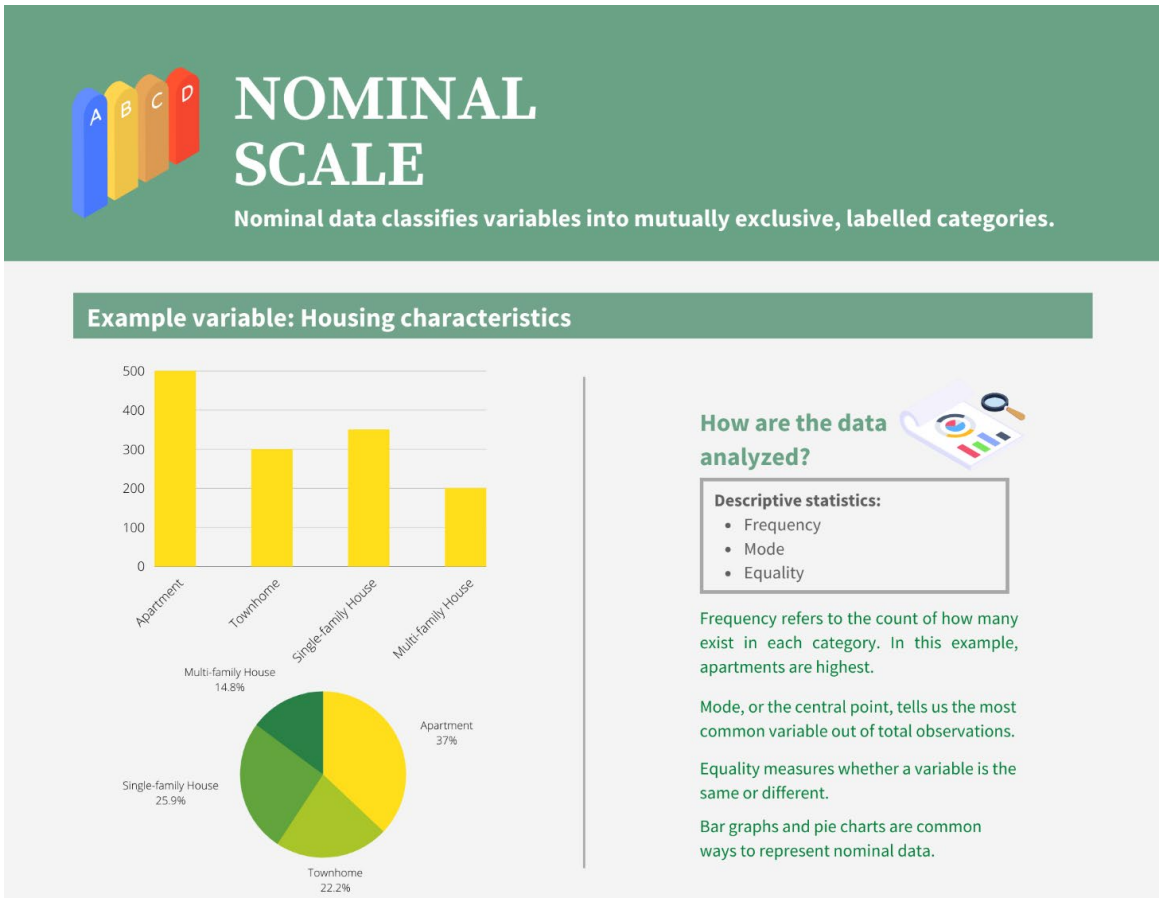
Table 2

OPERATION	NOMINAL	ORDINAL	INTERVAL	RATIO
EQUALITY	✓	✓	✓	✓
ORDER		✓	✓	✓
ADD / SUBTRACT			✓	✓
MULTIPLY / DIVIDE				✓
ARITHMETIC MEAN			✓	✓
MEDIAN		✓	✓	✓
MODE	✓	✓	✓	✓
VARIANCE			✓	✓

Qualitative Data

Qualitative, or categorical data represents data that can be grouped based off specific characteristics. Housing characteristics, such as the distinction between a house, apartment, or townhome is an example of qualitative data.

Nominal Scale



Nominal data represent discrete data that can be labelled or classified into distinct categories. Nominal data can be expressed as either text, or numbers, but are the most basic data type as they have no quantitative value. Additionally, nominal variables have no order, meaning regardless of how variables are arranged, the outcome would not change. Lastly, nominal data can be either dichotomous (two categories), or multichotomous (more than two categories).

Applications of nominal data

Examples of dichotomous nominal variables include:

- Gender (male/female)
- Pass/fail or yes/no
- Democrat or Republican
- Under age 65 or over age 65

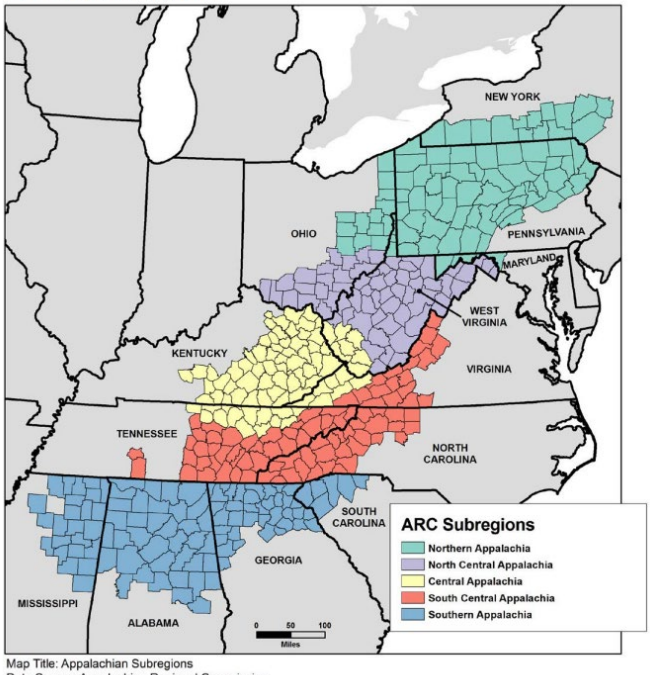
Examples of multichotomous nominal variables include:

- Marital status (single, married, widowed)
- Geographical regions (Northern, Central, Southern Appalachia)
- Race/ethnicity
- Housing characteristics

Mapping example:

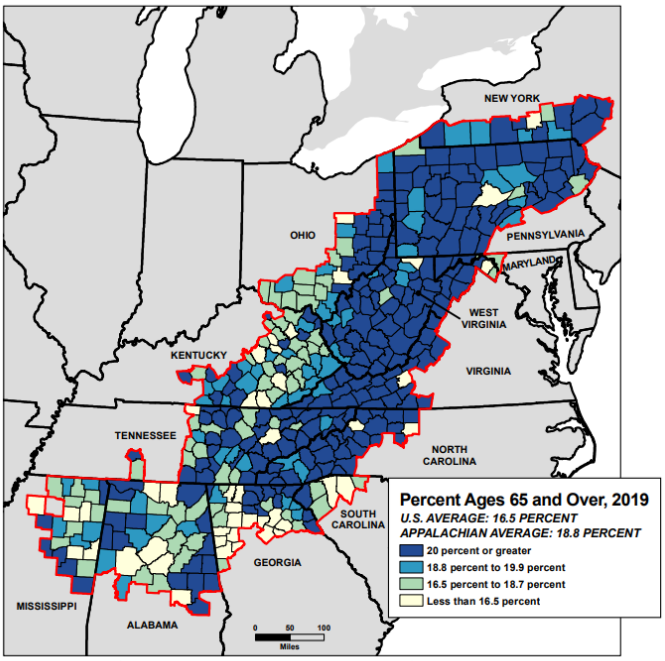
Nominal variables are useful for comparisons between groups to determine equality or inequality. For example, Appalachian counties are grouped by region based off different relatively homogenous economic, geographic, and demographic characteristics. Using this information, we can make more meaningful comparisons between groupings, such as, population distribution by age between subregions. In the example below, we can see that parts of Northern and Central Appalachia have a greater concentration of population aged 65 and over than Southern Appalachia.

Figure 1.1: Appalachian Subregions



Map Title: Appalachian Subregions
Data Source: Appalachian Regional Commission.

Figure 2.4: Percent of Population in the Appalachian Region Ages 65 and Over, July 1, 2019



Map Title: Percent of Population in the Appalachian Region Ages 65 and Over, July 1, 2019
Data Source: U.S. Census Bureau, Vintage 2019 Population Estimates.

Conclusion

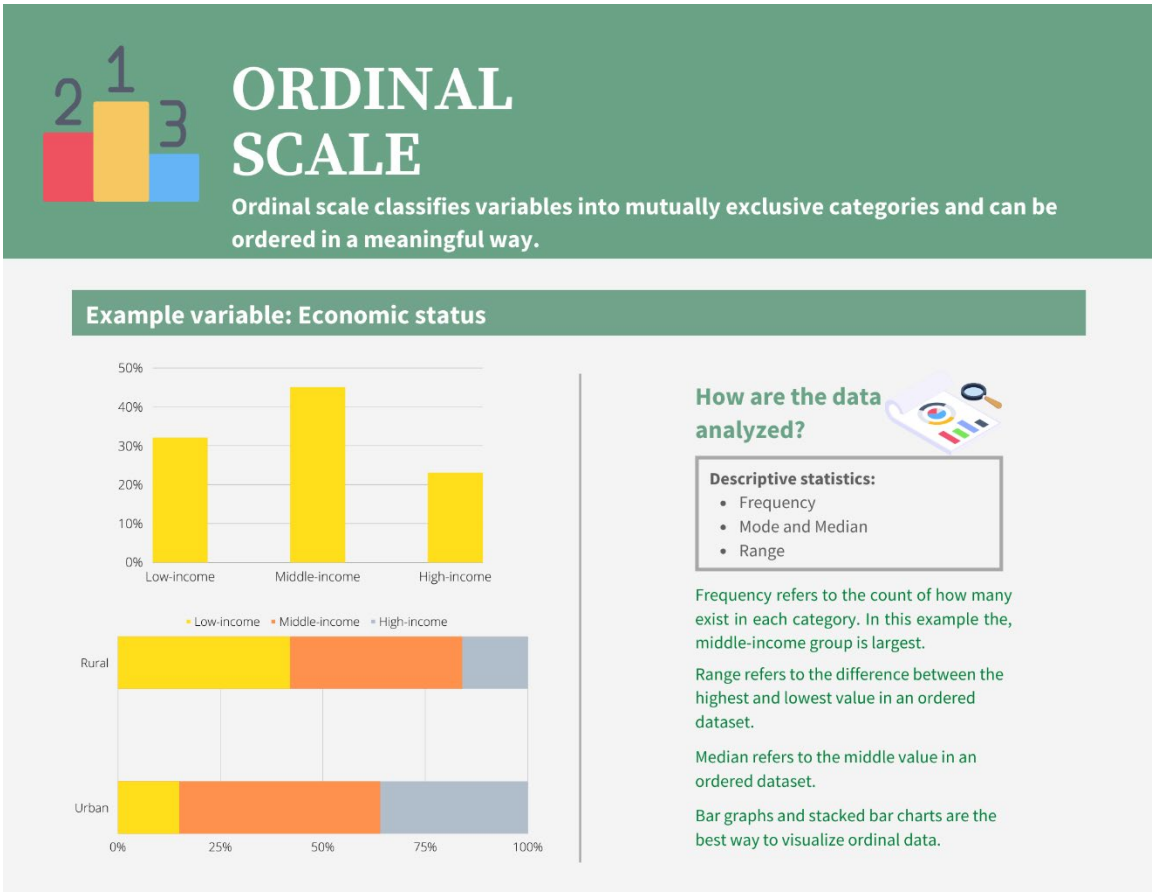
Advantages nominal data

- Fast and easy to use – most often nominal data is collected through closed-ended survey questions, which allow for larger amounts of responses to be collected quickly. Because many responses can be collected over a shorter period of time, reliability also increases.

Disadvantages of nominal data

- Not quantifiable – nominal data are the lowest level of measurement and there is no quantitative value. Nominal data only *describes* categories.

Ordinal Scale



Ordinal data represents values that are discrete, i.e. countable, and can be ordered in a meaningful way. Like nominal data, ordinal values are divided into mutually exclusive categories, but are sorted based on position or rank. Because there is an inherent numerical order, ordinal data can be used to measure if something is greater or less than a baseline. However, we cannot measure absolute magnitude, e.g. the actual difference between ranks, nor can we perform arithmetic operations (+ / -, × / ÷).

Applications of ordinal scale

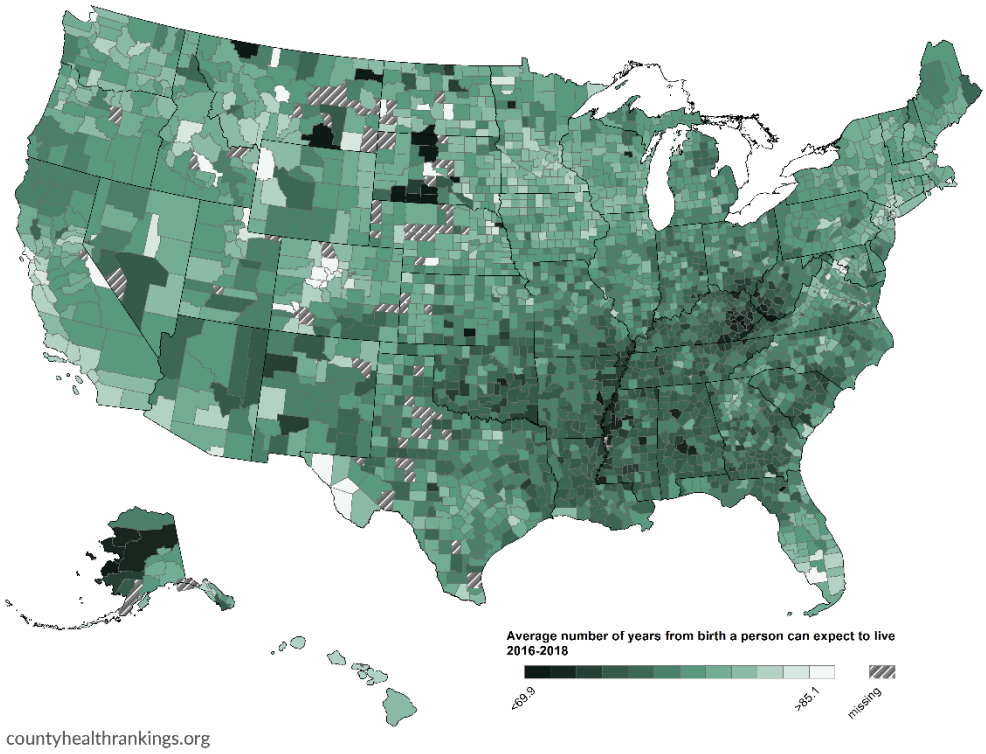
Examples of ordinal variables

- Likert scale questionnaire (strongly disagree, disagree, neutral, agree, strongly agree)
- Competition placements (1st, 2nd, 3rd)
- Test scores (low, average, high)
- Economic status (low, medium, high)

Mapping example:

Ordinal data are useful for comparison between variables, but provide more meaningful results than nominal data because values are organized based on a rating scale. For example, the County Health Rankings map below displays overall rankings of life expectancy between all U.S. counties based off several county-level measures. While these ranks highlight disparities of health, we cannot determine the true difference between rankings. In other words, it would be misleading to claim that overall life expectancy in the Southeast is 10x worse than life expectancy in the West.

Life Expectancy Among U.S. Counties (Rankings 2020)



Conclusion

Advantages

- Easy to use for comparisons between variables. Organization of variables is convenient once they have been ordered.
- More meaningful comparisons than nominal scale.

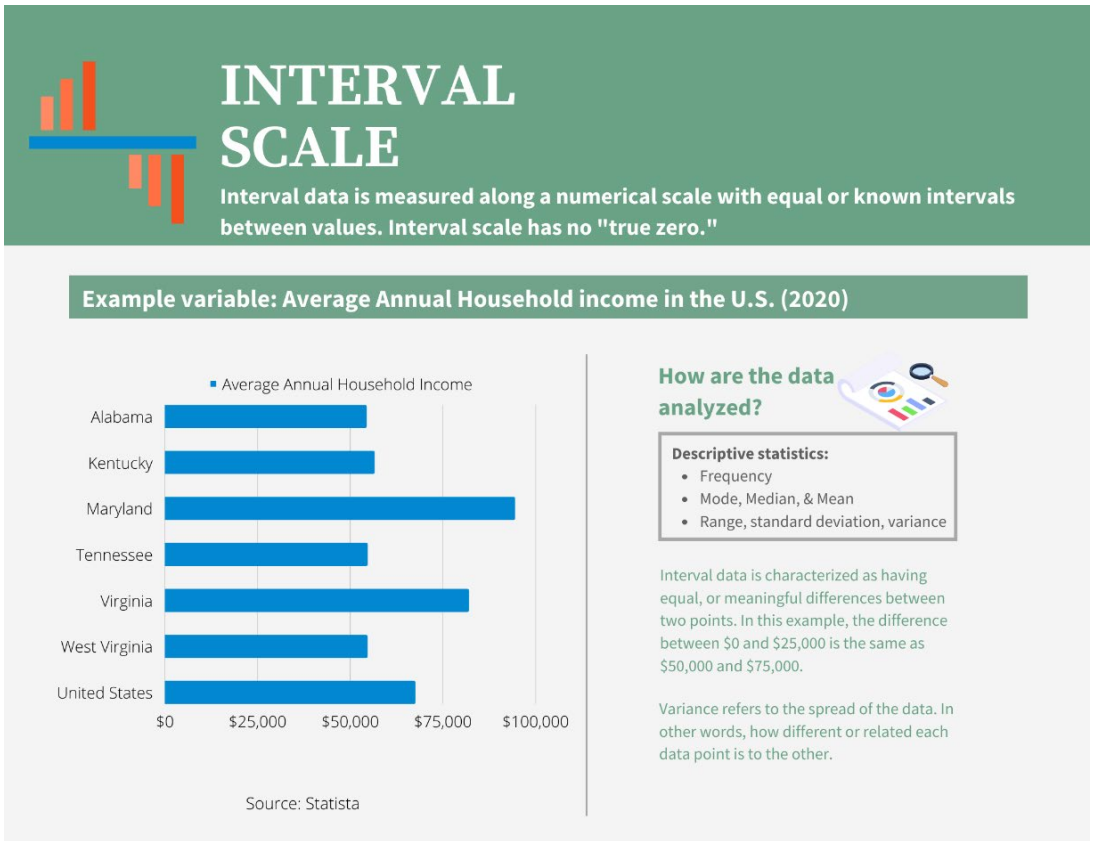
Disadvantages

- Differences between groups are not intrinsically known, so comparison between rankings can create a bias that is not representative of a larger population.

Quantitative Data

Quantitative data are any variable that can be counted or measured in some way, and assigned a numerical value. Quantitative analysis seeks to answer questions like "how much" or "how often", for example "how many people migrated in and out of county X in 2020?"

Interval Scale



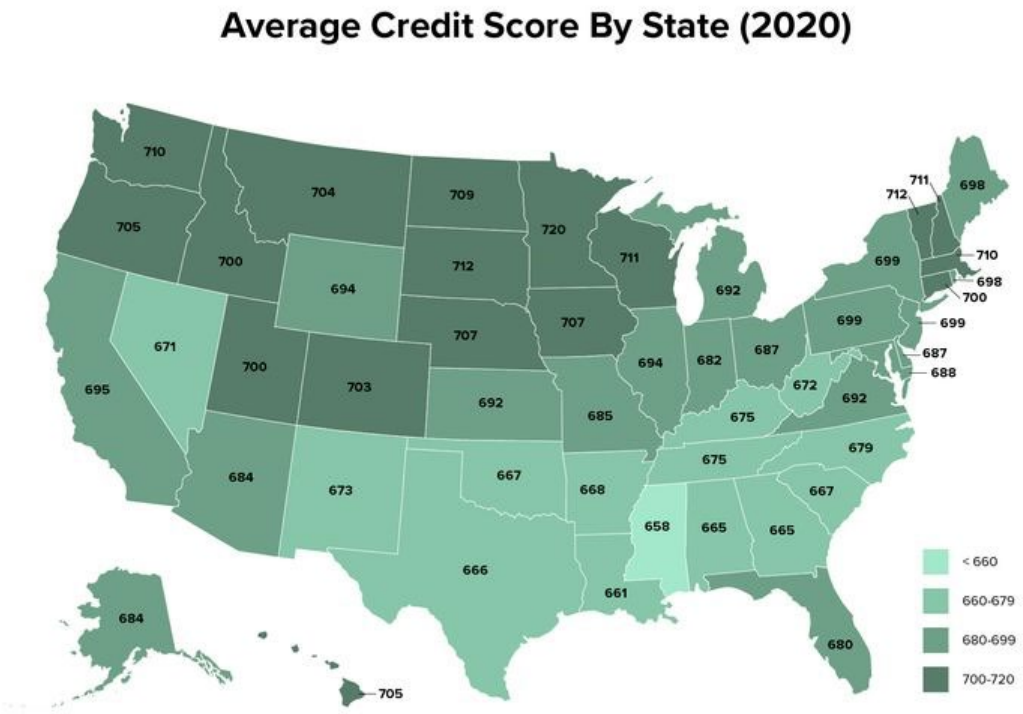
Interval scale classifies values into unique, and ordered categories. Interval variables are normally continuous, measured along a numerical scale where the degree of difference between each value is meaningful, or fixed. Most importantly, interval scale has no true zero, meaning that a value of zero does not indicate an absence of the measured variable. Most arithmetic and statistical operations can be performed with interval scale (mean, median, mode; +/ -), but ratios cannot be determined.

Applications of interval scale

Examples of interval variables

- Temperature (°F or °C)
- Credit scores
- Average income level expressed as a range ($10k - $20k, $30k - $40K… $100K)
- Time (on a 12-hr clock)

Mapping example: When interval scale is plotted on a map, or graph, classes are displayed as a range of values. Because the zero point is arbitrary, values can be negative, but the important thing to remember is that the difference between adjacent values are fixed across the entire scale. The example below displays a range of average credit scores by state, where the lowest scores are ***aggregated*** and represented as anything below 660. This represents interval scale because a credit score below 660 is an arbitrary threshold assigned to the range of values to indicate "low" scores, making it easier to compare between states. Nonetheless, an individual state average compared to another is not meaningful because a value of zero doesn't signify 'zero credit', therefore it is only a relative measure of people's capacity to repay a loan.



Conclusions

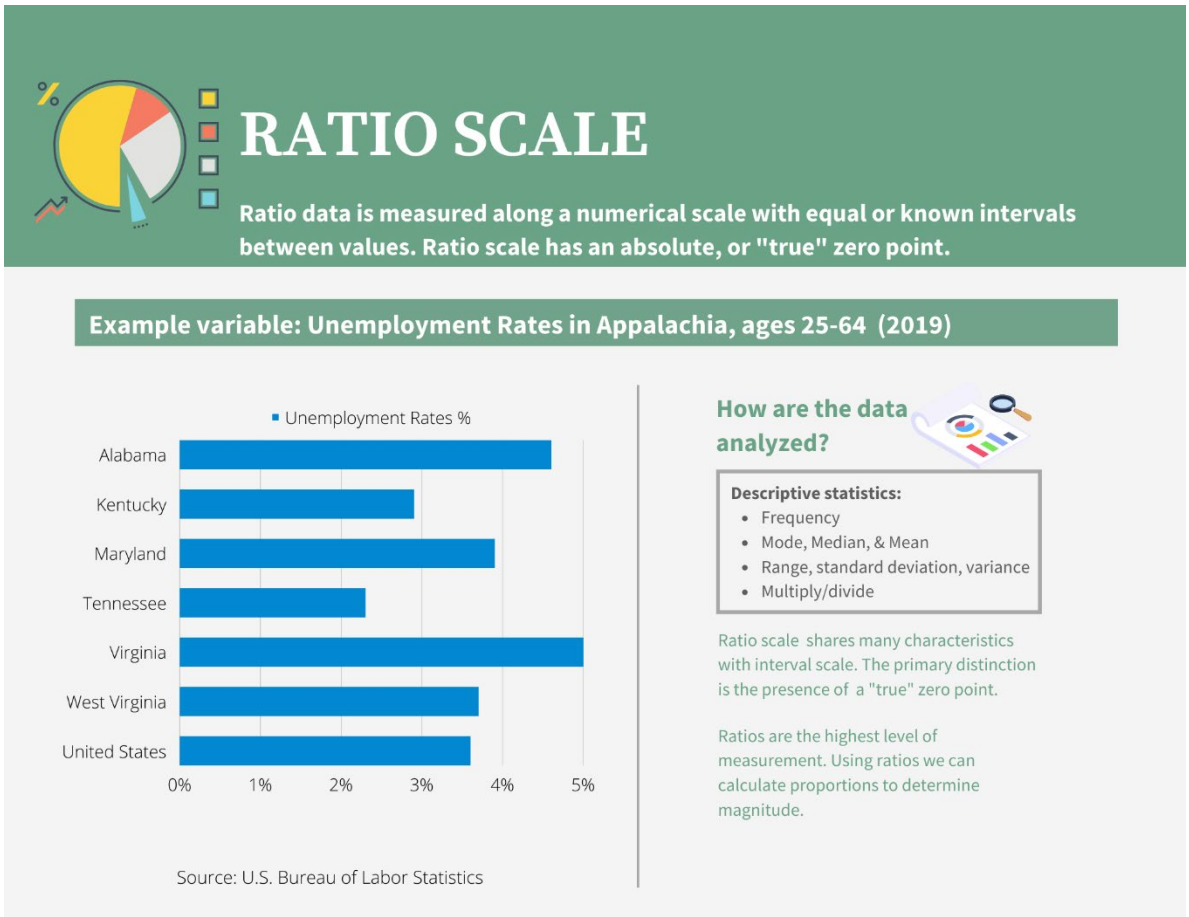
Advantages

- ❖ Determines order and direction of data. Differences between adjacent values are equal and meaningful. Values are quantifiable.

Disadvantages

- ❖ No true zero therefore ratios (proportion/magnitude) cannot be calculated.

Ratio Scale



Ratio scale is the most comprehensive measurement scale. Ratio variables are either continuous or discrete and share several of the same characteristics as interval scale. Ratio scale assigns categories and order, and is measured along a numerical scale with fixed or known intervals. The primary distinction between ratio and interval, however, are that ratio scale does have a 'true' or absolute zero. In other words, 'zero' indicates the absence of the measured variable and can *never* be a negative value. This allows us to perform all possible operations (add/subtract, multiple/divide, etc.) using ratio.

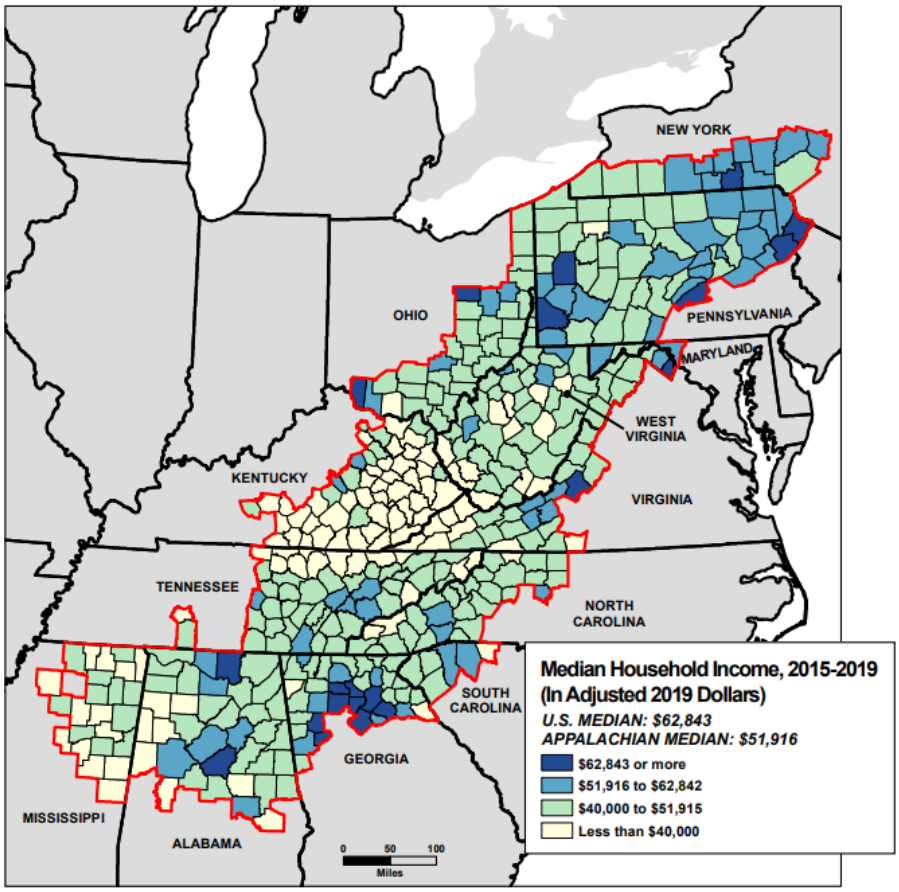
Applications of ratio scale

- Total income earned in a given year
- Total population
- Temperature (°K)
- Unemployment rate

Mapping example:

Like interval scale, ratio variables are displayed as categorial ranges with meaningful intervals between adjacent values, however these ranges are based on the absolute zero point. If it is ratio scale, values cannot fall below zero and are more precise measures than interval scale. The following example is a map of median household income for years 2015-2019. Because these values were determined based off *total income earned* per household over 5 years, it is a more precise measure of actual income earned by county, versus assigning a range to measure average *income levels.* Thus, comparisons to the U.S. median income are meaningful because values are based off a "true zero" which allows us to measure the magnitude of income disparities between places. In this example, it would be fair to claim that on average, median household income in Appalachia is about 17% less than the U.S. median income.

Figure 8.2: Median Household Income in the Appalachian Region (In Adjusted 2019 Dollars), 2015-2019



Map Title: Median Household Income in the Appalachian Region (In Adjusted 2019 Dollars), 2015-2019
Data Source: U.S. Census Bureau, 2015-2019 American Community Survey.

Conclusions

Advantages

- Determines order and direction of data. Differences between adjacent values are equal and meaningful.
- Absolute zero exists, therefore all mathematical/statistical operations may be performed with ratio scale. Proportions and magnitude can be determined.

Disadvantages

- Ratio scale does not consider external factors (e.g. additional variables that may affect the outcome). Ratios highlight a problem or pattern, but do not offer solutions.

Data interpretation – The Basics

Data interpretation is the process in which you critically evaluate data products to inform, build knowledge and share potential wisdom. Without data interpretation, data is simply data, however proper data interpretation requires some understanding of different data analysis procedures.

We use statistics and parameters to communicate population characteristics. ***Descriptive statistics*** is a form of data analysis that characterizes raw data in a meaningful way, revealing hidden trends occurring in a sample population that otherwise may be obscured. Like the name suggests, descriptive statistics help us describe our data by *summarizing existing patterns*. At the most basic level, descriptive statistics measures:

1. Central tendency – estimates where the "center" of the ***distribution*** of a dataset is located.

Measure of central tendency	**Definition**	**Formula**
Mean	Average value in a dataset. It informs the 'balance point'	$= \frac{sum\ of\ all\ values}{total\ \#\ of\ values}$
Median	Middle value in a dataset.	$For\ odd\ numbered\ datasets: M1 + M2 / 2$
Mode	Most frequent value in a data set.	$None$

- ***Why it matters:*** Descriptive statistics ensure context is accounted for. Depending on the goal or context – using the mean versus median can alter what is being reported.
    - For example, medians are generally resistant to outliers in a dataset that skew the distribution of the data left or right (we also like to examine the ***shape*** of the data to explain patterns – the curve of a skewed distribution is called a *kurtosis*).
    - As a result, medians are often used to report ***income levels and housing prices*** because they do better at describing the actual 'middle' point.

2. Variability, spread or dispersion – a measure of how spread or clustered values are relative to the mean e.g. how evenly (or unevenly) data is distributed.

Measure of Variability	**Definition**
Variance	Measure the degree of spread in a dataset.
Interquartile range	Segmented by quartiles ordered low to high, IQR measures the range for middle 50% of a dataset.
Ranges	Difference between highest and low value.
Standard deviation	Square root of variance

- Measures of spread are used in conjunction with measures of central tendency to gain a comprehensive understanding of how a particular set of values behaves. Measures of dispersion help to test the reliability of a central tendency measures, e.g. how reliable a mean is at describing a dataset. In social science sectors, measures of variability are particularly important in measuring 'inequality' in the distribution of income and wealth.

Inferential statistics are techniques that allow users to make generalizations about a population through what is known as sampling. Because collecting actual data from an entire population is near impossible, a sample population is determined using random, unbiased sampling methods. Using inferential statistics, we can identify parameters of a whole population and test hypotheses to draw conclusions.

In short: A ***statistic*** is a simple measure of a sample – a ***parameter*** describes the whole population. The uncertainty between a statistic and a parameter is called a ***sampling error.*** Parameters are tested using ***confidence intervals***, which is the variability of a statistic or the probability of a parameter falling within the expected statistical range.

Why it's important: U.S. Census Bureau American Community Survey Tables are developed using inferential statistics methods. Because it is difficult to survey such large populations, ACS tables provide estimates of sample populations. The uncertainty associated with ACS data is *unavoidable*, but as data consumers its imperative to recognize limitations of the data to avoid misinterpretation or spread of misinformation.

Learn more about descriptive and inferential statistics to cultivate your statistical literacy skills!

→ Confidence Intervals, levels, and margin of error
→ Descriptive statistics in Excel
→ Understanding Hypothesis Testing and Statistical Significance
→ What is Data Interpretation? + [Types, Method & Tools]

Data Contextualization: Translating Numbers into Narratives

What is data context: Data without context is misleading, creating barriers for communication and effective decision-making. Each data point tells a different story, but context provides the audience a more comprehensive picture of what the data points are representing, and is essential component of building trust in the data. In short, People define ***data*** to *learn new* ***information*** *– asking different questions and connecting the dots transform newly derived information that into* ***knowledge***; *context transforms knowledge into* ***wisdom***. Critical consumers of data will always ask questions about how data was collected, the decisions that led to the outcome, and how it is communicated.

Data sources matter: Reliable data is determined by the degree to which the raw data was correctly processed, and the quality of data analysis. It is also dependent on its relevance to your need and whether the data is transparent, and accessible. *Reliable data is intrinsic to the context in which it was produced*.

How to verify data and their source: Metadata is as valuable as the data itself – it is the "data about the data." Consumers, producers and decision-makers alike should rely on metadata to establish data context, and better understand how to work with data. Metadata can be broken down into three broad categories:

> *"Data is the content. Metadata is the context. Metadata can be much more revealing than data…" – Bruce Schneier, Data and Goliath*

Metadata Type	Purpose	Examples
Descriptive	Describes the content in a dataset so users can more easily identify items and make connections. It helps the user identify the "who" and "what" of the resource.	• Titles and abstracts • Bibliographical attributes (author, keywords, definitions) • Physical attributes (media type, e.g. blog, map or spreadsheet / table)
Structural	Identifies and describes how information is organized and stored, and how those different elements are related to each other.	• Table of contents • Data structure (format, type, sequence)
Administrative	Information about data management, processing, and sharing including, data quality and validity, use requirements, access and restrictions.	• Rights and permissions • Privacy restrictions • Date and time of creation • Data accuracy

Challenges: A common problem in data analysis and interpretation is the prevalence of ***statistical fallacies***. A statistical fallacy occurs when the statistical procedure is either accidentally or intentionally misleading as a result of bias or misuse. There are many kinds of statistical fallacies, but for the purposes of the consumer pathway, the most crucial to consider is the ***ecological fallacy***. An ecological fallacy occurs when aggregated data is generalized to individual people, usually by omitting a variable.

Example scenario: Low-income Appalachian counties have high crime rates. Therefore, it is a result of low-income residents.

Explanation: On an aggregated level there are high crime rates in several low-income Appalachian counties. However, one cannot assume that the correlation between low income communities and high crime rates is a cause-and-effect relationship. We must consider other factors that may correlate with instances of high crime rates, for example, there may be limited funding for police departments in places where crime is higher.

Conclusions: Data context is essential, but sometimes metadata may not be available, complete or up-to-date as a result of high costs or lack of capacity to maintain over time. While a lack of complete metadata is common, it also illustrates the ongoing challenges of data literacy where *uncertainty is inevitable*. Much of what we understand about data management, analysis and sharing is comprised of trial and error – the important thing to bear in mind is that all data has limitations and must be acknowledged when reporting and sharing to reduce bias, and the spread of misleading information.

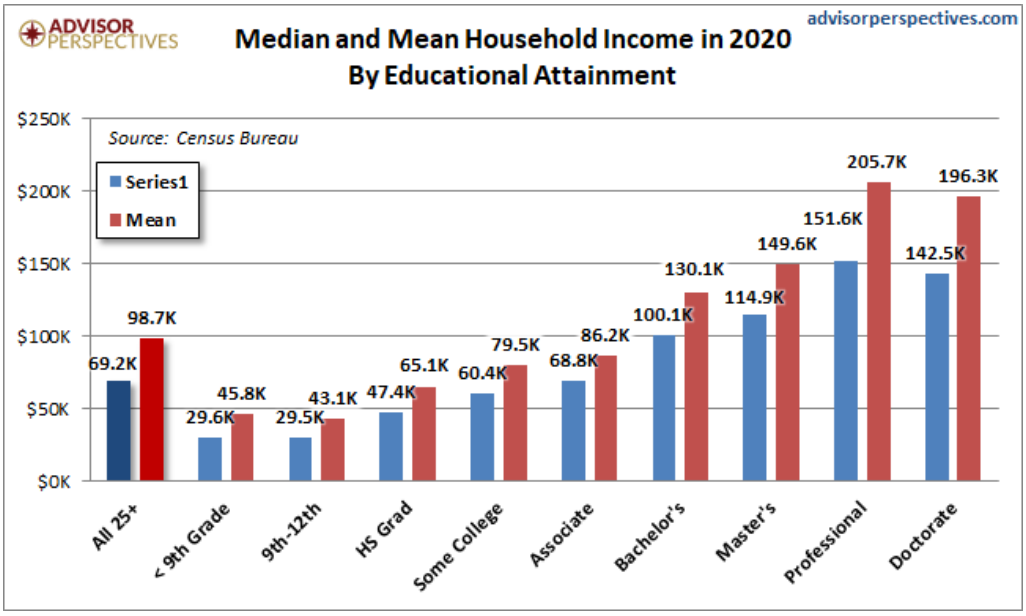
Key takeaways:

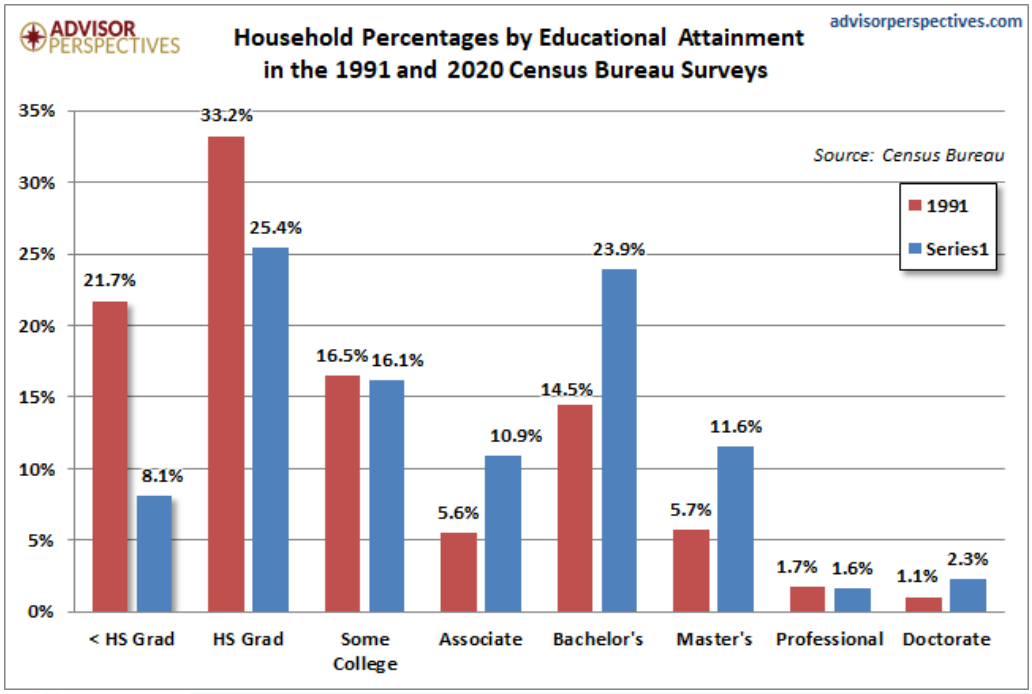
- Data is the foundation of information gathering and knowledge building.
- Data can be expressed as either text or numbers and is either discrete or continuous.
- There are FOUR scales of measurement (NOIR). Ratio is the most comprehensive scale of measurement.
- Depending on the scale of measurement, there are specific permissible arithmetic operations (+ / -, × / ÷, ≥ / ≤) and statistical operations (mean, median, mode, range, std. dev.) for each.
- Whether is it a map, or graph, context is critical. The context, framing, and visualization choices affect how information is perceived.
- Metadata is an important and useful tool for filtering out unreliable data. If it is not available, reflect on its limitations and cross reference with other sources.
- Avoid an ecological fallacy by critically assessing correlation versus causation.
- Uncertainty is inevitable because all data has limitations.

Activity – Unpacking representations of data

Household Incomes 2020: The Value of Higher Education

The following activity lists a series of graphs depicting different parameters of household incomes by educational attainment. The data is based on the Census Bureau's annual survey data for 2019 and the graphs were published in Advisor's Perspectives digital newsletter in 2021. Take some time exploring the graphs and think about your initial reactions – what information can you gather by what is being shown? After reviewing each graph and reviewing the source publication, critically reflect on what the data is demonstrating. Use the following list of questions to guide your analysis.





1. Is the source a reputable or reliable source?
2. Does the source have a particular agenda or perspective that can influence the content?
3. How was the data was collected?
4. Are the graph titles and labels clear and appropriate?
5. What kind of graphs are shown?
6. What is the scale of measurement? How do you know?
7. What is the time frame?
8. What is the unit of analysis? HINT: Look at the horizontal (X) and vertical (Y) axis.
9. How are the numbers shown and what do they represent? How are they similar or different? HINT: Percentages vs mean vs median
10. Describe the relationship between the variables. Can you identify a trend or pattern? HINT: What is the independent/dependent variable? Is the relationship direct, or inverse? Is it correlational or causative?
11. What other factors may affect the result? HINT: Is enough context provided? Cross-reference with other sources to compare.
12. Are there other possible means of conveying this information?

For more information and examples of misleading data visit any of the following resources: Statistics How To; ISAE Good and Bad Graphs; When Maps Lie

Source: Advisor Perspectives Newsletter

Answer key can be found end of document in Appendix.
