



# K-12 Data Science Learning Outcomes

## TOPIC GROUP

## CONTENT OUTCOMES AND EXAMPLES

### Data for Daily Life

**Regularly leverage data in daily life**, including to inform personal decision-making, address societal or community problems, or create solutions for others. Examples may include leveraging spreadsheets to compare investment or healthcare options, examining data as a "first impulse" to make a policy or civic argument, or helping others solve a difficult problem by sourcing data on an unknown or tricky phenomenon.

### Data Structures

**Know the basic types and structures of data**, including quantitative, qualitative, image, text, and other types of data; data tables and tidyverse formats; sources of digital data, including primary (sensors, web-traffic, etc.) and secondary (pre-collected or previously assembled data), and examples of applications which leverage data (music, product, or social recommendation algorithms, weather forecasting models, autonomous vehicles, large-language models, etc.). Examples may include a student being able to identify several examples of devices or objects and how they relate to data; students in older grades can discuss ethical implications about the use of data, personal privacy, and other social trade-offs.

### Assessing Generalization Claims

**Assess whether data generalizes or not in context**, including whether the data sufficiently represents the population or question of interest, how the collection or analysis process introduces bias, what the data says about causality, and how to check for validity issues in analysis or code. Examples may include a student assessing two competing headlines by examining the underlying data used for research or conducting a power analysis to understand the statistical significance of a result.

### Art of Critique

**Practice the art of critique for how data analyses may mislead, exaggerate, or mis-represent**, including changing axes scales, inflated data graphics, mis-communicating correlation vs. causation, confusing the difference between mean / median, creating an intentional response bias via survey question design, failing to include controls, overfitting a model, and many other examples. Students should regularly seek other sources of information, knowing that any one analysis may be flawed or that quantitative data may not tell the full story. In a classroom example, a student may learn the "data tricks" of misrepresentation and then be asked to identify them in the media for a homework assignment.

## TOPIC GROUP

## CONTENT OUTCOMES AND EXAMPLES

### Producing and Tailoring Visualizations

**Deploy data visualizations relevant to the problem and audience**, including knowing which visualization technique is best for different types of data (categorical, numeric, different distributions, etc.), knowing a toolkit of visualization types (box-plots, histograms, other visual types, etc.), confidently leverage the most appropriate software tool (e.g. spreadsheets, PowerBI, Tableau, R-Shiny) for the task, instinctively make data visuals accessible (e.g. alt-text, accessible colors), and design the visualization for the existing knowledge of their audience. Examples may include students presenting survey data on school-lunches to their student council, who have not taken an introductory data science course, wherein categorical data could be arranged in a simple bar-chart. Rather than memorizing data visualization types, students would consider the type of data they have and their audience before building or coding a chart.

### Iteration and Validation

**Comfortably iterate and validate within a data analysis cycle**, wherein students collect or source, explore multiple dimensions, produce interim summary values or visualizations, contextualize, model, analyze, communicate, and corroborate their findings from one or more datasets. This process is defined by frequent double-checking and rethinking, and includes "Exploratory Data Analysis," pre-processing, and both "unplugged" and technology-assisted validation steps. The emphasis here is "check your work, question yourself, and check it again."

### Storytelling

**Clearly tell a story or an argument with data**, including with effective presentation and speaking skills, the ability to write about a data analysis with plain-language vocabulary and any additional problem-specific terms, the ability to adapt to different audiences technical and non-technical audiences, necessary caveats and limitations of the analysis, a clear explanation for "why" their audience should care, and multiple representations of the data relevant for individual arguments (visualizations, summary statistics, process or methodology descriptions, etc.)

### Bayesian Probability

**Practice probabilistic thinking**, including comfort with uncertainty, the ability to explain and account for variation, basic rules of probability in context, and the ability to model simple expected value functions to aid decision-making. Example: a student may review a raw distribution of financial returns from their class playing a stock-market game, calculate the average return across the students in class, and then apply that average probability to their own savings for a future year via a simple expected value calculation, holding other potential changes constant. Younger students may learn the basics of "a prior guess, new data, and an updated guess informed by the new data" to apply Bayesian probability principles to their everyday life. Students in older grades may practice with multiple-variable prediction functions for a variety of problems or examine Bayes Rule itself.

## TOPIC GROUP

## CONTENT OUTCOMES AND EXAMPLES

### Hypothesis Construction

**Construct questions and hypotheses for data analyses that can be answered by data**, including a clear hypothesis or set of hypotheses, an analysis plan for individual or collaborative tasks, a draft plan to source or gather additional datasets that may be needed. All analysis plans should be deeply rooted in the domain or context of the question. Questions should be authentic and ideally student-driven in project-based formats that introduce new methods as needed. Classroom examples could include the creation of formal statistical hypothesis tests AND a collaborative project plan for task management within the same step.

### Bias in AI

**Detect and troubleshoot bias in data-driven tools, including AI models.** Students should regularly question data sources used for technology tools; evaluate the social, economic, and demographic makeup of training data; and test the output against real-world examples. Students should come to terms with the reality that all data is biased (either by collection, by observation, or through other means), but that some bias may reproduce specific social norms when employed in technology tools. Classroom examples may include research projects into training data for AI tools. Older students may experiment with testing an output of an AI model against new or rebalanced training data.

### AI Model Intuition

**Understand how data "powers" AI tools**, including machine-learning approaches, large language models, recommendation algorithms, autonomous technology, and other tools. This may include the "under-the-hood" intuition that allows students to leverage existing toolkits from statistics or other fields to question AI outputs (sample size, outliers, generalization issues of underlying data), detect errors or "hallucinations" from AI tools, and better understand AI tools' use-cases based on the selected training data. For older students, this may include practice with customizing existing AI tools or packages in existing open-source coding platforms (R, Python, etc.) on new training data to solve specific problems.

### Tool Selection

**Select and transfer between the most appropriate software tool for the problem at-hand.** Students should have a high-level knowledge of currently available tools, including their "pros," "cons," and best use-cases (spreadsheets, scripting languages such as R or Python, visualization tools such as Tableau or PowerBI, data management tools such as SQL, etc.), and classroom appropriate tools for younger learners. Students should also have practice with or exposure to multiple tools as they progress through their K-12 and postsecondary education experiences, and the ability to more easily transition between data analysis tools over time.

## TOPIC GROUP

## CONTENT OUTCOMES AND EXAMPLES

### Exploratory Analysis

**Conduct Exploratory Data Analysis (EDA) to summarize key insights from a dataset** at an interim or mid-analysis stage, including basic summary statistics, simple merges, subsetting, or filtering, common tests to analyze data for differences and similarities, distributions of key variables of interest, identification of outliers, and review or validation of end-to-end data pipelines.

### Multi-variable Tradeoffs

**Employ multi-variable modeling**, drawing upon a toolkit of multiple modeling approaches (linear functions, exponential functions, logistic functions, linear regression, polynomial regression), a number of potential "control" variables, and knowledge of tradeoffs from modeling choices (e.g. overfitting, covariates) to make either descriptive claims or predictions from data. By the end of high school, students should have the opportunity to work with complex datasets with many variables and many observations.

### Sourcing and APIs

**Easily source new data into an analysis, both informally in a tool such as a spreadsheet (e.g. candidate information on a ballot, consumer-goods comparisons, insurance policies, etc.), and automatically for a more formal analysis (e.g. via an API, web-scraping, or other advanced techniques)**

### Statistical Significance

**Construct simulations and tests for statistical significance**, with introductory "unplugged" approaches and more advanced technology-assisted approaches. This includes simulated probability distributions, bootstrapping, and conversion between traditional statistical tests (p-values and z-tests or t-tests) to technology-assisted approaches. Teachers should not allocate significant amounts of time to the "famous formulas" or other hand-written approaches, but rather introduce their techniques, the foundational mathematics, and then translate to modern techniques.

### Societal Implications

**Assess ethical trade-offs related to societal data use**, including the storage and security of personal data, data privacy law and individual rights, societal or research benefits from open data, intellectual property of data, and what may be expected for transparency and accountability in data collection. Classroom examples may include in-class, seminar-style discussions surrounding recent data breaches, Supreme Court cases, or technology regulations. Older students may be given the opportunity for a "Technology Ethics & Policy" elective in high school.

## TOPIC GROUP

## CONTENT OUTCOMES AND EXAMPLES

### Ethics During Analysis

**Consider ethics when producing and making decisions throughout all phases of the data investigation process**, including when using personal data, choosing representations, or categories; leveraging data that represents people responsibly (with attention to the social stakes of representation); utilizing secondary data or existing online data; and testing and validating output of a model for anonymity, privacy, discrimination, or other social implications. Students should be equipped with historical examples of how data analysis has been used to harm people (e.g. Tuskegee Experiment, Stanford Prison Experiment) and the technical validation techniques to prevent bias or favoritism, including data for the training and use of AI tools.

### Demystifying Careers

**Know potential career opportunities for data-related skills**, including how data skills apply to different careers across sectors, financial benefits or tradeoffs that are made possible by associated technical skills, and the level of specialization and degree obtainment required for different types of data-intensive roles.

### Research Design

**Know research and survey design best practices**, including the major differences between experiments, natural experiments, observational, and correlational analysis (and their implications); types of survey questions and design, including how to minimize response or observation bias; and other domain-specific techniques for research approaches. Students should connect the ideas of correlation vs. causation with research design choices.

### Reproducibility

**Document data sources, analysis steps, and assumptions made for reproducibility and validation of their work by others.** Practice sharing documentation through a collaboration tool or platform (e.g. an analysis report, Github). For all students, this may include creating a simple Data Sheet for a dataset if one does not already exist, describing the general topic, variables, date collated, attribution, and other basic information. For older students, this may include exposure or practice with collaboration tools like Git.Hub for sharing and validating code run on datasets.

### Software Experience

**Practice programming for data analysis**, including the basics of Python, R, SQL, SAS, Stata, or other scripting languages. This draft learning outcome goes beyond spreadsheets or drag-and-drop tools.

### Data Wrangling

**Employ cleaning, merging, and data transformation techniques**, especially for "messy" or "real-world" data, including treatment of outliers, missing data, unknown or poorly-labeled variables, and other issues.

## TOPIC GROUP

## CONTENT OUTCOMES AND EXAMPLES

### Alternate Explanations

**Identify alternative explanations and confounds for any result**, checking sources, other datasets, qualitative experiences, and prior research. Students should recognize that data and data visualizations are important but insufficient sources of evidence and information, and should be corroborated with and integrated with other sources of information, including alternate explanations or confounding factors, other samples / sampling techniques, or other plausible explanations or interpretations. Example: in a classroom activity, students should practice identifying why their analysis may be right and why it may be wrong at the same time.

### Automation

**Implement analysis automation and simple machine-learning techniques**, leveraging existing software packages with a large dataset. Students should practice training a custom model, holding a subset of data for testing / validating the model, and improving its predictive power. Examples may include an end-of-year project or competition where students train their own AI model, or implement a software package