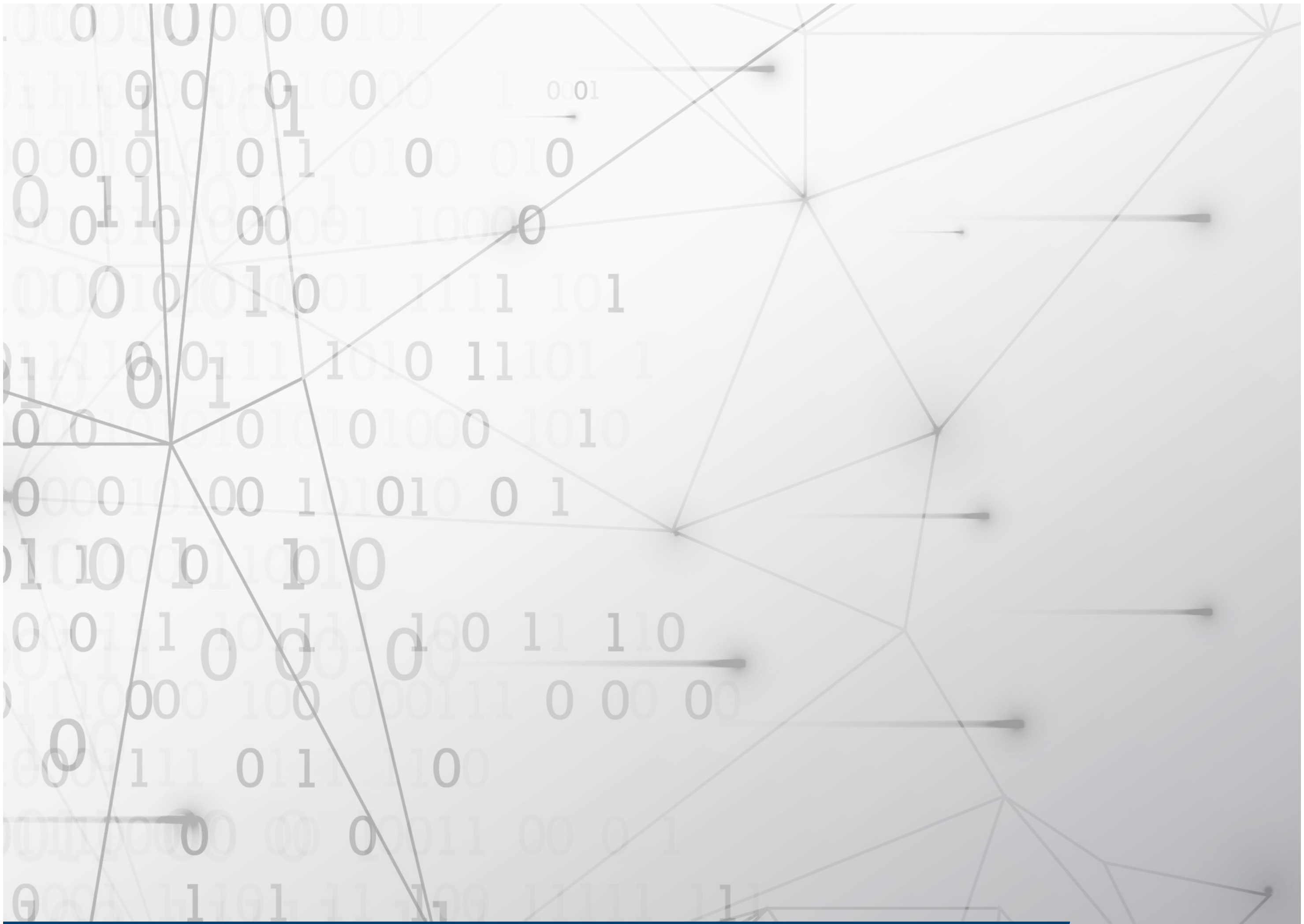

Data Science Standards and Resources

West Virginia DEPARTMENT OF EDUCATION

Created July 1, 2025

Data Science Standards and Resources

Domains

Explore Data: Exploration is to identify the potential relationships between variables. Explorations allow learners to try different hypotheses, methods, and strategies. These insights help guide decisions about how to prepare the data and which types of models are most appropriate for capturing those relationships.

Visualize Data: Visualization is any technique for creating images, diagrams, or animations to communicate a message.

Analyze Data: Analysis of data to gather useful information can be used to draw conclusions, make predictions, and make informed decisions.

Communicate Using Data: Communication about data involves using summary statistics for both categorical and numerical variables. These summaries can describe individual variables, multiple variables of the same type, or how numerical variables vary across categories. Knowing the full spectrum of data visualization potentials, not just histograms but see many examples of data visuals. Good data visualization practices include the selection of the chart and/or graph to visualize the data. Good explanation practices include defining and explaining the variables on the x and y axis, practicing presentations on data in classroom, etc. Good written explanations of data visualizations include argumentation with data.

Application of Data Science: Apply data science techniques to explore, visualize, analyze, and communicate data.

Additional Resources for Data Science

- *Teach | Data Science for Everyone*
- *Learn | Data Science for Everyone*
- *Gr. 9-12 Resources for Data Science - Google Docs*
- *Bootstrap: Data Science*

Data Sets Resources and Links

- *http://oceansofdata.org/node/202#high*
- *Datasets for K-12 | Data Science for Everyone*
- *https://guides.lib.virginia.edu/data/teachandlearn*
- *https://nces.ed.gov/*
- *Find Open Datasets and Machine Learning Projects | Kaggle*

Domain	**Explore Data**	
Cluster	**Ask and develop questions; collect data; and consider ethics and bias.**	
Standard(s)	M.DSHS.1	Describe techniques for locating and collecting small and large-scale data sets.

Content Examples and Resources

What is Big Data
https://www.khanacademy.org/computing/ap-computer-science-principles/data-analysis-101/big-data/a/what-is-big-data

Sources of Big Data
https://www.khanacademy.org/computing/ap-computer-science-principles/data-analysis-101/big-data/a/sources-of-big-data

The Power of Big Data
https://www.khanacademy.org/computing/ap-computer-science-principles/data-analysis-101/big-data/a/the-power-of-big-data

Background Knowledge for Educators

Big Data Collection
https://www.techtarget.com/searchdatamanagement/feature/Big-data-collection-processes-challenges-and-best-practices

How Big Data is Collected
https://computools.com/how-is-big-data-collected/

Domain	**Explore Data**	
Cluster	**Ask and develop questions; collect data; and consider ethics and bias.**	
Standard(s)	M.DSHS.2	Recognize a question that can be explored or answered using data science, including statistical questions.

Vocabulary

- **Classification:** a supervised learning method where the goal is to assign data points to a predicted category or class. Classification techniques include algorithms like logistic regression, decision trees, neural networks, etc.
- **Regression:** refers to estimating a continuous dependent variable based on a set of input variables.
- **Optimization:** the process of finding the maximum or minimum value of a real-valued function by systematically selecting input values from an allowed set and evaluating the function.

Content Examples and Resources

Hurricanes as Heat Engines Story Map
https://nasa.maps.arcgis.com/apps/MapSeries/index.html?appid=abc5591aaa944c9ebc7b5ea6102c73c2

Statistical Questions
https://portal.mathmedic.com/lesson-plans/course/Intro-Stats

Statistical Questions
https://youtu.be/OjzfQDFf7Uk

The Five Questions Data Science Can Answer
https://learn.microsoft.com/en-us/shows/supervisionnotrequired/8

Background Knowledge for Educators

What Kind of Questions Can Data Science Solve
https://scientistcafe.com/ids/what-kind-of-questions-can-data-science-solve.html

What Types of Questions Can Data Science Answer
https://thelead.io/data-science/what-types-of-questions-can-data-science-answer/

Domain	**Explore Data**	
Cluster	**Ask and develop questions; collect data; and consider ethics and bias.**	
Standard(s)	M.DSHS.3	Use technology to informally describe the shape, variability, and center of a distribution of data.

Vocabulary

- **Uniform distribution:** a distribution where all values have nearly equal frequency.
- **Normal distribution:** a particular distribution completely defined by its mean and standard deviation that is symmetric about the mean and creates a bell curve.
- **Skewed distribution:** a distribution that has the majority of the data values on one side.
- **Bimodal distribution:** a distribution with two modes or peaks.
- **Variability:** how spread out the data is, which can be described by range, interquartile range, standard deviation, and variance.
- **Range:** the difference between the highest and lowest values in a data set.
- **Standard deviation:** the average distance of a typical point from the mean of a data set.
- **Variance:** average of the squared distances from the mean.

Content Examples and Resources

Classifying Distributions
https://www.khanacademy.org/math/ap-statistics/quantitative-data-ap/xfb5d8e68:describing-distribution-quant/v/classifying-distributions

Summarizing Quantitative Data
https://www.khanacademy.org/math/ap-statistics/summarizing-quantitative-data-ap

The Binomial Distribution (3Blue1Brown)
Binomial distributions | Probabilities of probabilities, part 1

Sampling distribution (Khan Academy)
Introduction to sampling distributions | Sampling distributions | AP Statistics | Khan Academy

Background Knowledge for Educators

Code Emporium 5 Probability Distributions you should know as a Data Scientist
https://www.youtube.com/watch?v=CF0l5zw4t9s

Domain	**Explore Data**	
Cluster	**Ask and develop questions; collect data; and consider ethics and bias.**	
Standard(s)	M.DSHS.4	Determine possible sources of statistical bias in a study and how such bias may affect the ability to generalize the results and evaluate a variety of resources used to collect data for accuracy, perspective, credibility, relevance, and privacy concerns.

Vocabulary

- **Convenience sampling:** selecting individuals who are easy to reach.
- **Non-response bias:** when individuals from a sample can't be contacted or refuse to participate.
- **Response bias:** when there is a systematic pattern of inaccurate responses to a question in a study.
- **Statistical bias:** is anything that leads to a systematic difference between the true parameters of a population and the statistics used to estimate those parameters.
- **Undercoverage:** when some members of a population are unable to or unlikely to become part of a sample.
- **Voluntary response bias:** when subjects self-select to be a part of a sample, usually leading to an over-representation of extreme opinions, particularly negative ones.

Content Examples and Resources

Sampling: Good and Bad
https://portal.mathmedic.com/lesson-plans/course/Intro-Stats/unit/4/day/3

Background Knowledge for Educators

Types of Statistical Bias
https://online.hbs.edu/blog/post/types-of-statistical-bias

Domain	**Explore Data**	
Cluster	**Ask and develop questions; collect data; and consider ethics and bias.**	
Standard(s)	M.DSHS.5	Understand that random sampling tends to produce representative samples that support valid inferences and generalizations about a population.

Vocabulary

- **Random sampling:** using a system of chance to select a sample of a given size from a population.
- **Sampling variability:** the fact that different random samples of the same size from the same population will produce different statistics.

Content Examples and Resources

Jelly Blubber Sampling Activity

https://www.stem.org.uk/resources/elibrary/resource/75969/jellyblubber-colony

Background Knowledge for Educators

Comparing Distributions with Dot Plots

https://www.khanacademy.org/math/cc-seventh-grade-math/cc-7th-probability-statistics/cc-7th-population-sampling/v/comparing-swim-times-at-the-olympics

Reasonable Samples

https://www.khanacademy.org/math/cc-seventh-grade-math/cc-7th-probability-statistics/cc-7th-population-sampling/v/reasonable-samples

Domain	**Explore Data**	
Cluster	**Research issues, access multivariable data, and clean data.**	
Standard(s)	M.DSHS.6	Explore and understand real-world issues and problems using multivariable data sets to hypothesize solutions.

Vocabulary

- **Big data:** collections of data that meet the 3 V's criteria – Volume, Variation, and Velocity. The volume of data is "so much" that it requires tools and techniques to handle and process. Big data is "so mixed" and potentially from a variety of sources. The rate at which big data can accumulate can be "so fast" it exceeds the ability of traditional techniques and applications (Excel, Sheets, etc.).
- **Data cleaning:** the process of detecting and correcting corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.
- **Data wrangling:** the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.

Content Examples and Resources

NASA Dataset Story Maps
https://mynasadata.larc.nasa.gov/basic-page/earth-system-story-map-collections-lesson-plans

Data Wrangling: What It Is & Why It's Important
Harvard Business School
https://online.hbs.edu/blog/post/data-wrangling

Domain	**Explore Data**	
Cluster	**Research issues, access multivariable data, and clean data.**	
Standard(s)	M.DSHS.7	Access data from a variety of digital sources and apply mathematical concepts and models to solve problems in mathematics and other disciplines.

Content Examples and Resources

Common Online Data Analysis Platform (CODAP) Tool for Modeling
https://codap.concord.org/app/static/dg/en/cert/index.html

The Collection of Really Great, Interesting, Situated Datasets (CORGIS)
https://corgis-edu.github.io/corgis/csv/

Five Thirty-Eight
https://data.fivethirtyeight.com/

Gapminder Tools
https://www.gapminder.org/tools/#$chart-type=bubbles&url=v1

Google Public Data
https://www.google.com/publicdata/directory

Google Trends
https://trends.google.com

Background Knowledge for Educators

Tidy and Tame Data
Hadley Wickham
https://vita.had.co.nz/papers/tidy-data.pdf

Five Thirty-Eight
https://escholarship.org/uc/item/0rx1231m

How to Analyze a Dataset: 6 Steps
Harvard Business School
https://online.hbs.edu/blog/post/how-to-analyze-datasets

Domain	**Explore Data**	
Cluster	**Research issues, access multivariable data, and clean data.**	
Standard(s)	M.DSHS.8	Using programming techniques and spreadsheet capabilities, clean, store, analyze, and model with data sets.

Vocabulary

- **Data cleaning:** the process of detecting and correcting corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.
- **Data wrangling:** the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.

Content Examples and Resources

Code.org Big, Open, and Crowdsourced Data Lesson Plan
https://studio.code.org/s/csp9-2021/lessons/5

Free Data Analysis Software
https://codap.concord.org/

Background Knowledge for Educators

Manipulating Data in Spreadsheets
https://deepblue.lib.umich.edu/bitstream/handle/2027.42/146731/Chapter_5_Stuit.pdf?sequence=1

What is Data Cleaning?
https://www.tableau.com/learn/articles/what-is-data-cleaning

Data Wrangling: What It Is & Why It's Important
Harvard Business School
https://online.hbs.edu/blog/post/data-wrangling

Domain	**Explore Data**	
Cluster	**Research issues, access multivariable data, and clean data.**	
Standard(s)	M.DSHS.9	Compare techniques (e.g., sorting, statistics, searching) for analyzing multivariable data sets.

Vocabulary

- **A/B testing:** comparing two groups together, for example a control group and a test group, or two test groups.
- **Data mining:** the process of finding or analyzing datasets in order to discover new patterns that might improve the model.
- **Machine learning:** using pattern recognition software to find trends in data, building models that explain the trends/patterns, and then using the models to predict something.
- **Natural language processing:** refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.
- **Statistics:** the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.
- **Regression analysis:** a statistical term used to describe linear, quadratic, exponential or other functional relationship between two variables.
- **Data pipeline:** a system or series of steps that automatically moves data from one place to another—like from a survey form to a spreadsheet—often cleaning, organizing, or combining it along the way so it's ready to be used for analysis.
- **Extract, Transform, Load (ETL):** a basic process used to move data from one place to another—first taking it from a source (extract), changing it into a useful format (transform), and then storing it in a central location like a spreadsheet or database (load) so it can be analyzed.
- **Data warehouse:** a large, organized storage system where data from different sources is collected and stored in one place, making it easier to find, combine, and analyze information—kind of like a digital library for data.

Background Knowledge for Educators

Big Data Analysis Techniques
https://www.getsmarter.com/blog/career-advice/big-data-analysis-techniques/

Data Analysis Techniques
https://careerfoundry.com/en/blog/data-analytics/data-analysis-techniques/

Data Moves
https://escholarship.org/uc/item/0mg8m7g6

Domain	**Visualize Data**	
Cluster	**Display data.**	
Standard(s)	M.DSHS.10	Use appropriate tools to represent data visually.

Vocabulary

- **Box plot:** a special type of diagram showing Quartiles 1, 2 and 3 (where the data can be split into quarters) in a box, with lines extending to the lowest and highest values.
- **Dot plot:** a graphical display of data using dots.
- **Histogram:** a graphical display where quantitative data is grouped into ranges (such as "100 to 149", "150 to 199", etc.), and then plotted as bars.
- **Real number line:** is a horizontal line with arrows on both sides. It consists of the origin "0". All the positive numbers are represented on the right side of the origin and all the negative numbers are represented on the left side of the origin, with a definite scale.

Content Examples and Resources

Health and Nutrition from Core-Plus Mathematics - YouCubed

Understanding plots and creating them by hand

- Box Plot - *https://youtu.be/oajrmwCALmc*
- Dot Plot - *https://youtu.be/gdE46YSedvE*
- Histogram - *https://youtu.be/gSEYtAjuZ-Y*
- Bar Chart - *https://youtu.be/woUQ9LLaees?feature=shared*
- Scatter Plot - *https://youtu.be/sHbX58y5D4U?feature=shared*

Using appropriate technology to create displays

- Microsoft Excel – Box Plot - *https://youtu.be/39lsUsJsc2c*
- CODAP – Dot Plot - *https://codap.concord.org/help/basics/graphs*
- Google Sheets – Histogram - *https://youtu.be/BABC1jktDIc*

Several charts that are designed to encode multivariate data in a single visualization, which may include, but are not limited to, scatter plots, bubble charts, heatmaps, treemaps, and radar/spider chart.

Background Knowledge for Educators

CODAP - *https://codap.concord.org/*
Code.org - *App Lab* – Data Visualizer Microsoft Excel
Google Sheets
Tableau - *https://www.tableau.com/academic/teaching*
EduBlocks (sample coding tool) - *https://edublocks.org/*

Domain	**Visualize Data**	
Cluster	**Display data.**	
Standard(s)	M.DSHS.11	Use appropriate tools and multiple representations to represent and model relationships of quantitative multivariable data consisting of at least four variables.

Vocabulary

- **Quantitative data:** the value of data in the form of counts or numbers where each data set has a unique numerical value.
- **Scatterplot:** shows the relationship between two quantitative variables measured for the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as a point on the graph.
- **Variable:** in statistics, a variable is a characteristic of interest that you measure, record, and analyze.
  - An explanatory variable is what you manipulate.
  - A response variable is what changes as a result.

Content Examples and Resources

Midges Activity
https://ncssm.instructure.com/courses/1067/files/197631/download?wrap=1

Use AI to generate a multivariable task specific to your community
Sample AI generated task

Background Knowledge for Educators

CODAP - https://codap.concord.org/
Code.org - App Lab – Data Visualizer - https://studio.code.org/s/explore-data-1-2021/lessons/1
Computational and Inferential Thinking: The Foundations of Data Science (Chapter 15. Prediction and Chapter 17. Classification) - *Computational and Inferential Thinking: The Foundations of Data Science — Computational and Inferential Thinking*
Data Moves - https://escholarship.org/uc/item/0mg8m7g6
Exploring two-variable quantitative data - https://www.khanacademy.org/math/ap-statistics/bivariate-data-ap
Google Sheets
Microsoft Excel
Numbers
Programming languages (Python, Pyret - https://code.pyret.org/, the datascience library - https://www.data8.org/datascience/ also makes working with data and visualizations accessible in Python)

Domain	**Visualize Data**	
Cluster	**Display data.**	
Standard(s)	M.DSHS.12	Describe visual patterns in quantitative data such as clustering, outliers, positive or negative association, linear association, and nonlinear association (e.g., determine form, strength, and direction).

Vocabulary

- **Clustering:** high concentrations of data in a given set.
- **Data processing:** the steps taken to clean, transform, and prepare raw data for analysis or modeling.
- **Linear association:** a statistical term used to describe a straight-line relationship between two variables.
- **Negative association:** an association trending downward.
- **Nonlinear association:** an association between two variables in which the direction and rate of change fluctuate.
- **Outliers:** a data point that lies outside the overall pattern in a distribution.
- **Positive association:** an association trending upward.
- **Quantitative data:** the value of data in the form of counts or numbers where each data set has a unique numerical value.

Content Examples and Resources

Blue Whales

https://www.youcubed.org/tasks/blue-whales/

Data Visualization

https://www.youcubed.org/resources/dataviz/

Something Fishy

https://www.statisticsteacher.org/2021/03/29/nonlinear-modeling-something-fishy/

Background Knowledge for Educators

Interpreting Scatter Plots

https://youtu.be/Jpbm5YgciqI

Shapes of Distribution

https://youtu.be/2oJldeE4JcU

Domain	**Visualize Data**	
Cluster	**Display data.**	
Standard(s)	M.DSHS.13	Visualize categorical data using appropriate models such as mosaic plots, stacked bar graphs, etc. Recognize possible associations and data trends.

Vocabulary

- **Categorical data:** data that can be divided into specific groups, such as favorite color, age group, type of food, sport, etc.
- **Mosaic plots:** a mosaic plot is a special type of stacked bar chart that shows percentages of data in groups.
- **Stacked bar graph:** a form of bar chart that shows the composition and comparison of a few variables, either relative or absolute, over time.

Content Examples and Resources

Data Talk – House, Education, Wages
https://www.youcubed.org/wp-content/uploads/2021/10/Housing-Education-Wages-Data-Talk.pdf

Skew the Script
Describing Categorical Data Mosaic Plots
https://www.youtube.com/watch?v=rqte4YQ0mKA

AP Statistics Lessons
https://skewthescript.org/ap-stats-curriculum

Background Knowledge for Educators

CODAP
https://codap.concord.org/

Two-way tables
https://www.khanacademy.org/math/ap-statistics/analyzing-categorical-ap#stats-two-way-tables

Domain	**Visualize Data**	
Cluster	**Display data.**	
Standard(s)	M.DSHS.14	Use methods of geospatial analysis to graphically or spatially represent natural phenomena.

Vocabulary

- **Modeling:** the process of creating a mathematical representation of a real-world scenario to make a prediction or provide insight.
- **Spatial analysis:** the process of studying entities by examining, assessing, evaluating, and modeling spatial data features such as locations, attributes, and their relationships that reveal the geometric or geographic properties of data, generally using a GIS (Geographic Information System).

Content Examples and Resources

Creating a Strong Foundation for Statistics with GIS
https://storymaps.arcgis.com/collections/abf49441153745b984a9c61d8c0e9682

The Language of Spatial Analysis E-book
https://www.esri.com/content/dam/esrisites/sitecore-archive/Files/Pdfs/library/books/the-language-of-spatial-analysis.pdf

What is GIS?
https://www.esri.com/content/dam/esrisites/sitecore-archive/Files/Pdfs/library/bestpractices/what-is-gis.pdf

Background Knowledge for Educators

Discover Patterns and Trends in Data using ArcGIS Online. - *https://learn.arcgis.com/en/paths/data-analysis/*
Global Navigation Satellite System for Agricultural Use - *https://www.youtube.com/watch?v=0951MdaqzxI&list=PLkV8CNVuB_rJWSFAHCUWeiYikRvnLZ6Js&index=5*
Ocean Animal Emergency - *NOVA Online | Teachers | Classroom Activity | Ocean Animal Emergency | PBS*
Simulations - *https://www.khanacademy.org/computing/ap-computer-science-principles/x2d2f703b37b450a3:simulations*
Spatial Analysis and Data Science: Big Data Analytics and Spatial Analysis - *https://www.esri.com/en-us/arcgis/products/spatial-analytics-data-science/capabilities/spatial-analysis*
Try ArcGIS Online - *https://learn.arcgis.com/en/paths/try-arcgis-online/*
West Virginia View - *https://wvview.org/*

Domain	**Visualize Data**	
Cluster	**Display data.**	
Standard(s)	M.DSHS.15	Understand the use of simulation to compare probabilities from a model to observed frequencies; explain possible sources of discrepancy.

Vocabulary

- **Simulation:** the process of imitating a real phenomenon with a set of mathematical formulas.
- **Bootstrapping:** simulation method that estimates the variability of a statistic by resampling with replacement from the original data.

Content Examples and Resources

Coding a Simulation – Outbreak Simulator: *Outbreak Simulator - Code.org*
https://studio.code.org/s/outbreak/lessons/1

Using simulation to make conclusions
https://www.khanacademy.org/math/ap-statistics/xfb5d8e68:inference-categorical-proportions/idea-significance-tests/v/estimating-p-value-from-simulation

https://www.khanacademy.org/math/ap-statistics/xfb5d8e68:inference-categorical-proportions/idea-significance-tests/a/p-value-conclusions

https://www.khanacademy.org/math/ap-statistics/xfb5d8e68:inference-categorical-proportions/idea-significance-tests/e/estimating-p-values-and-making-conclusions

https://www.khanacademy.org/computing/ap-computer-science-principles/x2d2f703b37b450a3:simulations

Background Knowledge for Educators

Understanding Probability
https://youtu.be/KFgvOQtH0Z0

StatsQuest
Bootstrapping Clearly Explained Main Ideas
https://youtu.be/Xz0x-8-cgaQ?feature=shared
Bootstrapping vs Traditional Statistics
https://youtu.be/wbGHLj5GAdE?feature=shared

Domain	**Analyze Data**	
Cluster	**Choose appropriate statistical values.**	
Standard(s)	M.DSHS.16	Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, standard deviation) of two or more different data sets.

Note: The intent of this course is to apply these concepts, not to calculate the values by hand.

Vocabulary

- **Five number summary:** a summary of the data including the minimum, first quartile, median (second quartile), third quartile, and maximum.
- **Interquartile range (IQR):** the difference between the third quartile and the first quartile in a data set, the middle 50% of the data set.
- **Left skewed:** a distribution that has the majority of its data values on the right with possible outliers on the left.
- **Mean:** the average of a data set. Found by adding up all numbers together and dividing the sum by the number of items in the data set.
- **Median:** a value or quantity at the midpoint of a frequency data of observed values or quantities.
- **Mode:** the most frequently occurring value in a data set.
- **Quartile:** the value that breaks data set up into 4 equal parts (25% each).
- **Range:** the difference between the minimum and maximum value.
- **Right skewed:** a distribution that has the majority of its data values on the left with possible outliers on the right.
- **Roughly symmetric:** a distribution that is roughly even on both sides.
- **Standard deviation:** the typical average distance that a value is from the mean in data set.

Content Examples and Resources

Describing and Comparing Data Sets (This lesson needs modified to have a data science lens.)
https://www.census.gov/programs-surveys/sis/activities/math/describing-and-comparing.html

- This resource can support modifying the above lesson: CODAP – MicroData Portal - https://codap.concord.org/app/static/dg/en/cert/index.html#shared=https%3A%2F%2Fcfm-shared.concord.org%2FnKKraPUHFL6uC6K7G18F%2Ffile.json

Background Knowledge for Educators

How to Compare Data Sets
https://study.com/skill/learn/how-to-compare-a-data-set-by-measures-of-center-variation-explanation.html

Domain	**Analyze Data**	
Cluster	**Choose appropriate statistical values.**	
Standard(s)	M.DSHS.17	Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).

Note: The intent of this course is to apply these concepts, not to calculate the values by hand.

Vocabulary

Shape:

- **Left Skewed:** a distribution that has the majority of its data values on the right with possible outliers on the left.
- **Right skewed:** a distribution that has the majority of its data values on the left with possible outliers on the right.
- **Roughly symmetric:** a distribution that is roughly even on both sides.

Center:

- **Mean:** the average of a data set. Found by adding up all numbers together and dividing the sum by the number of items in the data set.
- **Median:** a value or quantity at the midpoint of a frequency data of observed values or quantities.

Spread:

- **Interquartile range (IQR):** the difference between the third quartile and the first quartile in a data set, the middle 50% of the data set.
- **Outlier:** an outlier is an extreme value in a data set that is either much larger or much smaller than all the other values.
- **Range:** the difference between the minimum and maximum value.
- **Standard deviation:** the typical average distance that a value is from the mean in data set.

Background Knowledge for Educators

Emory Oxford College
https://math.oxford.emory.edu/site/math117/shapeCenterAndSpread/

Measures of center and spread
https://youtu.be/OZXMEp8ZKvI

Domain	**Analyze Data**	
Cluster	**Fit bivariate data to functions using regression.**	
Standard(s)	M.DSHS.18	Use technology to create a regression for data that suggests a linear association. Compute the correlation coefficient, coefficient of determination, and residual plot, and interpret the results in the context of the problem.
Note: The intent of this course is to apply these concepts, not to calculate the values by hand.		

Vocabulary

- **Coefficient of determination:** the coefficient of determination measures the percentage of variability within the y-values that can be explained by the regression model.
- **Correlation coefficient:** a measure of strength and direction of a linear relationship with values ranging between -1 and 1 with 1 denoting a strong positive relationship and negative 1 denoting a strong negative relationship.
- **Least squares regression line:** the line of best fit for a set of data.
- **Residual:** the difference between the actual data point and the predicted value.
- **Residual plot:** a plot of all the residuals from the distribution.

Content Examples and Resources

Coefficient of determination
https://youtu.be/lng4ZgConCM

Computer outputs for linear regression
https://youtu.be/sIJj7Q77SVI

Correlation coefficient
https://youtu.be/-Y-M9aD_ccQ

Midges Activity
https://ncssm.instructure.com/courses/1067/files/197631/download?wrap=1

Regression analysis
https://www.khanacademy.org/math/probability/xa88397b6:scatterplots

Residual plot and using residuals
https://youtu.be/VamMrPZ-8fc

Background Knowledge for Educators

Computational and Inferential Thinking: The Foundations of Data Science **(Chapter 15. Prediction)**
https://inferentialthinking.com/chapters/intro.html

Domain	**Analyze Data**	
Cluster	**Fit bivariate data to functions using regression.**	
Standard(s)	M.DSHS.19	Fit a function to the data that does not suggest a linear association; use algebraic re-expression of the function to fit the data to solve problems in the context of the situation.

Note: The intent of this course is to apply these concepts, not to calculate the values by hand.

Vocabulary

- **Non-linear regression:** a form of regression analysis in which data is fitted to a model that does not suggest a linear association.

Content Examples and Resources

Models fit to data
https://www.khanacademy.org/math/statistics-probability/advanced-regression-inference-transforming/nonlinear-regression/v/comparing-models-to-fit-data

Statistics: Ch 3 Bivariate Data (8 of 25) Linear vs Non-Linear Correlation
https://www.youtube.com/watch?v=g-fF4vto9ws

Background Knowledge for Educators

Decaying Skittles
https://teacher.desmos.com/activitybuilder/custom/59ad6c1963c6c3546e27be89

Domain	**Analyze Data**	
Cluster	**Fit bivariate data to functions using regression.**	
Standard(s)	M.DSHS.20	Interpret key features such as intercepts, rate of change, and turning points of models in context of the data.
Note: Values to be estimates from the graphs of regression models.		

Vocabulary

- **Explanatory variable:** the variable that is manipulated or observed.
- **Intercepts:** a point where the graph touches the x or y axis.
- **Rate of change:** the ratio of the change in the response variable compared to the change in the change in the explanatory.
- **Response variable:** the variable that changes as a result of the explanatory variable.
- **Turning points:** the place where a function changes from increase to decreasing or vice versa.

Content Examples and Resources

Interpret slope and y-intercept in context
https://www.youtube.com/watch?v=PplggM0KtJ8&feature=youtu.be

Background Knowledge for Educators

Regression and prediction
https://www.oreilly.com/library/view/practical-statistics-for/9781491952955/ch04.html

Computational and Inferential Thinking: The Foundations of Data Science **(Chapter 15. Prediction)**
https://inferentialthinking.com/chapters/intro.html

Domain	**Analyze Data**	
Cluster	**Understand the use of algorithms in statistical tests.**	
Standard(s)	M.DSHS.21	Examine existing algorithms and describe connections to algebraic and statistical functions, sets, and logic.

Vocabulary

- **Algorithm:** a process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer.

Content Examples and Resources

CODAP - Dynamic Data Science Activities
https://learn.concord.org/dynamic-data-science

NASA Data Stories
https://mynasadata.larc.nasa.gov/basic-page/earth-system-story-map-collections-lesson-plans

Domain	**Analyze Data**	
Cluster	**Understand the use of algorithms in statistical tests.**	
Standard(s)	M.DSHS.22	Develop algorithms in order to solve mathematical problems

Vocabulary

- **Algorithm:** a process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer.
- **Computational thinking:** breaking down problems into smaller parts and solving them logically.
- **Custom data wrangling:** create functions, formulas, or logic rules to organize, clean, or transform data—rather than relying only on built-in tools.
- **Mathematical problems:** these could include anything from basic arithmetic to modeling real-world data (like calculating growth rates, probability, or trends).

Background Knowledge for Educators

EduBlocks
https://app.edublocks.org/

Flowgorithm
http://www.flowgorithm.org/index.html

Domain	**Analyze Data**	
Cluster	**Understand probability in relation to decision-making.**	
Standard(s)	M.DSHS.23	Use the concepts of independent events and conditional probabilities to calculate and interpret outcomes of chance events to make data-informed decisions. Recognize and explain the concepts of conditional probability and independence to multiple audiences and contexts.

Note: This course is not designed to test for independence.

Vocabulary

- **Conditional Probability:** the probability of an event out of a subset of the population instead of out of the whole.
- **Independent Events:** two events are independent if one event occurring does not change the probability of the other event occurring.

Content Examples and Resources

Expected Value
https://www.khanacademy.org/math/probability/xa88397b6:probability-distributions-expected-value

Probability
https://www.khanacademy.org/math/probability/xa88397b6:probability

Background Knowledge for Educators

Conditional probabilities and independent events
https://portal.mathmedic.com/lesson-plans/course/Intro-Stats/unit/5/day/4

Two-way tables and Venn diagrams
https://portal.mathmedic.com/lesson-plans/course/Intro-Stats/unit/5/day/3

Domain	**Analyze Data**	
Cluster	**Understand probability in relation to decision-making.**	
Standard(s)	M.DSHS.24	Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages, using the area beneath the curve to make estimations of frequencies. ***Instructional Note:*** *Emphasize that only some data are well described by a normal distribution.*

Note: The intent of this course is to apply these concepts, not to calculate the values by hand.

Vocabulary

- **Mean:** the average of a data set. Found by adding up all numbers together and dividing the sum by the number of numbers.
- **Standard deviation:** the typical average distance that a value is from the mean in data set.
- **z-score:** the number of standard deviations a value is from the mean.

Content Examples and Resources

Define normal distributions
https://www.khanacademy.org/math/statistics-probability/modeling-distributions-of-data/normal-distributions-library/a/normal-distributions-review

Empirical Rule and Normal distribution calculations
https://www.khanacademy.org/math/probability/xa88397b6:analyze-quantitative (Last 2 sections)

Background Knowledge for Educators

The Normal Distribution, Clearly Explained!!!
https://www.youtube.com/watch?v=rzFX5NWojp0

Normal distribution activity
https://www.census.gov/programs-surveys/sis/activities/math/the-new-normal.html

Domain	**Communicate Using Data**	
Cluster	**Compare distributions.**	
Standard(s)	M.DSHS.25	Assess the degree of visual overlap of two numerical data distributions with similar variabilities, measuring the difference between the centers by expressing it as a multiple of a measure of variability (e.g., The mean height of players on the basketball team is 10 cm greater than the mean height of players on the soccer team, about twice the variability on either team; on a dot plot, the separation between the two distributions of heights is noticeable).

Vocabulary

- **Mean:** the value obtained by dividing the sum of a set of quantities by the number of the quantities in the set.
- **Range:** the difference between the lowest and highest values.
- **Spread:** the extent to which values in a dataset vary, either from one another or from the center.
- **Standard deviation:** the typical average distance that a value is from the mean in data set.
- **Variability:** describes how far apart data points lie from each other and from the center of a distribution.
- **Variance:** the average of the squared differences from the mean.

Content Examples and Resources

Comparing Distributions
https://youtu.be/lT3LRjtSbJc

Measures of Spread
https://youtu.be/zjmyBk9eJ-c

Measures of Variability (Range, Standard Deviation, Variance)
https://youtu.be/s7WTQ0H0Acc

Background Knowledge for Educators

Variability| Calculating Range, IQR, Variance, Standard Deviation
https://www.scribbr.com/statistics/variability/

Domain	**Communicate Using Data**	
Cluster	**Compare distributions.**	
Standard(s)	M.DSHS.26	Analyze and communicate the benefits and limitations of data visualization tools to solve a real-world problem.

Vocabulary

- **Data visualization:** representation of information in the form of a chart, diagram, picture, etc.
- **Infographic:** collection of imagery, data visualizations like pie charts and bar graphs, and minimal text that gives you an easy-to-understand overview of a topic.

Content Examples and Resources

Crash Course: Data & Infographics
https://youtu.be/OiND50qfCek

Types of Data Visualizations
https://youtu.be/umfpuYKtlK0

What's going on in this graph: NY Times
https://www.nytimes.com/column/whats-going-on-in-this-graph

Background Knowledge for Educators

Canva free through login with Office 365 account

Story Maps (*https://storymaps.arcgis.com/***) free through login with Office365 at** *WV Student Maps (arcgis.com)*

Domain	**Communicate Using Data**	
Cluster	**Evaluate claims.**	
Standard(s)	M.DSHS.27	Distinguish between correlation and causation. ***Instructional Note:*** *The important distinction between a statistical relationship and a cause-and-effect relationship is the focus.*

Vocabulary

- **Causation:** one event causes another event to occur.
- **Correlation:** describes the relationship or pattern between the values of two variables.
- **Negative correlation:** as one set of values increases, the other set tends to decrease.
- **Positive correlation:** as one set of values increases, the other set tends to increase.
- **Zero correlation:** if the change in values of one set doesn't affect the values of the other.

Content Examples and Resources

Correlation vs Causation
https://youtu.be/rzfEj8jjs24

Background Knowledge for Educators

Correlation does not imply Causation
https://www.statology.org/correlation-does-not-imply-causation-examples/

Correlation vs Causation
https://www.scribbr.com/methodology/correlation-vs-causation/

Data Demystified: Correlation vs. Causation
https://www.datacamp.com/blog/data-demystified-correlation-vs-causation

Domain	**Communicate Using Data**	
Cluster	**Evaluate claims.**	
Standard(s)	M.DSHS.28	Evaluate claims based on data reports gathered from a variety of sources such as the media, scientific journals, census data, etc.

Vocabulary

- **Claim:** state or assert that something is the case, typically without providing evidence or proof.
- **Credible source:** written by someone who is an expert in their discipline and is free of errors.
- **Peer-reviewed:** the peer-review process subjects an author's scholarly work, research, or ideas to the scrutiny of others who are experts in the same field (peers) and is considered necessary to ensure academic scientific quality.
- **Scholarly:** written by academics and other experts and contribute to knowledge in a particular field by sharing new research findings, theories, analyses, insights, news, or summaries of current knowledge. Scholarly sources can be either primary or secondary research.

Content Examples and Resources

Evaluating Evidence
https://youtu.be/hxhbOvR2TGk

Background Knowledge for Educators

Evaluating Claims you find on Websites
https://libguides.pima.edu/credible/claim

Evaluating Sources
https://researchguides.library.brocku.ca/external-analysis/evaluating-sources

Is my Source Credible?
https://libguides.umgc.edu/credibility

Domain	Communicate Using Data	
Cluster	**Report conclusions in multiple formats.**	
Standard(s)	M.DSHS.29	Report results using an appropriate format (digital presentation, verbal, textual, etc.) and to a particular audience using the relevant language of mathematics and data science. Use data displays and interpret results in terms of the question studied.

Vocabulary

- **Data storytelling:** the ability to effectively communicate insights from a dataset using narrative and visualizations.
- **Infographic:** a visual image such as a chart or diagram used to represent information or data.

Content Examples and Resources

Code.org – Project to Tell a Data Story Parts 1 and 2
https://studio.code.org/s/csp9-2022/lessons/8
https://studio.code.org/s/csp9-2022/lessons/9

Background Knowledge for Educators

How to Tell a Story with Data
https://hbr.org/2013/04/how-to-tell-a-story-with-data

How to Tell a Story with Data
https://www.lucidchart.com/blog/how-to-tell-a-story-with-data

Storytelling with Data website- request access for collaboration
https://www.storytellingwithdata.com/blog/teaching-data-storytelling

Data Visualizations, Charts, and Graphs
https://accessibility.huit.harvard.edu/data-viz-charts-graphs

Write Helpful Alt Test to Describe Images
https://accessibility.huit.harvard.edu/describe-content-images

Domain	**Application of Data Science**	
Cluster	**Understand security and ethics.**	
Standard(s)	M.DSHS.30	Explore various legal and ethical standards for data ownership and the implications of the standards to the study and application of data science.

Vocabulary

- **Anonymization:** removing personal details from data so that individuals cannot be identified, even if the dataset is shared or analyzed.
- **Data ownership:** the possession of and responsibility for information.
- **Data masking:** hiding real data by replacing it with fake but realistic-looking information.
- **Personally Identifiable Information (PII):** information about an individual that identifies, links, relates, or describes them.
- **Secure storage:** keeping data safe using methods like encryption, passwords, and restricted access, so only the right people can see it.

Content Examples and Resources

YouCubed Data Science- Data Privacy and Ethics
https://hsdatascience.youcubed.org/section/unit-1-section-4/#activity3

Background Knowledge for Educators

Code.org – Data Policies and Privacy
https://studio.code.org/s/csp10-2022/lessons/3

Domain	**Application of Data Science**	
Cluster	**Understand security and ethics.**	
Standard(s)	M.DSHS.31	Describe and understand how data is collected from both individuals and groups of individuals, shared, and used.

Vocabulary

- **Big data:** collections of data that meet the 3 V's criteria: Volume, Variation and Velocity. The volume of data is "so much" that it requires tools and techniques to handle and process. Big data is "so mixed" and potentially from a variety of sources. The rate at which big data can accumulate can be "so fast" it exceeds the ability of traditional techniques and applications (Excel, Sheets, etc.).
- **Data mining:** the process of finding or analyzing datasets to discover new patterns that might improve the model.
- **Data set:** collection of information that can be numeric or character based.
- **Training data:** all the data used during the process of training a machine learning algorithm, as well as the specific dataset used for training rather than testing.

Background Knowledge for Educators

Data Mining
https://www.elephango.com/index.cfm/pg/k12learning/lcid/13228/Data_Is_Mined!

How is Big Data Collected?
https://codeit.us/blog/how-is-big-data-collected

How is Big Data Collected by Companies?
https://computools.com/how-is-big-data-collected/

Domain	**Application of Data Science**	
Cluster	**Explore artificial intelligence.**	
Standard(s)	M.DSHS.32	Know and identify examples of real-world or societal machine learning applications.

Vocabulary

- **Deep learning:** a type of machine learning where computers learn to recognize patterns using layers of processing—kind of like how the human brain works. It's especially good at working with images, sound, and large amounts of data.
- **Machine learning:** using pattern recognition software to find trends in data, building models that explain the trends/patterns, and then using the models to predict something.
- **Neural network:** a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain.
- **Training data:** all the data used during the process of training a machine learning algorithm, as well as the specific dataset used for training rather than testing.
- **Test data set:** unlabeled data used to check that a machine learning model can perform its assigned task.
- **Validation data set:** used to test a recently trained model against new data and to analyze performance, with a particular focus on checking for overfitting

Content Examples and Resources

AI Oceans Activity
https://hourofcode.com/ai-oceans

Machine Learning and Bias
https://www.khanacademy.org/computing/

Background Knowledge for Educators

9 Examples of Machine Learning in Action
https://www.codecademy.com/resources/blog/machine-learning-examples/

Code.org – AI
https://code.org/ai

Machine Learning, explained
https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

What is artificial intelligence (AI)?
https://www.ibm.com/topics/artificial-intelligence

LLMS Briefly Explained 3Blue1Brown (The 3Blue1Brown YouTube Channel has several videos on neural networks and LLMs)
https://youtu.be/LPZh9BOjkQs?feature=shared

What is machine learning (ML)?
https://www.ibm.com/topics/machine-learning

Domain	**Application of Data Science**	
Cluster	**Explore artificial intelligence.**	
Standard(s)	M.DSHS.33	Describe basic machine learning concepts such as training a model and evaluating model performance

Vocabulary

- **Generative Adversarial Network (GAN):** has two parts:
  - The generator learns to generate plausible data. The generated instances become negative training examples for the discriminator.
  - The discriminator learns to distinguish the generator's fake data from real data. The discriminator penalizes the generator for producing implausible results.

    When training begins, the generator produces obviously fake data, and the discriminator quickly learns to tell that it's fake.
- **Neural network:** a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain.
- **Reinforcement learning (RL):** is the science of decision making. It is about learning the optimal behavior in an environment to obtain maximum reward.
- **Self-learning:** is artificial intelligence that can train itself using unlabeled data. On a high level, it works by analyzing a dataset and looking for patterns that it can draw conclusions from. It essentially learns to "fill in the blanks."
- **Supervised learning:** also known as supervised machine learning, is a subcategory of machine learning and artificial intelligence. It is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately.
- **Training a model:** a machine learning training model is a process in which a machine learning (ML) algorithm is fed with sufficient training data to learn from.
- **Unsupervised learning:** also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabeled datasets.

Content Examples and Resources

Machine Learning and Bias
https://www.khanacademy.org/computing/ap-computer-science-principles/data-analysis-101#x2d2f703b37b450a3:machine-learning-and-bias

Background Knowledge for Educators

Code.org - AI
https://code.org/ai

Generative Adversarial Networks (GANS):
https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/
https://developers.google.com/machine-learning/gan/gan_structure

Machine Learning algorithms train on large data to update an algebraic function and make it fit data better automatically.

- If they're doing this, students should learn the basic math mechanics underneath. Ideally, teachers would introduce the basics of what's going on underneath (a weighted average that's optimized, intuition about calculus and optimization problems, etc.).

Neural Network Resources:
https://www.ibm.com/think/topics/neural-networks
https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414
https://youtu.be/CqOfi41LfDw?feature=shared

Reinforcement Learning:
https://www.synopsys.com/ai/what-is-reinforcement-learning.html
https://youtu.be/Z-T0iJEXiwM?feature=shared

Self-Learning:
https://www.udacity.com/blog/2021/08/self-learning-ai-explained.html

Supervised and Unsupervised Learning:
https://www.ibm.com/topics/supervised-learning
https://www.simplilearn.com/tutorials/machine-learning-tutorial/supervised-and-unsupervised-learning
https://www.ibm.com/think/topics/unsupervised-learning

Training a model:
https://developers.google.com/machine-learning/crash-course/linear-regression/loss
https://oden.io/glossary/model-training/

LLMS Briefly Explained:
3Blue1Brown (The 3Blue1Brown YouTube Channel has several videos on neural networks and LLMs)
https://youtu.be/LPZh9BOjkQs?feature=shared

Domain	**Application of Data Science**	
Cluster	**Explore artificial intelligence.**	
Standard(s)	M.DSHS.34	Know and identify examples of natural language processing and its connection to mathematics and probability.

Vocabulary

- **Natural language processing:** refers to the branch of computer science—and more specifically, the branch of *artificial intelligence or AI*—concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.

Background Knowledge for Educators

8 NLP Examples
https://www.tableau.com/learn/articles/natural-language-processing-examples

Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review
https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13370

Natural Language Processing
https://www.ibm.com/topics/natural-language-processing

Domain	Application of Data Science	
Cluster	Explore artificial intelligence.	
Standard(s)	M.DSHS.35	Review ethical issues and the impact of machine learning and natural language processing.

Vocabulary

- **Machine learning:** using pattern recognition software to find trends in data, building models that explain the trends/patterns, and then using the models to predict something.
- **Natural language processing:** refers to the branch of computer science—and more specifically, the branch of *artificial intelligence or AI*—concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.

Content Examples and Resources

Code.org - AI
https://code.org/ai

Code.org – Our AI Code of Ethics
https://studio.code.org/s/ai-ethics-2021/lessons/1

Background Knowledge for Educators

Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward
https://www.nature.com/articles/s41599-020-0501-9

Ethics in NLP
https://dida.do/blog/ethics-in-natural-language-processing

Domain	**Application of Data Science**	
Cluster	**Utilize a coding language.**	
Standard(s)	M.DSHS.36	Evaluate the appropriateness of programming languages and applications as they relate to data science.

Vocabulary

- **Programming language:** system of notation for writing computer programs.
- **Python:** a very popular programming language that is versatile and diverse in its uses. Python has several very good libraries of procedures that are for data handling.
- **R:** programming language popular with statisticians and analysts. Developed specifically for statistics and data processing.
- **SQL:** a programming language made to work with multiple databases and data sets.

Background Knowledge for Educators

5 Python Libraries You Need for Data Science- CodeEmporium
https://youtu.be/9E-lsou160g

Collection of videos related to R and Python used for data processing
https://www.youtube.com/@aaronmaxwell7222/videos

Introduction to SQL
https://www.w3schools.com/sql/sql_intro.asp

Intro to SQL: Querying and Managing Data
https://www.khanacademy.org/computing/computer-programming/sql

Python- MySQL
https://www.w3schools.com/python/python_mysql_getstarted.asp

Python vs SQL Comparison
https://streamsets.com/blog/python-vs-sql/#:~:text=Python%20and%20SQL%20are%20popular,used%20to%20communicate%20with%20databases

R Tutorial
https://www.w3schools.com/r/

SQL You Need as a Data Science- CodeEmporium
https://youtu.be/CM2zjiud5PI

Domain	**Application of Data Science**	
Cluster	**Utilize a coding language.**	
Standard(s)	M.DSHS.37	Select a programming language to explore, display, and analyze data.

Vocabulary

- **Ascending order:** arranging numeric data from least to greatest. Arranging string data by Ascii table values, character by character, least value to greatest (capital letters come before lower case letters, hence they have a smaller value), etc.
- **Descending order:** arranging numeric data from greatest to least. Arranging string data by Ascii table values, character by character, greatest value to least, etc.

Background Knowledge for Educators

Matplotlib Tutorial
https://www.w3schools.com/python/matplotlib_intro.asp

max() and min() in Python
https://www.geeksforgeeks.org/max-min-python/

Replit- Python Starter For Plotting Data
https://replit.com/@wgibson/PythonStarterForPlottingData#main.py

Sorting A List Alphabetically in Python with sort()
https://learnpython.com/blog/sort-alphabetically-in-python/#:~:text=In%20Python%2C%20sorting%20a%20list,their%20first%20letter%20(A%2DZ)

Additional Resources
ASCII Table
Pandas tutorial
10 Python data visualization libraries

Domain	**Application of Data Science**	
Cluster	**Utilize a coding language.**	
Standard(s)	M.DSHS.38	Identify types of information that can be stored as variables, classify variables, and utilize variables in programs that store data in appropriate ways (e.g., Booleans, characters, integers, floating points, strings).

Vocabulary

- **Boolean:** a data type that is related to the truth of an expression. It usually stores a 1 for true or a 0 for false.
- **Characters:** a data type consisting of a single letter, number or keyboard character. They are stored as an integer value identified on the ASCII or Unicode tables. Python does not use the character type.
- **Floating point:** a decimal or fractional number. In Python it is referred to as a float.
- **Integer:** a whole number, positive or negative, without decimals, of limited length. In Python that limit is 263 - 1 (the largest number that can be stored in 64 bits of memory).
- **String:** a collection of 1 or more characters. Strings can consist of letters, numbers, words and keyboard characters. Most programming languages treat Strings as a list or array of individual characters.
- **Variable:** a name for a pointer to a location in memory in which data is stored. It can be a single value such as an integer or it can be to an object that is storing multiple types of data.

Content Examples and Resources

Assigning Variables
https://www.khanacademy.org/computing/ap-computer-science-principles/programming-101/storing-variables/a/assigning-variables

Programming Mathematical Expressions
https://www.khanacademy.org/computing/ap-computer-science-principles/programming-101/numbers-and-math/a/programming-mathematical-expressions

Python Data Types
Python data types

String Variables
https://www.khanacademy.org/computing/ap-computer-science-principles/programming-101/strings/a/storing-strings-in-variables

Background Knowledge for Educators

Fostering Better Coding Practices for Data Scientists
https://hdsr.mitpress.mit.edu/pub/8wsiqh1c/release/4

Coding Best Practices – variable naming
Coding best practices

Python Variables Practice
Python Variables Practice

Domain	**Application of Data Science**	
Cluster	**Utilize a coding language.**	
Standard(s)	M.DSHS.39	Interpret relational and logical expressions of level-appropriate complexity using comparison and Boolean operators.

Vocabulary

- **DeMorgan's Law:**
  - The complement of the union of two sets is the same as the intersection of their complements or not (A or B) = (not A) and (not B). In code: !(A||B) == !A && !B
  - The complement of the intersection of two sets is the same as the union of their complements or not (A and B) = (not A) or (not B). In code: !(A&&B) == !A || !B
- **Logical "and":** a Boolean operator that requires both operands to be true in order for the logical expression to evaluate to true. If only one operand is true or if neither is true, the expression will evaluate to false. The "and" operator is often coded as "&&"in many programming languages.
- **Logical "not":** a Boolean operator that causes the expression to evaluate to the opposite value. If the original expression is true, the not would be false. The "not" operator is often coded as "!" in many programming languages.
- **Logical "or":** a Boolean operator that requires only one of the operands to be true in order for the logical expression to evaluate to true. If both operands are false, the expression will evaluate to false. The "or" operator is often coded as "||" in many programming languages.
- **Truth table:** a truth table is a tabular representation of all the combinations of Boolean values for inputs and their corresponding outputs.

Content Examples and Resources

Conditionals
https://www.khanacademy.org/computing/ap-computer-science-principles/programming-101/boolean-logic/a/conditionals-with-if-else-and-booleans

Logic Exercises for Python
Codingbat Logic Exercises for Python

Background Knowledge for Educators

Other logical operators (think of these as asking a question whose answer is either true or false): < (is less than), > (is greater than), <= (is less than or equal to), >= (is greater than or equal to), == (is equal to), != (is not equal to)

Domain	**Application of Data Science**	
Cluster	**Utilize a coding language.**	
Standard(s)	M.DSHS.40	Create programming solutions by reusing existing code to perform analysis or retrieve data (e.g., libraries, APIs, publicly shared code).

Vocabulary

- **Application Programming Interface (API):** generally, it is a set of defined rules that allow programs to "talk" to each other. Programming languages use an API to help create applications that can run on various hardware.
- **Forking:** using an exact copy someone else's code as a starting point to allow the developer to change, add to, or otherwise augment the original code to same time and effort. This is a common practice in open source projects. A common platform for forking projects is Git via GitHub.
- **Inheritance:** an inherent property in object-oriented programming that allows for the reuse of code by programs (classes) that inherit the properties of the "parent" class. For example, an Animal object may be defined to have the properties of "age", "size", and "name" as well as the behaviors of "breathing", "reproducing", and "eating". A Bear object would inherit these and add more that better defines a bear. A Dog object would also inherit the same from Animal and add its own properties or behaviors.
- **Library:** a collection of prewritten code that can be used by the programmer to complete a task efficiently. It saves time and effort as new code need not be recreated with every program.

Background Knowledge for Educators

How to Fork a Repository
https://docs.github.com/en/get-started/quickstart/fork-a-repo

How to Import modules in Python
https://www.digitalocean.com/community/tutorials/how-to-import-modules-in-python-3

Inheritance
https://www.w3schools.com/python/python_inheritance.asp

Inheritance in Python
https://www.geeksforgeeks.org/inheritance-in-python/

Python Modules Tutorial
https://docs.python.org/3/tutorial/modules.html

R notes and Examples
https://www.tutorialspoint.com/r/r_packages.htm

The Most Popular Python Packages
https://learnpython.com/blog/most-popular-python-packages/

The Top R Libraries for Data Science
https://www.knowledgehut.com/blog/data-science/top-r-libraries-for-data-science

Domain	**Application of Data Science**	
Cluster	**Utilize a coding language.**	
Standard(s)	M.DSHS.41	Write code or functions that can programmatically manipulate data sets (e.g., slice, merge, subset, sort, fit, summarize, analyze).

Vocabulary

- **Algorithm:** a step-by-step of commands that produce a desired outcome.
- **Iteration:** the process of repeating code for a set number of times or until a condition has been met.
- **Append:** add contents to a list at the end or beginning (depending on the language used).
- **Conditionals:** statements that control the flow of a program based on the truth of a given parameter.
- **Insert:** add contents to a list at a chosen location.

Content Examples and Resources

Algorithms
https://www.khanacademy.org/computing/computer-science/algorithms

Programming
https://www.khanacademy.org/computing/ap-computer-science-principles/programming-101

Background Knowledge for Educators

How to Write Pseudocode
https://www.geeksforgeeks.org/how-to-write-a-pseudo-code/

Machine Learning Algorithms for Python
https://data-flair.training/blogs/machine-learning-algorithms-in-python/

Purdue Algorithm Notes
https://www.cs.purdue.edu/cgvlab/courses/177/Spring2012/Algo0Notes.pdf

Data Moves
https://escholarship.org/uc/item/0mg8m7g6

Domain	**Application of Data Science**	
Cluster	**Apply data science to a capstone project.**	
Standard(s)	M.DSHS.42	Choose a problem or issue of interest. Throughout the program of study, research and use existing data set(s) to explore, visualize, analyze, and communicate findings to tell a data story.

Vocabulary

- **Capstone Project:** a multifaceted body of work that serves as a culminating academic and intellectual experience for students.
- **Data Storytelling:** the process of transforming data analyses into an understandable storyline for a wider audience.

Background Knowledge for Educators

Note: These resources are meant to be examples of problems that can be used as a Capstone but are not meant to be prescriptive.

12 Data Science Projects to Try
https://www.springboard.com/blog/data-science/data-science-projects/

Data Science Capstone Projects
https://samchaaa.medium.com/data-science-capstone-ideas-and-how-to-get-started-46d607194ce2

Appendix

AI Generated Task for M.DS.11, which includes other standards in the Data Science course.

Exploratory Data Analysis (EDA) on Housing Prices in West Virginia

Objective:

Students will conduct exploratory data analysis on a housing dataset for West Virginia, focusing on understanding the relationships between various variables (e.g., house size, location, number of bedrooms) and how they influence house prices. The goal is to develop insights into regional differences, economic factors, and potential investment opportunities.

Dataset:

Use a West Virginia housing dataset (or a similar dataset if unavailable) that includes features like:

- **Price:** The selling price of the house.
- **Number of Bedrooms:** The count of bedrooms in each house.
- **Square Footage:** The total area of the house.
- **Location:** The region of West Virginia, such as urban (Charleston, Morgantown) or rural areas (Appalachian regions, small towns).
- **Year Built:** The year the house was constructed.
- **Distance to Major Cities:** Proximity to Charleston or Morgantown.
- **Proximity to Natural Features:** Distance to mountains, rivers, or national forests.
- **Economic Factors:** Variables such as unemployment rate or coal mining presence in the area.
- **Condition:** Age and upkeep of the house (e.g., renovated vs. original).
- **Property Tax Rate:** The local tax rate on the property.
- **Neighborhood Type:** Urban, suburban, rural, or coal mining town.

Steps:

1. **Introduction (15 minutes)**
   - Dataset Introduction:
     - The housing data for West Virginia can reveal trends in rural vs. urban real estate markets, with emphasis on price variations across different regions and proximity to natural resources or economic hubs.
     - Objective: To understand the relationship between housing prices and key features such as square footage, number of bedrooms, and geographic location (e.g., proximity to coal mines, rural/urban divide).
   - EDA Concepts:
     - Correlation: Determine if higher square footage or more bedrooms generally increase house prices.
     - Outliers: Identify if any properties, especially in certain regions (e.g., coal mining areas), are priced much higher or lower than expected.
     - Distributions: Visualize how house prices are distributed across urban and rural areas.

2. **Data Preparation (30 minutes)**
   - Load the Data
   - Check for Missing Values
   - Basic Statistical Summary
   - Data Cleaning:
     - Handle missing values, for example, by filling in missing values for “year built” with the median or dropping rows with excessive missing data.
     - Convert categorical variables like “neighborhood type” into a more analyzable form, using encoding if necessary.

3. **Exploratory Analysis (45 minutes)**
   - Scatter Plots: Visualize relationships between housing prices and relevant features.
     - Price vs. Square Footage
     - Price vs. Number of Bedrooms
     - Price vs. Proximity to Mountains or Rivers
   - Box Plots
     - Price Distribution by Region: Analyze how prices differ between urban (Charleston, Morgantown) and rural (Appalachian, coal mining areas) neighborhoods.
     - Price by Year Built
   - Correlation Matrix:
     - Visualize correlations between key variables (e.g., price, square footage, number of bedrooms, proximity to key areas).

4. **Interpretation (30 minutes)**
   - Key Insights from Visualizations:
     - Does the square footage correlate with price across urban and rural regions?
     - How do prices vary between houses near natural features (e.g., mountains, rivers) versus those further away?
     - Do older houses (built in the early 1900s) have a significant price difference compared to newer houses?
   - Notable Correlations:
     - Identify key relationships, such as a potential correlation between proximity to economic hubs (Charleston, Morgantown) and higher prices.
   - Outliers:
     - Investigate any outliers (e.g., a very high-priced house in a small coal mining town) and hypothesize why they might exist.
   - Emerging Questions:
     - Why might houses in coal mining regions be priced lower despite being relatively large?
     - How does proximity to nature or national parks affect house prices in rural West Virginia?

5. **Group Discussion (30 minutes)**

- Sharing Findings: Students will present their analysis on key factors that influence housing prices in different regions of West Virginia.
  - Example questions to discuss: “What factors contribute to higher prices in Morgantown compared to rural areas?” or “Why are houses near national parks priced higher?”
- Implications for Buyers and Investors:
  - Discuss how buyers could use this data to make informed decisions based on location, house size, and the age of the property.
  - How might real estate investors use this information to predict future housing market trends in rural vs. urban areas?

6. **Presentation (Optional, 30 minutes)**

If time permits, students can prepare a 5-minute presentation highlighting:

- Visualizations and key insights.
- Specific regional findings about housing prices in West Virginia.
- Implications for real estate investments or decisions for homebuyers.

Assessment Criteria:

- Completeness: Did the student explore the relationship between housing price and multiple factors, including square footage, location, and economic conditions?
- Clarity: Were the insights easy to understand and well-supported by visuals?
- Engagement: How actively did the student engage in the group discussion and respond to questions?
- Technical Proficiency: Did the student effectively use Python (or another tool) to conduct the analysis and produce meaningful results?

Tools and Libraries:

- Google Colab: Free cloud-based Jupyter notebooks for coding, demos, notebooks.
- Pandas: Data manipulation and preparation.
- Matplotlib & Seaborn: For visualization (scatter plots, box plots, heatmaps).
- Jupyter Notebook or Python IDE: For code execution and visualization.

This customized version focuses on West Virginia’s unique regional characteristics and would help students understand the local dynamics of housing markets, focusing on factors like rural vs. urban differences, economic conditions, and proximity to natural features like mountains.

Michele L. Blatt
West Virginia Superintendent of Schools

wvde.us