

Livermore Valley Joint Unified School District

Course Title: Explorations in Data Science A/B
Grade Level(s): 11-12
Length of Course: Two semesters or equivalent term
Credit: 10 units
Prerequisite: Algebra I A/B and Geometry A/B

Course Overview:

This curriculum will introduce students to the main ideas in data science through free tools such as Google Sheets, Python, Data Commons and Tableau. Students will learn to be data explorers in project-based units, through which they will develop their understanding of data analysis, sampling, correlation/causation, bias and uncertainty, probability, modeling with data, making and evaluating data-based arguments, the power of data in society, and more! At the end of the course students will have a portfolio of their data science work to showcase their newly developed abilities.

Schools Offering: Del Valle High School
Granada High School
Livermore High School
Vineyard High School

Meets University of California Entrance Requirements: Seeking "c" approval

Board Approval: Pending Board Approval

Course Materials: Students will use a variety of no-cost programs and applications via their ChromeBooks including Google Sheets, CODAP, Kaggle, Edublocks, and Tableau. Additional resources will be utilizing District approved software.

Supplemental Materials: None

Explorations in Data Science

COURSE CONTENT:

In this course students will learn to understand, ask questions of, and represent data through project-based units. The units will give students opportunities to be data explorers through active engagement, developing their understanding of data analysis, sampling, correlation/causation, bias and uncertainty, modeling with data, making and evaluating data-based arguments, and the importance of data in society. At the end of the course, students will have a portfolio of their data science work to showcase their newly developed knowledge and understanding. The curriculum will be adaptable so that teachers can either use the data sets provided or bring in data sets most relevant to their own students. This course can be taken as an alternative to Algebra 2, or in addition to Algebra 2.

This data science course will provide students with opportunities to understand the data science process of asking questions, gathering and organizing data, modeling, analyzing and synthesizing, and communicating. Students will work through this process in a variety of contexts. Students learn through making sense of complex problems, then through an iterative process of formulation and reformulation coming to a reasoned argument for the choices they will make. All of the Standards of Mathematical Practice (SMP) will be addressed in this course.

Unit 1 - Data Tells a Story

In this unit, students will be introduced to data science through a reflection on their own experiences through data, an exploration of a larger dataset of people's media use, and an analysis of business data. Through these activities, students will learn about the data science process, begin using data to tell stories, and think about the ethics involved in working with data. Students will also familiarize themselves with data science software they will use throughout and beyond the course such as CODAP and Google Sheets.

Topical Outline

- What are variability, data, and models?
- Data ethics
- Data science inquiry: asking questions of data
- Univariate, bivariate and multivariate data
- Creating visual representations
- What is the story I can tell from this data?
- Data cleaning
- Key Assignments

Summary of Key Assignments and/or Activities

The key assignment in this unit is called "Dear Data." Students will learn that data can be collected and represented in creative ways. Students will consider the model that represents their data, and the part of their story that the data shows. Students will also explore what interests them in a large data set on people's media usage. Then, they will again tell a story about the facet they investigate in the large data set.

Unit 2 - The Data of Our Community: Learning from Data Distributions

In this unit, students will explore different ways of modeling data distributions, starting with shape, center, and spread, and moving to considerations of sampling. Students will likely already be familiar with the calculations needed to find measures of center and spread for small data sets, but this unit takes a deeper dive into the shapes of distributions, what the measures mean and their limitations, and considers them in the context of data modeling. Additionally, students will collect their own data and compare it to a larger data set. In doing, this they will consider their sampling choices and those of the larger data set and how they affect the comparisons they are making.

Topical Outline

- Using measures of center and spread to model data
- Distributions and normal distributions
- Data representations
- Sampling and variability
- Probabilistic thinking

Summary of Key Assignments and/or Activities

Students will analyze the measures of center and spread and consider what they can tell us about the data set. Students will learn that these measures of center don't tell the full story, and data sets with the same measures of center can look very different from each other. Students will collect their own data to compare the measures of center of their collection to a larger set of data. What differences appear between the measures of center in their smaller sample compared to a larger, more general one? How do the students' chosen population and sampling methods affect what they see in the data?

Unit 3 - Water in Your Life: Bivariate Data and Causality vs. Spurious Correlation

In this unit, students will learn about bivariate data through discussions and data explorations around the theme of water usage. Students will explore scatter plots as a visual way to represent the relationship between two variables; within them, they will create their own lines of best fit as well as learn about the ways in which these are usually determined and analyzed in data science work. Throughout the unit, they will use the analytic tools they are learning to make and refine claims about water usage based on both self-collected data and large, publicly available data sets. During the unit, students will work in Google Sheets, CODAP, and Tableau.

Topical Outline

- Linear regression and bivariate data
- Using probability to analyze the fit of a regression
- Make connections between the trend and the context to make predictions
- Spurious correlations, confounding and mediating variables and data ethics
- Evaluating claims: spurious correlation vs causality

Summary of Key Assignments and/or Activities

Students will collect and analyze data about water usage based on the number of people in their household. From this data, they will estimate a line of best fit, describe where they place the line, why, and what it tells them about the data. They will make connections to finding the line of best fit by using this line and their data to find the residuals and squares of residuals. They will compute “r squared” for their data and communicate what this tells them about their least squares line and their data. They will then make claims about their data based on their analysis, and consider a model of the process to indicate how much of the story their data analysis tells. Students will then analyze a larger data set of water usage by city that includes additional variables. Students will explore and analyze this data set in Tableau, make statements based on their findings, and draw connections between different variables and water usage across cities.

Unit 4 - Shuffling Songs: Probabilistic Modeling

In this unit, students will again consider the modeling process and the role played by variation, reflecting on the data collected from simulations and the ways data can help answer probabilistic questions and leverage this power for decision-making. In the process of creating powerful simulations, students will learn the basics of programming, which will continue to be a powerful tool for data analysis. Additionally, students are introduced to conditional probability in the context of the unit question. During this unit, students will use Python in EduBlocks and Google Colab.

Topical Outline

- Algorithmic Thinking
- Basics of programming
 - Variables
 - Loops
 - If-then statements
- Variability
- Simulation
- Probability
 - Theoretical and Experimental Probability
 - Conditional Probability

Summary of Key Assignments and/or Activities

Students consider the probability of different genres being played when a playlist is shuffled. They build a class playlist and discuss the theoretical probabilities of each genre, and then in order to calculate experimental probabilities, students program their own simulations using block-based coding in Python. In order to prepare students to program their shuffling simulation, they build and analyze a series of simpler programs that help them become familiar with the key ideas behind basic programming. They compare their experimental probabilities to the theoretical probabilities of each song having a given genre.

Unit 5 - Skin Tones and Representation: Categorical Data and Introduction to Linear Algebra

In this unit, students explore the issues around skin tone representation in the media through a data-based exploration of skin tone representation in magazines. Students conduct both a categorical and a numerical analysis and compare the benefits and drawbacks of both. In their categorical analysis, students create two-way tables based on their interpretation of the skin tones of the people pictured, and in the numerical analysis they use the RGB values of the images themselves. After both analyses, students chose an audience for whom the information would be relevant and write a data-supported piece to share their

findings with that audience. During the unit students, will work in Google Sheets and Google Colab (Python).

Topical Outline

- Pros and cons of different ways of data collecting
- Collecting categorical data
- Two-way tables
- Foundations in Linear Algebra: Working in higher dimensional spaces
- Introduction to clustering
- Probability

Summary of Key Assignments and/or Activities

Students will discuss different ways of collecting and analyzing data. First, students will collect and analyze categorical data on the representation of different skin tones in the media. After a categorical analysis using two-way tables, they learn to conceptualize color as points in multi-dimensional space and use a numerical/linear algebraic approach to analyze the same data using clustering. Students will create a piece of writing to communicate their data-supported findings around media representation of skin tones to an audience of their choosing.

Unit 6 - What's the Best Place for Me? Modeling with Data and Understanding Bias

In this unit, students will build a prioritization model to create a ranking. In this process, students will decide what they value, collect variables based on their values, gather and clean data, create functions to combine variables, normalize data, and create a weighting system for prioritizing their data. Students will perform a sensitivity analysis on their weighting system. During this process, students will discuss how bias impacts mathematical models. They will use reasoning, justifications, and visualizations to explain their decisions. During this unit students will use Google Sheets, Google Data Commons, and Tableau.

Topical Outline

- Bias
- Data collection and cleaning
- Normalization and weighting of data
- Forming mathematical models
- Sensitivity analysis
- Writing reports and communicating findings

Summary of Key Assignments and/or Activities

In this assignment, students will analyze the bias of a published list of best places to live. Students will analyze the attributes that publishers value. Students will then create their own ranking and prioritization. Students analyze data available via the Google Data Commons "application programming interface" (API) to create a list of criteria for what is most important to them regarding the place(s) in which they would like to live. This will be an inquiry driven unit of study. They will then use those key characteristics along with Data Commons and Google Sheets to gather, analyze, and prioritize that data to formulate a model through which they will generate a set of countries or cities wherein they might travel.

Unit 7 - Predicting My Preferences: Introduction to Machine Learning

In this unit, students will be introduced to the big ideas behind machine learning. They will build two different machine learning algorithms to make predictions on whether they will like a song. In this process

they will learn about using vectors and matrices as data structures as well as applying conditional probability and exercising their basic programming abilities. Students will also consider how machine learning impacts their lives and others' lives and will share their newly gained understandings of machine learning with a member of their community. During the unit, students will work in Colab and Edublocks.

Topical Outline

- Predictive modeling
- Machine learning
- Basic programming
- Linear Algebra
- Conditional Probability

Summary of Key Assignments and/or Activities

Students consider the basic ideas behind machine learning. They explore and adapt algorithms to predict song ratings based on song attributes and their peers' ratings. As part of their work in these algorithms, students explore the concepts of train/test split of datasets, complexity of modeling functions, conditional probability as a measure of similarity, and weighted averages. Students use their knowledge of basic programming to work in EduBlocks and Google Colab. Additionally, they consider the ethical implications of the use of machine learning in the context of music recommendations and beyond. The assignment concludes with students sharing their knowledge of machine learning and how it impacts their lives with a member of their community.

Unit 8 - Being a Data Scientist

This unit will bring together all that the students have been working on. Students will have an opportunity to work through the full cycle of data science: making their own decisions about the questions they are interested in exploring, finding data to answer that question, cleaning the data, creating and analyzing a model, communicating with the data visually and reflecting on their process. This will be an iterative process mirroring how data scientists work on a project. Students will gather their own data. They will make decisions about how to work with it and describe the choices they have made including what technology tools to use, cleaning moves, visualization selection, univariate or bivariate data choices, combining data, and other content relevant to their project of choice.

Topical Outline

- Asking questions
- Gathering and organizing data
- Modeling
- Analyzing and synthesizing
- Communicating

Summary of Key Assignments and/or Activities

In their final assignment, students will write a question on a topic they are interested in learning more about. Students will collect local data from different partner groups (for example: teachers, students, parents, local business, community members, school administrators) or find a dataset of interest and make a model based on the data. Students will decide on their audience and create a product of their choice to communicate their findings. Their product will include data visualizations along with clear justifications. In this project, students will choose which technology tools will best support their analysis and explain their choices.

Interpret Categorical and Quantitative Data**S-ID Summarize, represent, and interpret data on a single count or measurement variable.**

1. Represent data with plots on the real number line (dot plots, histograms, and box plots).
2. 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.
3. Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).
4. Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages. Recognize that there are data sets for which such a procedure is not appropriate. Use calculators, spreadsheets, and tables to estimate areas under the normal curve. Summarize, represent, and interpret data on two categorical and quantitative variables.
5. Summarize categorical data for two categories in two-way frequency tables. Interpret relative frequencies in the context of the data (including joint, marginal, and conditional relative frequencies). Recognize possible associations and trends in the data.
6. Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.
 - a. Fit a function to the data; use functions fitted to data to solve problems in the context of the data. Use given functions or choose a function suggested by the context. Emphasize linear, quadratic, and exponential models.
 - b. Informally assess the fit of a function by plotting and analyzing residuals.
 - c. Fit a linear function for a scatter plot that suggests a linear association. Interpret linear models.
7. Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.
8. Compute (using technology) and interpret the correlation coefficient of a linear fit.
9. Distinguish between correlation and causation. Making inferences and justifying conclusions.

S-IC Understand and evaluate random processes underlying statistical experiments.

1. Understand statistics as a process for making inferences about population parameters based on a random sample from that population.
2. Decide if a specified model is consistent with results from a given data-generating process, e.g., using simulation. For example, a model says a spinning coin falls heads up with probability 0.5. Would a result of 5 tails in a row cause you to question the model? Make inferences and justify conclusions from sample surveys, experiments, and observational studies.
3. Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.
4. Use data from a sample survey to estimate a population mean or proportion; develop a margin of error through the use of simulation models for random sampling.
5. Use data from a randomized experiment to compare two treatments; use simulations to decide if differences between parameters are significant.

6. Evaluate reports based on data.

S-CP Conditional Probability and the Rules of Probability

Understand independence and conditional probability and use them to interpret data.

1. Describe events as subsets of a sample space (the set of outcomes) using characteristics (or categories) of the outcomes, or as unions, intersections, or complements of other events (“or,” “and,” “not”).
2. Understand that two events A and B are independent if the probability of A and B occurring together is the product of their probabilities, and use this characterization to determine if they are independent.
3. Understand the conditional probability of A given B as $P(A \text{ and } B)/P(B)$, and interpret independence of A and B as saying that the conditional probability of A given B is the same as the probability of A, and the conditional probability of B given A is the same as the probability of B.
4. Construct and interpret two-way frequency tables of data when two categories are associated with each object being classified. Use the two-way table as a sample space to decide if events are independent and to approximate conditional probabilities. For example, collect data from a random sample of students in your school on their favorite subject among math, science, and English. Estimate the probability that a randomly selected student from your school will favor science given that the student is in tenth grade. Do the same for other subjects and compare the results.
5. Recognize and explain the concepts of conditional probability and independence in everyday language and everyday situations. For example, compare the chance of having lung cancer if you are a smoker with the chance of being a smoker if you have lung cancer. Use the rules of probability to compute probabilities of compound events in a uniform probability model.
6. Find the conditional probability of A given B as the fraction of B’s outcomes that also belong to A, and interpret the answer in terms of the model.
7. Apply the Addition Rule, $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$, and interpret the answer in terms of the model.
8. (+) Apply the general Multiplication Rule in a uniform probability model, $P(A \text{ and } B) = P(A)P(B|A) = P(B)P(A|B)$, and interpret the answer in terms of the model.
9. (+) Use permutations and combinations to compute probabilities of compound events and solve problems.

S-MD Using Probability to Make Decisions

Calculate expected values and use them to solve problems.

1. (+) Define a random variable for a quantity of interest by assigning a numerical value to each event in a sample space; graph the corresponding probability distribution using the same graphical displays as for data distributions.
2. (+) Calculate the expected value of a random variable; interpret it as the mean of the probability distribution.
3. (+) Develop a probability distribution for a random variable defined for a sample space in which theoretical probabilities can be calculated; find the expected value. For example, find the theoretical probability distribution for the number of correct answers obtained by guessing on all five questions of a multiple-choice test where each question has four choices, and find the expected grade under various grading schemes.
4. (+) Develop a probability distribution for a random variable defined for a sample space in which probabilities are assigned empirically; find the expected value. For example, find a current data

- distribution on the number of TV sets per household in the United States, and calculate the expected number of sets per household. How many TV sets would you expect to find in 100 randomly selected households? Use probability to evaluate outcomes of decisions.
5. (+) Weigh the possible outcomes of a decision by assigning probabilities to payoff values and finding expected values.
 - a. Find the expected payoff for a game of chance. For example, find the expected winnings from a state lottery ticket or a game at a fast-food restaurant.
 - b. Evaluate and compare strategies on the basis of expected values. For example, compare a high-deductible versus a low-deductible automobile insurance policy using various, but reasonable, chances of having a minor or a major accident.
 6. (+) Use probabilities to make fair decisions (e.g., drawing by lots, using a random number generator).
 7. (+) Analyze decisions and strategies using probability concepts (e.g. product testing, medical testing, pulling a hockey goalie at the end of a game).

Instructional Methods and/or Strategies

- **Student Concept Organizer:** This is a running document for the entire course where students have the option to keep track of new topics/concepts/vocabulary words. Students can use this document every time a new concept is introduced, or for the larger components of each unit. As students progress through the course and more examples of using the concepts are worked through, students are encouraged to go back and expand on their organizer definition.
- **Student Portfolio:** At the end of every unit students will build their portfolio in the form of Google Slides. Using this template, students will add a brief description of their project(s), link the projects and optionally link reflections and visuals.
- Students will utilize Google Sheets and a variety of other free applications in order to analyze data.
- **Student Weekly Reflections:** In each unit, students write reflections in Google Docs. These tasks are an opportunity for students to reflect on their learning from the week.

Assessment Methods and/or Tools

Explorations in Data Science is a Project Based Learning (PBL) program and, as such, students keep a detailed portfolio with links to all the work and extension reflections of each unit. Traditional unit exams are not part of this course. Students are assessed with specific rubrics for each unit objective.

Summative Assessments

Students are assessed with specific rubrics for each unit's summative assessment of objectives.

Assessment Criteria (grading breakdown)

A = 100-92%	B- = 80-82%	D+ = 66-69%
A- = 89-91%	C+ = 77-79%	D = 60-65%
B+ = 87-88%	C = 72-76%	D- = 55-59%
B = 82-86%	C- = 70-71%	F < 55%