Data Science Syllabus

High School - One Semester (85 hours)

Course Overview and Goals

Industries of all types are hiring data scientists to analyze and highlight the hidden patterns in data. This course equips students with the essential skills of a data scientist which include data collection, cleanup, transformation, analysis, and visualization. Students will write algorithms, tell data stories, and build statistical models using Python libraries. They will use the same tools that data scientists use to draw meaningful insights and solve organizational problems.

Learning Environment

This course utilizes a blended classroom approach. The content is fully web-based, with students writing and running code in the browser. Each module of the course is broken down into lessons. Lessons consist of video tutorials, short quizzes, example programs to explore, and written programming exercises.

Programming Environment

Students write and run Python programs in the browser using the CodeHS editor.

Projects and Assessments

Students complete a project within each module while also learning and applying new material. The majority of lessons include a formative short multiple-choice quiz to check for understanding. At the end of each module, students take a summative multiple choice quiz that assesses their knowledge of the concepts covered in the module.

Prerequisites

The Data Science course is designed for intermediate computer science students with at least some knowledge of programming (not language specific) and an interest in computer science. The course is highly visual, dynamic, and interactive, and engaging.

More Information

Browse the content of this course at https://codehs.com/course/12135

Course Breakdown

Module 1: The Data Science Life Cycle (4 weeks/20 hours)

Students will learn and apply the process of the data science life cycle. This includes asking statistical questions, collecting or obtaining reliable raw data, analyzing the data using measures of central tendency and spread and interpreting, and summarizing the results.

Objectives / Topics Covered	• What is Data Science? • Gathering Data

	○ Quantitative/Qualitative ● Exploring Data Using Python ● Modules and Libraries ● Using the Pandas Library ○ Series ■ Measures of Central Tendency ■ Measures of Spread ○ DataFrames ■ Selecting Columns ■ Using Functions
Example Assignments / Projects	● **Mini-Project:** Students will go through the first two steps of the data cycle using data of their choosing. ○ Ask Questions: Formulate a statistical question that can be answered with data. ○ Consider Data: Collect or find data that will aid in answering your question. ○ Analyze Data: Perform statistical analysis, run calculations and/or create data displays to identify patterns and relationships ○ Interpret Data: Answer questions and summarize the results. ● **Hot Dog Plots:** Use the correct Python functions to create a boxplot of the data. Using the graph, determine the summary statistics and the spread. ● **Roller Coaster Rankings:** Define a function that will compute a score for each roller coaster. Use this function to store the results in a new column. ● **Student Test Scores:** Create a function that finds the maximum test score between test one and test two for each student. Create a function that finds the maximum test score between all three tests for each student. Decide which calculations, along with these two new columns, can help you answer the original statistical question? Explore and further analyze your data until you come to a conclusion.

Module 2: Data Science for Change (3 weeks/15 hours)

Students will use and analyze data to better understand a problem, measure the scope of a problem, or understand how people are affected by the problem. They will learn more about cleaning a dataset and filtering by column, rows, and conditions.

Objectives / Topics Covered	● What is Big Data? ○ Cognitive Bias ● Importing and Filtering ○ `loc` ○ `iloc` ○ By a condition ● Data Cleaning ○ Dropping Data ○ Fixing Data Types ● Exploring with Data Visualizations
Example Assignments	● **Project - Data Science for Change:** Students will run through the data

/ Projects	science life cycle with the intent to use data to better understand a problem, to measure the scope of a problem, and to understand how people are affected by the problem. ● **Instagram Filters:** This dataset consists of popular Instagram accounts and their number of followers (in millions). Use conditional filtering to print the rows where followers are greater than 230 (million), print only the account columns of those from the United States, print the account and followers columns of the row with the maximum number of followers. ● **Book Conditions**: This data was acquired from the Google Books store. It includes the title of each book, the author(s), the rating (from 1-5), the total voters, the price, the publisher, the page count, and the date the book was published. Print the title, rating, and voters columns for books that have a rating of 4.0 or higher and over 9000 voters. ● **Cleaning Book Data:** Check the data types of the dataset. Do they look okay? Do they need to be changed at all? Permanently drop the publisher and published_date columns. Print the shape of the data and check for duplicate rows. How many are there? Permanently drop duplicate rows from the dataset. Determine the number of missing values in the dataset. What would be the best decision for dealing with the missing values? Make the call and change the data.

Module 3: Data Storytelling (4 weeks/20 hours)

Students will use and analyze data to tell a data story. They will create a visually appealing infographic that displays important data visualizations. The infographic will also tell a story based on their interpretation after exploring, analyzing, and visualizing the data.

Objectives / Topics Covered	● Types of Data Stories ● Data Visualizations ○ Univariate and Bivariate Data ● Normal Distribution ● Trends and Correlations ● Linear Regression
Example Assignments / Projects	● **Project - Data Storytelling:** Students will get to tell their own data stories. They will create a visually appealing infographic that displays important data visualizations. The infographic will also tell a story based on their interpretation after exploring and analyzing data. ● **State Education:** The pie chart displayed has an error, which makes it a misleading visual. Find the bug in the program and fix it so that the program displays an accurate pie chart. ● **Precipitation:** There are two datasets in this activity that each list the average temperature and precipitation information for different states. Compare the precipitation averages for both states by plotting a line chart. What conclusions can you make from the chart?

	• **Professor's Salaries:** This dataset lists professor salaries collected in a survey. Other fields collected were the type of degree, years since their Ph.D. was earned, years of service, and sex (gender). Create a filtered table that lists only female professors and one that lists only professors with over 10 years of service. Do any of the filtered tables follow a normal distribution? • **Swim Time Regression:** This dataset lists the gold medalist time for the women's 400-meter freestyle swimming finals. Plot the scatterplot. Create a model using the `polyfit()` function and use the model to plot the line of best fit. Use the model equation to make predictions based on different values.

Module 4: Data Science for Business (5 weeks/25 hours)

Students will gather business data that can be used to make decisions about how to better the company or product. They will present their findings in a business report that suggests several action items that they predict will help the business's performance and growth.

Objectives / Topics Covered	• Determining Dataset Quality
	• Aggregating Data
	  ○ Grouping
	  ○ Sorting
	• Combining Datasets
	  ○ Concatenating
	  ○ Joining/Merging
	• Bias in Data Analytics

Example Assignments / Projects	• **Project - Data Science for Business:** Students will gather (or create) business data that can be used to make decisions about how to better the company or product. They will present their findings in a business report that suggests several action items that they predict will help the business's performance and growth. • **Determining Completeness:** The raw datasets (without any cleaning) are provided. Your task is to check the completeness of at least two different datasets to compare and contrast them. • **Fuzzy Book Titles:** Use the FuzzyWuzzy library to extract all titles that are close to matching the word "batman". Print out the results. Are they unique book titles or do you notice errors in the titles? • **Billionaire Sort:** Group the dataset by gender. Only display the gender count. What insights does this grouping give you? Group the dataset by industry. Only display the industry count. What insights does this grouping give you? Sort by age and then net worth. You will want the oldest person to be listed first. If there are two people of the same age, you'll want the person with the highest net worth to be listed first. • **Cereal Production:** Let's determine who would get the "Most Improved" award for a ten-year span. ○ Create a function that will return the difference between two years' values.

	○ Create a new column listing the growth from the year 1990 to the year 2000. ○ Sort by your new column. You will want the highest growth values to be listed first. Only print the Country Name and the new growth columns. Only print the first five rows. ● **Concatenating Cats:** The cat shelter uses two different databases - one for male cats and one for female cats. They'd like to combine the two datasets. Explore and concatenate the two datasets. ● **School Nurse:** Listed in this activity are the two datasets that resulted from the nurse's two visits. The roster has changed from his first visit (first.csv) to his second visit (second.csv). The current roster of students can be found in the second file. How can we keep all of the students in the second roster but add in missing information that might be found in the first roster?

Module 5: What's Next? (1 week/5 hours)

Students will explore the next chapter in learning about data science and the careers that are available and growing.

Objectives / Topics Covered	● Data Science Pathways ● Artificial Intelligence ● Python Programming