DATA SCIENCE 4 EVERYONE

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 a complex 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 (e.g. sensors, web-traffic) and secondary (e.g. pre-collected or previously assembled data); and significant present-day applications which leverage data (e.g. recommendation algorithms for music or products; weather forecasting models; autonomous vehicles or robotics; large-language models). In a classroom example, a student may identify how devices or objects relate to data; students in older grades may learn formal tidyverse norms for structuring data, or discuss ethical implications regarding the use of different online data types and sources.
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 can say about causality, and how to check for validity issues in analysis. In a classroom example, a student may assess two competing headlines by examining the underlying data used for research, or conducting a power analysis to understand the significance of a particular result.
Art of Critique	**Practice the art of critique for how data analyses may mislead, exaggerate, or mis-represent,** including examples such as manipulating the x or y axis display, mis-communicating correlation vs. causation, confusing the difference between mean / median / mode, creating an intentional response bias in a survey, failing to include controls, overfitting a model, and many others. 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 examples 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 (e.g. categorical, numeric) or different distributions, possessing a toolkit of visualization types (box-plots, histograms, density plots, etc.), confidently leveraging the most appropriate software tool for the task (e.g. spreadsheets, PowerBI, Tableau, R-Shiny, etc.), instinctively making data visuals accessible (e.g. alt-text, accessible colors, etc.), and tailoring visuals for the existing knowledge of an audience. In a classroom example, students presenting survey data on school-lunch quality to their student council may think strategically about how to introduce the issue, with visuals that move from simple to complex. 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 data, produce interim summary values or visualizations, model, analyze, contextualize, 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 a combination of "unplugged" and technology-assisted validation steps. In a classroom example, the emphasis for students is to "check your work, question yourself, and check it again" as a general habit and with specific techniques.
Storytelling	**Tell a story or make an argument with data,** including with effective presentation and speaking skills, writing about a data analysis with plain-language vocabulary and any problem-specific terms, adapting to different technical and non-technical audiences, explaining necessary caveats and limitations of the analysis, and including clear reasons for "why" their audience should care. In a classroom example, students may be encouraged to include multiple representations of their analysis relevant for making individual arguments (visualizations, summary statistics, process or methodology descriptions, etc.)
Probabilistic Thinking	**Practice probabilistic thinking,** including comfort with uncertainty, explaining and accounting for variation, applying basic rules of probability in context, and the ability to model simple expected value functions to aid decision-making. A classroom example may include a student reviewing financial returns from their class playing a stock-market game, calculating the average return across the students over multiple years, and then applying a simple expected value calculation to their own personal savings. 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 everyday life. Students in older grades may practice with multi-variable models for a variety of prediction problems, or examine Bayes Rule itself.

TOPIC GROUP	CONTENT OUTCOMES AND EXAMPLES
Inquiry Construction	**Construct questions and hypotheses that can be answered by data,** including a clear hypothesis or set of hypotheses test(s), an analysis plan for individual or collaborative tasks, and a method to source additional datasets that may be needed. In a classroom example, students may create a formal statistical hypothesis test AND a collaborative project plan for task management within the same stage. 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.
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. In a classroom example, students may be asked to research the training data used for an AI tool deployed in a previous project, and to then document potential issues or effects on the output. Older students may experiment with testing an output of an AI model against new or rebalanced training data. 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 in data may reproduce harmful norms or outcomes.
AI Model Intuition	**Understand how data "powers" AI tools,** including machine-learning approaches, large language models, recommendation algorithms, autonomous technology, and other tools. In a classroom example, an educator may connect existing toolkits from statistics or other fields to question AI outputs (e.g. sample size, outliers, bias, or generalization questions of underlying training data). Students may be asked to explain how a given AI model was trained, and then match appropriate or inappropriate use-cases of the AI tool in question. For older students, this may include practice with “customizing” existing AI packages in open-source coding platforms (R, Python, etc.) on new training data to solve a problem.
Tool Selection	**Select and transfer between the most appropriate software tool for the problem at-hand,** including a high-level knowledge of currently available tools, including their "pros," "cons," and best use-cases (including but not limited to: spreadsheets, scripting languages, visualization tools, querying tools, and classroom-appropriate tools for younger learners). Students should have practice with or exposure to multiple tools as they progress through their K-12 and postsecondary education experiences, and the confidence to more easily transition between data analysis tools over time. School leaders and educators may plan “tool progressions” over many grade-levels.

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; subsetting or filtering to identify patterns; common tests to analyze data for differences and similarities; visualizing distributions of key variables of interest; identification of outliers; and review or validation of data pipelines into the analysis.
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 exploring tradeoffs from modeling choices (e.g. overfitting, covariates) to make either descriptive claims or predictions from data. In a classroom example or project, students should have the opportunity to work on complex datasets with many variables and many observations by the end of high school, if not sooner.
Sourcing and APIs	**Source new data into an analysis,** informally in a tool such as a spreadsheet for quick comparisons (e.g. adding a column and deploying a pivot table), formally for a large comparison or more complex research question (e.g. via merging techniques and validation), and automatically for advanced or real-time analysis (e.g. via an API or web-scraping). In a classroom or project example, this may entail adding per capita population data by state (simple), by county (several), or by a less-defined population category (foot-traffic data) as students learn more advanced techniques.
Statistical Significance	**Construct simulations and tests for statistical significance,** with introductory "unplugged" approaches and technology-assisted approaches. This includes simulated probability distributions, bootstrapping, and translation between traditional statistical tests (z-tests or t-tests) to technology-assisted simulations. In a classroom example, educators should not allocate significant time to students memorizing the "famous formulas," but rather introduce their logic, identify problems that showcase their foundational mathematics, and then identify problems that require technology to complete.
Societal Implications	**Examine 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. In a classroom example, students may engage through in-class, seminar-style discussions surrounding currents events on major data breaches, new AI tool releases, 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 responsibly; and testing and validating output of a model for anonymity, privacy, discrimination, or other social implications. In a classroom example, students may be equipped with both historical examples of how data analysis has been used to harm individuals (e.g. Tuskegee Experiment, Stanford Prison Experiment) as framing, and then the technical validation techniques to assess bias in a dataset or an AI models' outputs.
Demystifying Careers	**Understand 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	**Understand research and survey design best practices,** including the differences between randomized experiments, natural experiments, observational, and correlational analysis; types of survey questions and design, including how to minimize response or observation bias; and other domain-specific techniques for research. In a classroom example, students should connect the ideas of correlation vs. causation with research design choices regardless of format.
Reproducibility	**Document data sources, analysis steps, and assumptions made** for reproducibility and validation of work by others. Practice sharing documentation through a collaboration tool or platform (e.g. an analysis report, Github, etc.). In a classroom example, this may include creating a simple Data Sheet: a short guide 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 that is run on datasets.
Software Experience	**Practice programming for data analysis,** including the basics of Python, R, SQL, SAS, Stata, or another scripting language. 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 common issues.

TOPIC GROUP	CONTENT OUTCOMES AND EXAMPLES
Alternate Explanations	**Identify alternative explanations for any result,** including reverse causation, missing or unmeasured effects, or historical context. Students should recognize that data and data visualizations are important but insufficient sources of evidence and information, and should be corroborated and considered with other inputs (e.g. qualitative data and lived experiences, checking other sources and datasets, prior research, other samples / sampling techniques, etc.). In a classroom example, students may practice identifying several reasons 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. In a classroom example, students may engage in an end-of-year project to train their own AI model. Students would implement a software package via R or Python, reserve a subset of their data for testing / validating the model, and try to improve its predictive power through iterating on model choices.