

## Coding Like a Data Miner: A Culturally Relevant Data Analytics Intervention for High School Students

The "[Coding Like a Data Miner](#)" project is a data science-based computer science (CS) curriculum funded by the [National Science Foundation](#) that uses culturally relevant and responsive pedagogies to constructively center diverse learners in STEM. The curriculum was co-created through participatory curriculum design sessions with educators and youth stakeholders, resulting in activities that begin with highly scaffolded learning to increase accessibility and eventually transition to free inquiry, allowing learners to pursue topics along their personal interests, cultural backgrounds, and/or sociopolitical histories.

The curriculum emphasizes student access to datasets from the social media platform Twitter, providing access to data sources that are both familiar to social media natives and expansive across topics and opinions, enabling users to participate along trajectories couched in their personally and culturally relevant interests. The curriculum was structured around the use of Twitter's Application Programming Interface (API), providing an authentic sandbox-like environment (sandbox data science) that retains material, personal and disciplinary agency for learners to go beyond their typical roles as consumers of information to actively serve as producers of knowledge on their own terms with real, complex, and messy data. The use of API access also provides an opportunity for practical computer science skill development. To read more about our project development, check out our [publications and presentations repository](#).

The curriculum consists of [17 modules](#) that iteratively guide learners through authentic data science practices organized by the amount of scaffolding and support offered: Introduction (7 modules), Guided Inquiry (5 modules), Scaffolded Inquiry (3 modules), and Free Inquiry (2 modules). The curriculum overview and learning objectives detailed below also align with the [CollegeBoard AP Computer Sciences Principles](#). For teacher resources, check out our repository of [lesson plans](#) associated with each module.

### Curriculum Overview and Learning Objectives

	Module	Objectives
Introduction	<b>1.1 Curriculum Overview</b>	<b>Provides an overview of curriculum structure and goals</b>
	<b>1.2 Introduction to Data Visualization</b>	<b>Students will be able to create simple visualizations in order to better understand data and identify trends</b>
	Module 1.2.1	Understand different types of data
	Module 1.2.2	Visualize data using Google Earth
	Module 1.2.3	Visualize data by creating a word cloud
	Module 1.2.4	Visualize one-dimensional data using Excel
	Module 1.2.5	Visualize two-dimensional data using Excel
	<b>1.3 Introduction to Data Analysis</b>	<b>Students will be able to perform basic data analysis in order to better understand data, its types, and structures</b>
	Module 1.3.1	Introduction to data analysis
	Module 1.3.2	Introduction to basic data analysis techniques
	Module 1.3.3	Introduction to basic quantitative data analysis
	Module 1.3.4	Introduction to basic qualitative data analysis
	<b>1.4 Introduction to Data Gathering</b>	<b>Students will be able to import and export data using Python in order to manipulate Data using different tools</b>

Guided Inquiry	Module 1.4.1	Introduction to data collection methods
	Module 1.4.2	Import and export data using existing code
	Module 1.4.3	Collect data using Python
	<b>1.5 Introduction to Statistics</b>	<b>Students will be able to use and manipulate existing Python code in order to extract basic statistical information from data</b>
	Module 1.5.1	Introduction to basic statistics
	Module 1.5.2	Introduction to descriptive statistics
	Module 1.5.3	Execute code fragments to calculate descriptive analysis
	Module 1.5.4	Introduction to data variability
	Module 1.5.5	Execute code to extract data range and interquartile range
	Module 1.5.6	Execute code to draw a boxplot diagram
	<b>1.6 Introduction to Data Pre-processing</b>	<b>Students will be able to perform basic data preprocessing in order to address missing data, inconsistent data, and noisy data</b>
	Module 1.6.1	Introduction to data pre-processing and its importance
	Module 1.6.2	Identify missing data and how to deal with it
	Module 1.6.3	Identify noisy data and how to deal with it
	Module 1.6.4	General data pre-processing methods: a. Data normalization b. Data attribute selection c. Data reduction
	<b>1.7 Introduction to Coding</b>	<b>Students will be able to understand basic Python code in order to use Python to manipulate, analyze, and visualize data</b>
	Module 1.7.1	Introduction to Google Colab
	Module 1.7.2	Introduction to basic coding in Python
	Module 1.7.3	Introduction to basic data types in Python (integers, float, dictionary)
	Module 1.7.4	Introduction to basic condition structures
	Module 1.7.5	Introduction to basic loop structures
	<b>2.1 Let's Gather Some Data</b>	<b>Students will be able to manipulate code run queries against Twitter in order to extract data from Twitter</b>
	Module 2.1.1	Understand different data gathering sources
	Module 2.1.2	Create a developer account on Twitter
	Module 2.1.3	Create an application on Twitter
	Module 2.1.4	Set up a connection to Twitter
	Module 2.1.5	Set up a search query
	Module 2.1.6	Extract data from Twitter
	<b>2.2 Let's Do Preprocessing</b>	<b>Students will be able to understand and manipulate Python code in order to</b>

Scaffolded Inquiry		<b>perform basic data preprocessing steps</b>
	Module 2.2.1	Use Python code to identify missing data
	Module 2.2.2	Use Python code to identify noisy data
	Module 2.2.3	Use Python to perform data normalization
	Module 2.2.4	Use Python to perform data reduction
	<b>2.3 Let's Analyze</b>	<b>Students will be able to apply descriptive statistics in order to analyze and manipulate datasets.</b>
	Module 2.3.1	Perform basic quantitative data analysis
	Module 2.3.2	Apply variability analysis
	Module 2.3.3	Conduct an outlier analysis
	Module 2.3.3	Conduct variance analysis
	<b>2.4 Let's Do Statistics</b>	<b>Students will be able to manipulate existing Python code in order to generate statistical visualizations</b>
	Module 2.4.1	Manipulate code to visualize statistical distributions on numerical data
	Module 2.4.2	Manipulate code to visualize statistical distributions on categorical data
	Module 2.4.3	Manipulate code to visualize a boxplot
	Module 2.4.4	Manipulate code to visualize statistical relationship among variables
	<b>2.5 Let's Visualize</b>	<b>Students will be able to use Python Libraries and IBM Watson Community Portal in order to visualize data using several methods</b>
	Module 2.5.1	Create data visualizations using Excel advance charts
	Module 2.5.2	Create and visualize a pivot data
	Module 2.5.3	Introduction to visualizations using Python Library 1 (Matplotlib)
	Module 2.5.4	Introduction to visualizations using Python Library 2 (Seaborn library)
	<b>3.1 Let's Mine Twitter Data</b>	<b>Students will be able to set up connection to twitter API in order to extract datasets</b>
	Module 3.1.1	Create a developer account on Twitter
	Module 3.1.2	Create an application on Twitter
	Module 3.1.3	Set up connection to Twitter
	Module 3.1.4	Set up a search query
	Module 3.1.5	Extract data from Twitter
	Module 3.1.6	Summarize data with a worksheet
	<b>3.2 Let's Analyze Twitter Data</b>	<b>Students will be able to identify and apply descriptive statistics and data variability analysis in order to objectively answer questions</b>
	Module 3.2.1	Identify and apply descriptive analysis to answer questions
	Module 3.2.2	Identify and apply variability analysis to answer questions
	Module 3.2.3	Examine data to determine whether outliers exist

Free Inquiry	<b>3.3 Let's Visualize Twitter Data</b>	<b>Students will be able to use advanced features in matplotlib and seaborn data visualization libraries in order to construct advanced visualizations</b>
	Module 3.3.1	Visualize Twitter data using word cloud
	Module 3.3.2	Visualize Twitter data using the Matplotlib library
	Module 3.3.3	Visualize Twitter data using the Seaborn library
	<b>4.1 Choose a Topic for Mining Twitter Data</b>	<b>Students will be able to independently design, implement and report on a line of inquiry based on unique primary source data sets</b>
	Module 4.1.1	Select a topic or domain
	Module 4.1.2	Identify questions
	Module 4.1.3	Mine datasets that pertain to the questions
	Module 4.1.4	Analyze datasets in order to objectively answer questions
	<b>4.2 Topic Sharing and Group Discussion</b>	<b>Students will be able to present findings for a line of inquiry based on a primary data source drawn from a social media platform</b>
	Module 4.2.1	Open discussion on topics
	Module 4.2.2	Open discussion on datasets
	Module 4.2.3	Open discussion on questions
	Module 4.2.4	Open discussion on produced results