Note to Reader – this is a working draft of our curriculum which will be updated at regular intervals.

Explorations in Data Science: Youcubed Adaptable Curriculum

Grades: 11,12

Length: Full Year

Environment: Classroom-based

Honors: Optional

Subject: Mathematics (C)

Discipline: Data Science and Statistics

Institution: Youcubed@Stanford University.

Course Overview

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. We will apply for A-G approval of the course, which would mean the 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.

This course is dependent upon the use and application of a variety of technologies. The appropriate and strategic use of these tools will be demonstrated and required throughout the course. The tools required will include CODAP (https://codap.concord.org/) for analyzing and visualizing data, Google Sheets for analyzing and visualizing large amounts of data (on the order of hundreds of data points), the Google Data Commons API (a website wherein students will gather, sort, visualize, and export country data that is freely available to the public, https://www.datacommons.org/), Tableau for analyzing data and creating visuals, and Python through Google Colaboratory, as students learn to use coding with larger data sets. Each tool required is widely accessible and web-based, downloading apps and software is not necessary for the use of this course.

This course has several opportunities for students to develop their explanatory writing skills across multiple platforms. Communication at every stage of the data science process is key in making sense of a context, its data, interpretation, and story. Students will revise and refine their writing using self, peer, and teacher feedback.

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

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

Key Assignments

In this unit, 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 does 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

Key Assignments

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

Key Assignments

In this assignment, 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).

Note: One of our goals throughout this curriculum is to provide teacher and student flexibility and choice. In support of that goal, here is a description of alternative topics you could use for this unit.

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

Key Assignments

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 do 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

Key Assignment

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 choose to live.

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

Key Assignments

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

Key Assignments

In this final assignment, students will write a question on a topic they are interested in learning more about. Students will collect local data from different stakeholders (for example: teachers, students, parents, local business, community members, administration) 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.