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

Youcubed Adaptable Curriculum: Explorations in Data Science

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 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 introductory unit students will make sense of the questions: What part of the story is told by data? What is variation? How is data generated? What data is gathered about themselves? They will consider the ways data can be used to model the world. They will also begin to think about data ethics, considering: What is the ethics behind all this data that is collected about people? As students learn about data, they will be introduced to many different ways to represent data. They will explore univariate, bivariate, and multivariate data. From the data visualizations they will consider what the story is they can tell from this data. During the unit, students will be learning to use 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”. In this assignment students will collect data from their own lives and represent it. They will learn that data can be represented in creative ways and will collect and represent it in their own way. Students will consider the model that represents their data, and the part of their story that the data shows. Students will also explore a large data set and find what interests them in that data set and tell a story about the large data set.

Unit 2 - Data Distributions

Unit 2 will explore different ways of modeling data, starting with the basic models of measures of center and spread, as well as considering 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 what they mean, their limitations, the impact of outliers and considering them in the context of data modeling. Students will explore distributions and the role of probability in understanding them. 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

- Modeling data using measures of center and spread
- 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 - Bivariate Data: 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 draw 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 variables and data ethics
- Evaluating claims: causality vs. spurious correlation

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, describing where they are placing the line, why, and what it tells them about the data. They will use this line and their data to find the residuals and squares of residuals. And make connections to finding the line of best fit. 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 how much of the story their data analysis tells, considering a model of the process. Students will then analyze a larger data set of water usage by city which includes additional variables. Students will explore and analyze this data set in Tableau and make statements based on their findings making connections between different variables and water usage across cities.

Unit 4 - Making Decisions with Data

In this unit, students will again consider the modelling 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. During this unit students will use Python in Edu-Blocks.

Topical Outline

- Algorithmic Thinking
- Basics of programming
  - Variables
  - Loops
  - If-then statements
- Simulation
- Variability
- Conditional Probability
- Probabilistic Modeling
- Decision Trees

Key Assignments

Students will use manipulatives to simulate a basic probability model and collect data from their experiments. They will then aggregate their data with the rest of the class' data and see how more simulations get them closer to the expected results. In order to harness this power, students program their own simulations using block-based coding in python. They will then be able to analyze the results of their simulations and use those to make decisions about a real-life problem.

Unit 5 - Categorical Data and Linear Algebra

In this unit, students will discuss different ways of collecting and analyzing data. Students will discuss surveys and questionnaires as data collection mechanisms. Students will learn how to create surveys, collect and analyze categorical data. They will also delve into the use of vectors to organize data into a multi-dimensional space to understand how data are similar or different to each other. During the unit students will work in Google Forms, spreadsheets, and Python.

Topical Outline

- Pros and cons of different ways of data collecting
- Design a survey and collect categorical data
- Two-way tables
- Vectors

- Linear Algebra
- Clustering
- Probability
- Data moves
    - Aggregating/grouping data
    - Filtering data

Key Assignments

Students will design class questionnaires, collect categorical data, and build two-way tables to compare them. Students will do this initial analysis using Google Sheets. They will create a model using vectors to measure how far the responses are apart from others. Students can plot these vectors in one, two, and three, dimensional space on the computer. Students will build clusters from this data.

Unit 6 - Modeling with Data and Understanding Bias

In this unit students will create 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, creating 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 justifications, explanations and visualizations to explain their decisions. During this unit students will use Google Sheets and Google Data Commons.

Topical Outline

- Bias
- Data collection and cleaning
- Normalization 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 - Data Predictions

In Unit 7, students will gather and clean data to make predictive models using the tools they have learned up until now. They will then be introduced to machine learning and will use machine learning on the same data to make more efficient and accurate predictions. During the unit, students will work in Google Sheets and Python.

Topical Outline

- Data cleaning
- Predictive modeling
- Probability
- Machine learning
- Basic programming

Key Assignments

In this assignment, students will be given a large data set with many different characteristics to analyze. Students will choose which attributes in the data would be interesting for them to make predictions about. Students could compare these to see which correlate the most. They will then find the five most correlated attributes. They will create a model for predicting their characteristic of choice and test and refine their model. They will then use machine learning to predict how good the model they created is compared to the machine learning model.

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 interactive 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
- Finding and collecting data
- Cleaning data
- Creating and analyzing a model
- Evaluating models
- Communicating with data

Key Assignments

In this final assignment, students will write a question on a topic they are you interested in learning more about. Students will collect local data from different stakeholders (for example: teachers, students, parents, local business, community members, administration) and make a model based on data collected. Students will make a recommendation to their local government, school, and others, based on the model. Their recommendations will include data visualization along with clear justifications. In this project, students will choose which technology tools will best support their analysis and explain their choices.