Accessible Integration: Empowering Biology Curriculum with Statistics and Computer Science on Posit Cloud

An Interactive Qualifying Project
Submitted to the Faculty of
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements for the
Degree of Bachelor of Science

By Vivek Kandasamy

Date:
11 August 2023
Project Presented To:
Professor Elizabeth Ryder, Advisor

Abstract

As time progresses, more life sciences organizations now seek applicants with more than just biology skills, especially in data science, statistics, and computer science. However, most schools in Massachusetts have not created a curriculum that integrates biology with statistics and computer science concepts. This project aimed to create a high school curriculum that can be incorporated into core math or biology classes and made accessible to teachers and students in Massachusetts and beyond. The curriculum incorporates the pedagogical techniques of"Use-Modify-Create" and "Abstraction, Automation, and Analysis" to teach students concepts interconnecting biology, statistics, and computer science using Posit Cloud.

Acknowledgments

I extend my gratitude to the individuals who offered invaluable support and guidance during the completion of this IQP. I am thankful to Professor Elizabeth Ryder, my advisor, for her unwavering support and constructive feedback that shaped the curriculum. She consistently made herself available, demonstrating patience as I navigated the challenges of curriculum development. Additionally, I am appreciative of Mrs. Maureen Chase, a Massachusetts high school teacher, for taking the time to meet with me and share insightful input on my curriculum proposal.

Table of Contents

Background

A Skills Gap in Life Sciences

Over the years, there has been a growing demand for statistics and computer science knowledge within life sciences industries including biotechnology, pharma, medicine and biological research (Attwood, 2019). The life sciences industry generates and uses growing volumes of data including clinical, pathological, and quality-of-life data (Willems et al., 2019). However, despite the growing demand, life sciences companies still find many applicants who lack these needed skill sets. The Biotechnology and Biological Sciences Research Council (BBSRC) and the Medical Research Council (MRC) did a joint report on their applicants' skill sets and found many lacked data analytics and bioinformatics (combines biology, computer science, and statistics) knowledge (BBSRC and MRC, 2021).

The skills gap may have its roots in K-12 education. According to the 2022 State of CS Report, 47% of U.S. high schools did not teach a computer science curriculum. Even in high schools that did offer computer science, students rarely enrolled in classes - in 36 states, just 5.6% of high school students enrolled in the computer science curriculum (Code.org Advocacy Coalition, 2022). While not every high school student will be proficient in statistics or computer science, it is crucial to offer them opportunities to learn these subjects so they can be exposed to the basics. Exposure to these skills can build students' confidence and increase their interest in these subjects.

Accessibility in STEM Education

STEM (Science, Technology, Engineering, and Mathematics) education is pivotal in preparing students for the dynamic technological landscape. Despite efforts to make high school STEM education more accessible, there still exists a lack of participation among students from lower-income neighborhoods and from minority groups. In a study of students ingifted education programs, 39.8% of students were enrolled in at least 1 AP course (Crabtree et al., 2019). This number dropped to 7.8% at high-poverty schools. High school students enrolled in AP Math consisted of 11% White, 16% Asian, 2.6% Black, and 2.0% Latinx. High school students enrolled in AP Science consisted of 11% White, 13% Asian, 2.8% Black, and 2.4% Latinx (Crabtree et al., 2019). Oftentimes, computer science classes are offered outside the core curriculum as stand-alone courses which are less accessible to many students (Cateté et al., 2018).

It has been shown that providing STEM-related education to minorities increases their interest in STEM topics. According to a study done by Palid on post-secondary school adults, STEM programs that aimed at improving outcomes of women and/or racially and ethnically minoritized students increased student's interest and they were more likely to pursue STEM careers when given equal opportunities. (Palid, 2023). Additionally, Kricorian studied post-secondary school adults from minority groups who pursued STEM-related jobs and found that 85% of respondents agreed that STEM topics excited their curiosity, suggesting that being exposed to STEM set them on a path to STEM-related fields (Kricorian et al., 2020).

Bio-CS Bridge

By embracing a more flexible and interdisciplinary approach to education, schools can better prepare students to adapt to an ever-changing world. The Bio-CS Bridge exemplifies this approach by creating and implementing a high school-level curriculum that integrates computer science, data visualization, and biology. Through this initiative, students are exposed to a wide range of computer science concepts rooted in the context of biology, enabling them to apply their learning to real-world problem-solving scenarios using simulations, data analysis, and visualization techniques. The curriculum was developed in collaboration with biology and computer science educators from Massachusetts and currently includes lessons utilizing various programming languages, including Python, HTML, CSS, JavaScript, Netlogo, Starlogo, and RStudio (now named Posit) The lessons are all tailored to biology and Beecology. Beecology, a citizen science project to collect data on pollinator species, serves as a foundational example of a real-world biological problem at the core of the Bio-CS Bridge curriculum. (WPI, 2023). As a result, the Bio-CS Bridge curriculum has successfully fostered the development of computational thinking skills among students in participating institutions.

Integrating Statistics and Computational Thinking into Biology Education

As the world becomes increasingly data-driven, there is a growing need to integrate statistics and computational thinking into high school-level biology classes to prepare students for the demands of the modern workforce. However, computational thinking opportunities are often segregated from the core curriculum and offered as separate courses, primarily accessible to individuals with preparatory advantages. To ensure equal opportunities for students to acquire essential skills, researchers have recognized the importance of integrating computational thinking (CT) into core classes (Cateté et al., 2018).

One of the few curriculums that incorporates statistics into biology class is Howard Hughes Medical Institute's (HHMI) Mathematics and Statistics in Biology Guide which was launched in 2015 (Strode, 2015). The curriculum includes lessons where students download biological data from a website and analyze the dataset through statistical tests. Also, the curriculum includes hands-on lessons for hypothesis testing like recording measurements of weeds they grew, and toothpick-breaking between dominant and non-dominant hands. Teachers hold in-class discussions with the students as a class to discuss their findings and the teachers weave statistical concepts into the discussions (Strode, 2015).

Another curriculum that incorporated statistics into biology was the Introduction to Data Science (IDS) course developed at UCLA (UCLA, 2022). This course provides a unique opportunity to explore student experiences in an unconventional learning path. Designed as an advanced high school mathematics course, IDS is taken after completing at least two years of high school mathematics, and teachers undergo extensive professional learning to adopt an inquiry-based approach and assist students in programming with R (Heinzman, 2022). The IDS curriculum encompasses descriptive and inferential statistics, data visualization, data collection methods, and the use of models for predictions (UCLA, 2022). As described by Gould and his colleagues (2016), the course involves classroom lessons, collaborative computer lab exercises using the statistical programming language R via RStudio, and participatory sensing campaigns. In these campaigns, students collect and analyze data directly related to their lives, such as monitoring stress levels over a few days in the Stress-Chill campaign and applying their knowledge to analyze the class data set (Gould et al., 2016; UCLA, 2022).

Pedagogy

Pedagogy refers to the theory, practice, and methods of teaching and education.It involves creating effective learning environments, engaging students in meaningful experiences, and tailoring instruction to meet individual learning needs.

One pedagogical technique for teaching Computational Thinking (CT) with STEM-related material is "Use-Modify-Create" (UMC). UMC is a pedagogical approach to teaching computational thinking (CT) that involves students gradually learning CT topics by first "Using" a given artifact, "Modifying" an existing one, and eventually "Creating" new ones. This approach helps students and teachers ease into CT topics by limiting anxiety and promoting the acquisition and development of CT skills. The UMC model also enables teachers to learn how programs represent their disciplinary knowledge, allowing them to make connections with students' learning (Lytle et. al., 2019).

This model was tested and showed supportive results. A study was performed in 2 separate middle schools where none of the teachers had experience teaching programming and they would teach a 4-day, CT lesson. Students were taught either under the "control group" pedagogical approach or UMC pedagogical approach. The test showed classroom observations where students under the UMC pedagogical approach found their lessons easier to understand and more engaging compared to students under the "control group" pedagogical approach (Lytle et. al., 2019).

Another pedagogical technique for teaching Computational Thinking (CT) with STEM-related material is the use of "Abstraction, Automation, and Analysis" (AAA) (Lee et al., 2011). This pedagogical approach teaches Computational Thinking by having students first break down a problem to the bare essentials (Abstraction), next use a computer to execute a set of tasks quickly (Automation), and finally, students reflect on whether the right assumptions were made in breaking down the problem (Analysis). This Computational Thinking process is a breakdown of preceding the “Create” phase of the UMC approach. This pedagogical technique was used in Project GUTS (Growing up Thinking Scientifically) where students created a simulation model of a school layout. Students first selected features of a real-world school that were important to model (Abstraction), next used a computer program to build and run the simulation (Automation), and then examined the model to see if the correct abstraction was made (Analysis). This approach helps students engage in CT topics and gain essential problem-solving and coding skills (Lee et al., 2011).

Methodology

When developing the curriculum, Massachusetts teaching standards were consulted. This was done so the material would be targeted for high school students that Massachusetts believes students need to learn and teachers are required to teach.

During consultations with a high school teacher from Massachusetts, valuable insights were gathered regarding the academic landscape of high school students, who typically engage in STEM courses encompassing Algebra and Calculus. It was emphasized that effective teaching should integrate both standards and practices, particularly in science education, where Massachusetts encourages the incorporation of modeling, algebra, statistics, and probability as part of their scientific practices. Given the favorable reception of RStudio (now renamed Posit) as an acceptable Integrated Development Environment (IDE), it was decided to build the curriculum around this platform. To make the curriculum accessible, the audience in mind for building the curriculum was students that do not have prior knowledge of statistics and computer science. Additionally, Chloe Byrne's IQP (Byrne, 2022), which explores the use of RStudio for her Bio-CS Project, served as an inspiring example to guide the development of the curriculum, ensuring its relevance and applicability in a high school setting.

Goals

Based on a discussion with the IQP advisor and a high school teacher from Massachusetts, the following goals were created as a baseline for creating the curriculum.

1. To create an inclusive and accessible learning environment that encourages all students, regardless of background or prior experience, to participate and excel in the curriculum, ensuring that no one is left behind in their educational journey.
2. To introduce high school students to the interdisciplinary nature of statistics, computer science, and biology in an engaging way.
3. To align curriculum to Massachusetts Department of Elementary and Secondary Education standards
4. To teach students fundamental statistical concepts and provide them practical experience with using a statistical programming language like R to analyze, manipulate, and visualize biological datasets, focusing on user-friendly tools and resources.
5. To equip students with the skills and knowledge necessary to pursue further studies or careers in statistics, computer science, or biology, while building their confidence in tackling real-world problems.

Choosing R and Posit Cloud

R was chosen as the computational platform for the curriculum. It is a simple and powerful statistical programming language that allows students to gain practical programming skills while exploring computer science and statistics concepts. R’s user-friendly syntax and extensive documentation make it understandable for students who have never done computer science in their lives, thus increasing the accessibility of the curriculum. R's ability to load datasets from various sources enables students to work with real-world data, enhancing their understanding of statistical analysis and computational thinking. R's widespread use in the field of biological sciences makes it a valuable tool that students can use to gain practical experience that aligns with the demands of the industry. Because R is an open-source programming language, all students can access its resources and utilize it without any financial burden. This democratizes access to computer science and statistics education, promoting equitable opportunities.

Several development environments were available to use R. For this curriculum, Posit Cloud was chosen primarily to ensure the accessibility of the curriculum to everyone. Students can access Posit Cloud through a website and an internet connection. This eliminates the barrier of not having a personal computer and expands access to students who do not have a computer at home and may not have a place to install a development environment like Posit.

Cost-effectiveness is another aspect that influenced the choice of Posit Cloud. Posit Cloud offers flexible subscription plans, including a free version, making it a cost-effective option for schools. All the lessons were designed and tested to ensure they can be run using the free version of Posit Cloud.

Aligning Curriculum Development to Massachusetts Educational Standards and Practices

The curriculum was built to align with the Massachusetts Curriculum Framework, specifically the 2017 Mathematics, 2016 Digital Literacy and Computer Science, and 2017 Science and Technology/Engineering curriculum frameworks. These frameworks provided a comprehensive guide that was used to create a curriculum to integrate statistics and computer science into biology.

The curriculum was aligned with the following 2017 Mathematics standards which are part of the Modeling Conceptual Category:

Standard	**Lesson where the standard is used**
S-ID (Interpreting Categorical and Quantitative Data)	
A.1. Represent data with plots on the real number line (dot plots, histograms, and box plots).	Lesson #2 Lesson #3 Lesson #5
A.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.	Lesson #3 Lesson #5
A.3. Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).	Lesson #2 Lesson #3 Lesson #5
A.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.	Lesson #2 Lesson #5
B.6. Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.	Lesson #3 Lesson #4 Lesson #5
S-IC (Making Inferences and Justifying Conclusions)	
B.3. Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.	Lesson #3 Lesson #4 Lesson #5
B.6. Evaluate reports based on data.	Lesson #2 Lesson #3 Lesson #4 Lesson #5

Along with the 2017 Mathematics standards, the curriculum was designed to align with the following standards from the 2016 Digital Literacy and Computer Science Framework:

Standard	**Lesson where the standard is used**
CT (Computational Thinking)	
Data	
9-12.CT.c 3 Create, evaluate, and revise data visualization for communication and knowledge.	Lesson #2 Lesson #3 Lesson #4 Lesson #5
9-12.CT.c 4 Analyze a complex data set to answer a question or test a hypothesis (e.g., analyze a large set of weather or financial data to predict future patterns).	Lesson #3 Lesson #4 Lesson #5
Programming and Development	
9-12.CT.d 9 Select and employ an appropriate component or library to facilitate programming solutions [e.g., turtle, Global Positioning System (GPS), statistics library].	Lesson #1 Lesson #2 Lesson #3 Lesson #5

Finally, the curriculum was designed to align with the 2016 Science and Technology/Engineering curriculum framework. The high school biology section states that "The high school biology standards place particular emphasis on science and engineering practices of developing and using models; constructing explanations; engaging in argumentation from evidence; and obtaining, evaluating, and communicating information" (Massachusetts Curriculum Framework, 2016). From Next Generation Science Standards (NGSS), the curriculum was designed to align with the following practices:

Biology Practice	**Lesson where the standard is used**
1. Asking questions (for science) and defining problems (for engineering)	Lesson #5
3. Planning and carrying out investigations	Lesson #5
4. Analyzing and interpreting data	Lesson #2 Lesson #3 Lesson #4 Lesson #5
5. Using Mathematics and Computational Thinking	Lesson #2 Lesson #3 Lesson #4 Lesson #5

The Massachusetts Curriculum Framework standards were chosen as the foundation for the lessons because they provide a comprehensive and well-structured high school framework for integrating statistics and computer science concepts with biology. These standards were applied to develop students' statistical and critical thinking skills, which are essential for understanding and interpreting data, making informed decisions, and solving real-world biology problems.

With Posit Cloud's capabilities, these standards can be effectively implemented and enriched with interactive and engaging learning experiences. Posit Cloud allows educators to access Posit and the ggplot2 package, which are powerful tools for data analysis and visualization. Students can perform statistical analyses, represent and interpret data using plots and visualizations, and draw inferences from real-world datasets, aligning perfectly with the S-ID (Interpreting Categorical and Quantitative Data) and S-IC (Making Inferences and Justifying Conclusions) standards in the Mathematics framework and the Computational Thinking and Data standards that are part of the Digital Literacy and Computer Science Framework.

Several prior works have used these Massachusetts standards in creating lessons for school students. For example, Mathspace is a free online mathematics website designed to give primary and secondary school students mathematical lessons. One of their lessons is called Descriptive Statistics and its lessons incorporate 2017 Massachusetts Curriculum Framework standards including A.1., A.2., and A.3. under S-ID (Mathspace, 2020). Another example is the Bootstrap World website which provides free integrated computer science and data science modules for grades 5-12 students. One of its modules called "Standard Deviation" incorporates 2017 Mathematics Curriculum Framework standards including A.1., A.2., and A.3. under S-ID (Bootstrap World, 2022).

Implementing Use-Modify-Create

Along with using the 2017 Massachusetts standards, the curriculum used the "Use-Modify-Create" (UMC) learning progression. The "Use-Modify-Create" (UMC) lesson progression is a Computational Thinking (CT) lesson design that has students ease into CT topics by first "Using" a given artifact, then "Modifying" an existing one, and eventually "Creating" new artifacts of their own (Lytle et. al., 2019).

In the curriculum, the "Use" phase was implemented so students were given example code that they could run with provided data. Sections of the code were explained to the students so they could grasp what the code does. Then the curriculum gives students free rein to "Modify" the code. The curriculum provides suggestions for students to try - for example, they could modify code to analyze different variables within the dataset. Finally, the curriculum would ask questions that led students to "Create" their own code to analyze data and answer the questions.

Implementing Abstraction, Automation, and Analysis

Another pedagogical technique the curriculum used was the "Abstraction, Automation, and Analysis" (AAA) Computational Thinking process. This pedagogical approach is used to guide students to analyze a dataset by first breaking down a problem to its bare essentials (Abstraction), using R programming to execute the required steps (Automation), and finally, guiding students to evaluate their results (Analysis) (Lee et. al., 2011).

In the last lesson of the curriculum when students "Create" new content in order to answer the questions, the lesson asks questions where “Abstraction” is performed by the students to come up with a plan to answer the questions. Then the lesson gives them free rein to perform “Automation” of the data on a computer. Finally, students perform an “Analysis” of their computations and explain how this approach allowed them to answer the curriculum questions.

Selection of Datasets

The selection of datasets was a critical aspect of the project as it directly impacts the experience students gained from their analysis. Four datasets were chosen to provide high school students with experience with real-world data analysis and inference in various fields of biology. Datasets were chosen to fulfill the following criteria:

1. Represent biological data from the real world
2. Interesting to high school students
3. Not too complicated for high school students to analyze and interpret within a single lesson
4. Not so voluminous that analyzing it would exceed the constraints on computing power allowed for the free version of Posit Cloud.

5. Available from a reliable source
6. Available in a format conducive to analysis with R, such as Excel or CSV.

Internet search was used as a start in searching for datasets using phrases like “Biological datasets”, “biological datasets csv”, or “biological datasets xlsx”. Sources including Data Carpentry, Data.world, Brown University’s Library, Vanderbilt University’s Department of Biostatistics (hbiostat), Kaggle, U.S. Government data (Data.gov), and R’s internal datasets were reviewed to find appropriate datasets. While the biological content in many datasets was interesting, many were too complicated for a high school student to understand, did not clearly state their source, or did not exist in a convenient format. Four datasets were found that fulfilled the selection criteria.

The Trees dataset was selected for introducing a curriculum tailored to high school students exploring statistics, computer science, and biology. The dataset contained information about the weight of 31 trees over time observed by Ryan (Ryan et al., 1976). The data is accessible from R’s internal datasets. By exploring this dataset, students were introduced to simple descriptive statistical analysis to calculate the mean, standard deviation, minimum, median, and maximum of variables. This dataset was chosen for its simplicity of access when introducing students to R and Posit Cloud.

The Diabetes dataset was an ideal choice for lessons aimed at high school students studying statistics, computer science, and biology. The dataset contains medical profiles related to over 400 individuals that were collected as part of a study of African Americans in Virginia in 1997 (J P Willems et al., 1997). The data is accessible from Vanderbilt University’s Department of Biostatistics (hbiostat.org). Analyzing this data provides real-world experience using health data for medical applications (Batko 2022). The dataset includes various attributes of the recorded individuals that are not too complicated for high school students to understand including patient ids, cholesterol levels, glucose levels, height, and weight (Appendix A). Given the numerical and categorical variables the dataset includes, it provides opportunities for students to “Use” and “Modify” example code to analyze and interpret categorical and quantitative data using statistical techniques, consistent with the Massachusetts Curriculum Frameworks.

The Iris dataset was chosen for the curriculum due to its simplicity, well-defined structure, and relevance to multiple disciplines. The dataset contains measurements of 150 Irises, a type of flower, that were collected as part of an Iris species study in 1936 (Fisher, 1936). The data is accessible from R’s internal datasets. The dataset consists of measurements of Irises, including sepal and petal length and width. By exploring this dataset, students can perform correlation analysis to uncover the relationships between the various measurements (features) of iris flowers. Given the shape of the data, it provides an opportunity for students to interpret the correlation coefficient and distinguish between correlation and causation.

The Heart Disease dataset contains health records related to over 300 individuals and is focused on factors related to heart disease. The data was collected at the Cleveland Clinic in 1988 (Janosi et al., 1988). The data was made available through Kaggle, a website focused on machine learning education. Students can relate to the data's importance in heart health, and learn the significance of a data-driven approach to analyzing and understanding real-world health-related issues. Given the number of interesting variables the dataset included, it provided a good dataset for creating a lesson that implements the “Create” phase of the "Use-Modify-Create" pedagogical technique where students analyze and interpret data independently to answer curriculum questions.

These datasets offer valuable insights into critical areas of study, ranging from diabetes prevalence and management to plant species characterization and cardiovascular health. By analyzing these datasets, students can enhance their understanding of techniques used in computer science and statistics.

Making the Curriculum Accessible

The curriculum is intentionally designed to provide greater accessibility compared to standard computer science programs. Unlike conventional educational materials that often require expensive hardware and software, the curriculum leverages freely available open-source tools such as R. Additionally, the curriculum takes into account the varying technological resources available to students and avoids assuming access to personal computers or advanced software, opting instead for cloud-based resources such as Posit Cloud. This approach eliminates financial and technological barriers that could hinder student participation. Second, the curriculum is intentionally designed to accommodate students with varying levels of background knowledge, beginning with foundational concepts to establish a comfortable learning path. By doing so, it ensures that students with little or no previous exposure to statistics and computer science can comfortably engage with the material.

Additionally, the curriculum's adaptability stands out as an advantage. Unlike stand-alone computer science courses, this curriculum can be seamlessly incorporated into core mathematics or biology classes as the curriculum aligns with multiple statistics and computer science standards and biology practices. Eliminating the need for separate elective courses in computer science expands access to computational thinking to students with a more diverse set of interests and abilities.

Results

This IQP successfully developed an integrative statistics curriculum to complement computational thinking within the existing Bio-CS Bridge material. The curriculum was built with the Massachusetts Curriculum Framework, the "Use-Modify-Create" approach, and the “Abstract-Automate-Analyze” approach in mind, supporting the pursuit of two key objectives. Firstly, the curriculum aimed to enhance data literacy among high school students while showcasing the diverse applications of biology. Secondly, it prioritized creating a friendly, informative, and approachable learning environment to counteract imposter syndrome among students from diverse backgrounds.

The Massachusetts Curriculum frameworks were used as a guide in developing an interdisciplinary curriculum to foster learning statistics and computational concepts and to apply these concepts to problems in biology. Lessons were developed to teach students to represent and interpret data, consistent with the Mathematics framework (S-ID), and to guide them to make inferences and justify conclusions (S-IC). Lessons were also designed to be consistent with the Literacy and Computer Science framework standards related to Data, Computational Thinking and Programming Development. Finally, the lessons were designed to be consistent with the goals stated in the Science and Technology/Engineering framework.

R and Posit Cloud were chosen to fulfill the curriculum's objective of fostering an inclusive and accessible learning environment regardless of students' background or prior experience. They provide a powerful combination of statistical tools and user-friendly interfaces, enabling students to engage in data-driven exploration, analyze biological datasets, and build essential skills for future studies or careers in statistics, computer science, or biology. The use of

Posit Cloud makes the curriculum accessible through a web browser, thereby expanding access to those who do not have access to a personal computer.

The curriculum incorporates statistical and programming techniques to analyze biological datasets to achieve its goals of introducing high school students to the interdisciplinary nature of statistics, computer science, and biology while fostering a deeper understanding of biological processes through data-driven exploration. By applying statistical methods and programming techniques to analyze biological data, students gain practical experience in data analysis, manipulation, and visualization, preparing them for further studies or careers in statistics, computer science, or biology. The complete lessons, R scripts, and an answer key to lesson 5 are available as a zip file archived with this report. The lessons created include:

1. Getting Started with R using Posit Cloud
2. Analyzing and Understanding Distribution of Data
3. Visualizing Data and Testing Hypothesis
4. Correlation Analysis
5. Analyzing Heart Disease Data

This curriculum's lesson flow thoughtfully aligns with the “Use/Modify/Create” pedagogical approach and each lesson builds upon the knowledge gained in the previous lesson. Lesson 1 provides an introduction to programming in R and shows the student how to *Use* the cloud-based Posit environment. Lesson 2 builds on that foundation, by asking the students to *Use* provided code to generate plots to visualize data and then *Modify* the code and finally *Create* their own code to plot and visualize data.This lesson guides the student to use visualization techniques to develop an intuition for data distribution and relationships through hands-on exercises. It sets the stage for subsequent lessons by equipping students with the necessary skills to assess data patterns.

Lesson 3 builds on the foundation of data visualization and guides students to test their observations using the statistical techniques of hypothesis testing. Students are provided the code to *Use* for hypothesis testing, then encouraged to *Modify* it and then *Create* their own R code, thus fostering adaptability and critical thinking. Lesson 4 delves deeper into statistical analysis, leading students to *Use*, *Modify*, and then *Create* their own code to visualize data and test for relationships. Data visualization is used to look for possible patterns in the data and then students are led to test for statistically-significant relationships using correlation analysis. This enhances their analytical skills.

The curriculum culminates in Lesson 5, the final lesson representing the pinnacle of the "Use/Modify/Create" approach. Here, students engage in the *Create* phase on a larger scale, independently performing a comprehensive statistical analysis of a real-world dataset related to heart disease. This lesson employs the "Abstraction, Automation, and Analysis" technique, where students break down a statistical problem, create their own code, and extract meaningful insights from the data. The curriculum's sequential order ensures that each lesson builds upon the knowledge and skills gained in previous ones, effectively leading students toward the final lesson, where they independently apply all they have learned to a real-world scenario. This flow of lessons offers a structured and progressively challenging learning journey, embodying the principles of the Use/Modify/Create pedagogical approach.

Lesson Descriptions

Lesson #1: Getting Started with R using Posit Cloud

Learning outcomes:

Upon completing this lesson, students:

1. Have created a Posit Cloud account for use in this and subsequent lessons
2. Are able to describe what R and Posit Cloud are and how they are useful
3. Are able to describe what a dataset, data frame, and vector are
4. Are able to refer to R help modules to learn more about R
5. Are able to navigate the Posit Cloud user interface and perform simple statistical analyses using R including calculating mean and standard deviation

Dataset: Trees Dataset

Standards addressed:

Massachusetts Curriculum Framework

1. Digital Literacy and Computer Science Framework
   a. 9-12.CT.d 9 Select and employ an appropriate component or library to facilitate programming solutions [e.g., turtle, Global Positioning System (GPS), statistics library]

R commands & concepts:

Command	Description
Print(“ “)	Displays the specified text or value inside the parentheses
<-	Assigns values to variables in R
#	Adds comment in R
+	Perform addition
-	Perform subtraction
*	Performs multiplication
/	Performs division
^	Performs exponent
data()	Lists available built-in datasets in R

help()	Opens the help system in R, providing information about functions and packages.
head()	Displays first few rows of a dataset
nrow()	Returns the number of rows in a data frame in R
?	To get help about a specific function or topic in R
mean()	Computes the mean of a numeric vector in R
sd()	Computes the standard deviation of a numeric vector in R
summary()	Provides a summary of a data frame, showing descriptive statistics for each variable

- Concepts
  - R
  - Posit Cloud
  - Creating a Posit Cloud account
  - Workspace
  - Working directory
    - Source panel
    - Console panel
    - Environment panel
  - Script
  - Variable
  - Comments
  - Dataset
  - Dataframe
  - Vector
  - Calculating summary statistics

In this lesson, students walk through the process of creating a Posit Cloud account and setting up their workspace. The lesson introduces students to the concepts of datasets, dataframes, and vectors. Moreover, the lesson introduces basic programming and statistical concepts to help students become familiar with executing code in the R and Posit Cloud environment. Students also learn to access R's help command to learn more. Finally, students learn the importance of using a script so that they can reproduce their results, as well as comments to document their code so that others can easily interpret it.

Lesson #2: Analyzing and Understanding Distribution of Data

Learning outcomes:

Upon completing this lesson, students:

1. Are able to load a dataset into Posit Cloud
2. Are able to visualize data using histograms and interpret what the plots convey
3. Are able to describe what normality is
4. Are able to identify if data is normally distributed or skewed
5. Are able to analyze and interpret the results of a normality test

Dataset: Diabetes Dataset

Standards addressed:

Massachusetts Curriculum Framework

Mathematics

1. S-ID A.1. Represent data with plots on the real number line (dot plots, histograms, and box plots)
2. S-ID A.3. Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers)
3. S-ID A.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

Digital Literacy and Computer Science Framework

1. 9-12.CT.c 3 Create, evaluate, and revise data visualization for communication and knowledge
2. 9-12.CT.d 9 Select and employ an appropriate component or library to facilitate programming solutions [e.g., turtle, Global Positioning System (GPS), statistics library]

Biology practices addressed:

Next Generation Science Standards

Biology

1. Analyzing and interpreting data
2. Using Mathematics and Computational Thinking

R commands & concepts:

Command	**Description**
View()	Displaying the contents of an object like a data frame into a different tab on R for easy exploration.
$	Used to access individual columns (variables) of a data frame in R
ifelse	Conditional function in R that returns a value based on a specified condition being true or false
ggplot()	Initializes a new ggplot object, allowing for the creation of data visualizations using the ggplot2 package in R
install.packages()	Used to download and install R packages from the internet
library()	Loads a specific R package into the current R script
geom_histogram	ggplot2 function that creates a histogram in R
shapiro.test	Performs the Shapiro-Wilk normality test in R to check if a dataset follows a normal distribution

- Concepts
    - Loading a dataset from an Excel file
    - Creating and saving a R script
    - Installing and loading R packages

The dataset used in this lesson is called the diabetes dataset. The variables of the dataset represent different health indicators such as cholesterol level, glucose level, and blood pressure and patient demographic information such as sex, age, and weight.

Understanding how data is distributed is crucial in biology as it allows scientists to gain insights into the underlying patterns in biological processes, leading to more informed and accurate conclusions in their studies. This lesson guides students to represent and visualize the distribution of variables in a dataset using histograms and density plots. It then introduces students to the concepts of normal distribution and skewness in data distribution. Students are guided to use R to create a histogram of the weights of patients to see if the data appears normally distributed or skewed. In preparation for future lessons, students create and derive two categorical variables representing if patients had hypertension or diabetes based on numerical data in the dataset.

Lesson #3: Visualizing Data and Testing Hypotheses

Learning outcomes:

Upon completing this lesson, students:

1. Are able to describe what median, quartiles, and outliers of a dataset are
2. Know when and how to use a scatter plot and a box plot
3. Are able to identify if there is a statistically significant difference between two groups of data
4. Know how to test a hypothesis using a t-test
5. Are able to analyze the results of a t-test and either accept or reject the null hypothesis

Dataset: Diabetes Dataset

Standards addressed:

Massachusetts Curriculum Framework

Mathematics

1. S-ID A.1. Represent data with plots on the real number line (dot plots, histograms, and box plots)
2. S-ID A.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. S-ID A.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. S-IC B.3. Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each
5. S-IC B.6. Evaluate reports based on data

Digital Literacy and Computer Science Framework

1. 9-12.CT.c 3 Create, evaluate, and revise data visualization for communication and knowledge
2. 9-12.CT.c 4 Analyze a complex data set to answer a question or test a hypothesis (e.g., analyze a large set of weather or financial data to predict future patterns)
3. 9-12.CT.d 9 Select and employ an appropriate component or library to facilitate programming solutions [e.g., turtle, Global Positioning System (GPS), statistics library].

Biology practices addressed:

Next Generation Science Standards

Biology

1. Analyzing and interpreting data
2. Using Mathematics and Computational Thinking

R commands & concepts:

Command	Description
geom_point()	A ggplot2 function that creates a scatter plot, representing data points as individual points on the plot
geom_boxplot()	A ggplot2 function that generates a box plot, representing the distribution of a numeric variable through its quartiles and outliers
t.test()	Computes a two-sample t-test in R, comparing the means of two groups to determine if there is a significant difference between them

- Concepts:
  - Adding a column to a dataset based on another column
  - Creating box and scatterplots
  - Performing the t-test

Lesson 3 uses the same diabetes dataset from the previous lesson. This lesson expands on representing and visualizing the distribution of variables in a dataset. Students use box plots to visualize the distribution of data, including seeing the median, quartiles, and outliers. This can help students understand the central tendency and spread of biological measurements, enabling them to identify potential anomalies or unique characteristics in their data that might signify abnormal biological phenomena, such as environmental influences that could significantly impact a population's behavior. Students were guided to use R to create and interpret box plots that compare data distributions of the weights between patients with and without diabetes. Similar analysis can be used in other areas of biology, such as comparing the gene expression levels of different groups of organisms, comparing risk factors across different age groups, etc.

Along with box plots, this lesson adds another method for visualizing the relationship between two different variables using scatter plots. Students create and interpret scatter plots to understand the relationship between patients' heights and weights. In biology, scatter plots can be used to investigate various relationships, for example, the concentration of a chemical compound in the soil and the growth rate of plants, helping researchers understand how soil composition affects plant development and productivity.

Lesson 3 then introduces students to the concept of hypothesis testing for statistically significant differences between two groups of data. Students are guided to create null and alternative hypotheses regarding whether individual weights differ between those with and without diabetes. They then test their hypotheses by performing and interpreting a t-test.

Lesson #4: Correlation Analysis

Learning outcomes:

Upon completing this lesson, students:

1. Are able to describe what correlation is
2. Are able to describe and interpret what the correlation coefficient is
3. Are able to describe the difference between correlation and causation
4. Are able to test if the relationship between 2 variables is significant

Dataset: Iris Dataset

Standards addressed:

Massachusetts Curriculum Framework

Mathematics

1. S-ID B.6. Represent data on two quantitative variables on a scatter plot, and describe how the variables are related
2. S-IC B.3. Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each
3. S-IC B.6. Evaluate reports based on data

Digital Literacy and Computer Science Framework

1. 9-12.CT.c 3 Create, evaluate, and revise data visualization for communication and knowledge
2. 9-12.CT.c 4 Analyze a complex data set to answer a question or test a hypothesis (e.g., analyze a large set of weather or financial data to predict future patterns)

Biology practices addressed:

Next Generation Science Standards

Biology

1. Analyzing and interpreting data
2. Using Mathematics and Computational Thinking

R commands & concepts:

Command	Description
names()	Retrieves the names of the elements in an R object like a data frame.
cor()	Calculates the correlation matrix for a set of numeric variables, showing the linear relationships between pairs of variables in a

	dataset
[:]	Used for subsetting data in R, allowing the selection of specific rows and columns from a data frame
cor.test	Perform a statistical test of correlation in R, determining if there is a significant correlation between two numeric variables
pairs()	Creates a scatterplot matrix in R, displaying scatter plots of multiple numeric variables against each other to explore their relationships

The dataset used in this lesson is the Iris dataset, which contains measurements of different characteristics of Iris flowers. Understanding correlation and the correlation coefficient is crucial in biology as it allows scientists to explore relationships between variables and distinguish between correlation and causation, leading to more informed and accurate conclusions in their studies. This lesson guides students to represent and visualize the relationship between two quantitative variables in the Iris dataset using scatter plots. It then introduces students to the concept of correlation and how to calculate the correlation coefficient. Students are guided to use R to create a scatter plot of the sepal length and petal length of Iris flowers and calculate the correlation coefficient to determine the strength and direction of the relationship between the two variables. In preparation for a future lesson, students learn to interpret the correlation coefficient and understand its significance in analyzing biological data.

Lesson #5: Analyzing Heart Disease Data

Learning outcomes:

Upon completing this lesson, students:

1. Are able to plot and visualize the variables in dataset
2. Are able to generate and test hypotheses
3. Are able to determine if variables of a dataset are correlated

Dataset: Heart Disease Dataset

Standards addressed:

Massachusetts Curriculum Framework

Mathematics

1. S-ID A.1. Represent data with plots on the real number line (dot plots, histograms, and box plots)
2. S-ID A.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. S-ID A.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. S-ID A.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
5. S-ID B.6. Represent data on two quantitative variables on a scatter plot, and describe how the variables are related
6. S-IC B.3. Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each
7. S-IC B.6. Evaluate reports based on data

Digital Literacy and Computer Science Framework

1. 9-12.CT.c 3 Create, evaluate, and revise data visualization for communication and knowledge
2. 9-12.CT.c 4 Analyze a complex data set to answer a question or test a hypothesis (e.g., analyze a large set of weather or financial data to predict future patterns)
3. 9-12.CT.d 9 Select and employ an appropriate component or library to facilitate programming solutions [e.g., turtle, Global Positioning System (GPS), statistics library]

Biology practices addressed:

Next Generation Science Standards

Biology

1. Asking questions (for science) and defining problems (for engineering)
2. Planning and carrying out investigations
3. Analyzing and interpreting data
4. Using Mathematics and Computational Thinking

The dataset used in this lesson is called the heart disease dataset. The variables of the dataset represent different health indicators such as cholesterol level, blood pressure, and maximum heart rate and patient demographic information such as sex and age.

This lesson reinforces the learning that students gained from prior lessons. Previous lessons implement the "Use-Modify-Create" phases with small versions of the “Create” phase. This lesson represents the "Create" phase on a larger scale where students are tasked with performing a comprehensive statistical analysis of the heart disease dataset by creating their own code without being given any code samples. This lesson also employs the "Abstraction, Automation, and Analysis” pedagogical technique. In the process, a statistical problem is stated and students break it down into steps they will complete (Abstraction). Next, they write and execute their own code to solve the problem (Automation). Finally, students extract meaningful insights from the dataset to solve the statistical problem (Analysis). This lesson can be found in Appendix B.

Conclusion

The Bio-CS Bridge R unit was thoughtfully designed with the aim of integrating computer science, data visualization, and biology to create a holistic learning experience for students. While developing the curriculum, we encountered various challenges, including limitations in the free version of Posit Cloud, which had a limitation in processing and memory usage. Despite these obstacles, time was invested in refining and aligning the materials to ensure a positive impact on student learning in these interdisciplinary fields. The finalized curriculum aims to empower students with valuable skills in computer science, data visualization, and biology, fostering their confidence in tackling real-world problems and finding connections between these disciplines. In today's technology-driven world, it is crucial for students to have exposure to statistics and computer science, and the Bio-CS Bridge R curriculum strives to provide a solid foundation in these areas, preparing students for future exploration and success.

Future Work

While this curriculum provides a robust foundation for integrating statistics and computer science into high school biology education using Posit Cloud, there are areas for potential future development and expansion. Firstly, considering Posit Cloud's accessibility, it is essential to acknowledge the limitations of the free version. To enhance accessibility for all students, exploring possibilities for more extended access or alternative platforms could be valuable.

Additionally, this curriculum primarily focuses on R for statistical analysis and data visualization. Future work could delve deeper into teaching advanced R techniques that are particularly relevant to high school biology, such as algebra, probability, and calculus. These skills can empower students to tackle more intricate biological questions and research projects.

Further, there is room to extend the curriculum's application in biology. Future work could apply the statistical analysis techniques learned in this curriculum in areas such as genetics, biochemistry, and other biology subdisciplines, broadening their understanding of the diverse applications of statistics in the biological sciences. These suggestions for future work aim to continually enhance the curriculum's effectiveness in preparing high school students for careers and further education in biology and related fields.

References

Attwood T. K., Blackford S, Brazas M. D., Davies A, Schneider M. V. (2019). A global perspective on evolving bioinformatics and data science training needs. Briefings in Bioinformatics, 20(2), 398–404. https://doi.org/10.1093/bib/bbx100

Batko K, Ślęzak A. (2022). The use of Big Data Analytics in healthcare. Journal of big data, 9(1), 3. https://doi.org/10.1186/s40537-021-00553-4

BBSRC and MRC review of vulnerable skills and capabilities. (2017). https://www.ukri.org/publications/bbsrc-and-mrc-review-of-vulnerable-skills-and-capabilities/

Byrne, C. (2023). Building the Bio-CS Bridge: R and Netlogo. digitalwpi.wpi.edu/concern/student_works/rj4307977?locale=en

Cateté V, Mott B, Boyer K, Lytle N, Dong Y, Boulden D, Akram B, Houchins J, Barnes T, Wiebe E, Lester J. (2018). Infusing computational thinking into middle grade science classrooms: lessons learned. 1-6. https://doi.org/10.1145/3265757.3265778

Code.org, CSTA, & ECEP Alliance. (2022). 2022 State of Computer Science Education: Understanding Our National Imperative. https://advocacy.code.org/stateofcs

Crabtree L. M., Richardson S. C., Lewis C. W. (2019). The Gifted Gap, STEM Education, and Economic Immobility. Journal of Advanced Academics, 30(2), 203–231. https://doi.org/10.1177/1932202X19829749

Crowder M, Hand D. (1990). Analysis of Repeated Measures, Chapman and Hall (example 5.3)

6.05 Comparisons of data sets | Math I Math | Massachusetts Math 1 - 2020 Edition. Mathspace. https://mathspace.co/textbooks/syllabuses/Syllabus-939/topics/Topic-19719/subtopics/Subtopic-262287/

Fisher R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, 7(2), 179-188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Gould R, Machado S, Ong C, Johnson T, Molyneux J, Nolen S, Tangmunarunkit H, Trusela L, Zanontian L. (2016). Teaching Data Science to Secondary Students: The Mobilize Introduction to Data Science Curriculum. https://iase-web.org/documents/papers/rt2016/Gould.pdf?1482484533

Hand D, Crowder M. (1996), Practical Longitudinal Data Analysis, Chapman and Hall (table A.2)

Heinzman E. (2022). "I love math only if it's coding": A case study of student experiences in an introduction to data science course. Statistics Education Research Journal, 21(2), 5–5. https://doi.org/10.52041/serj.v21i2.43

Janosi A, Steinbrunn W, Pfisterer M, Detrano R. (1988). Heart Disease. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.

Kricorian K, Seu M, Lopez D, Ureta E, Equils O. (2020). Factors influencing participation of underrepresented students in STEM fields: matched mentors and mindsets. IJ Stem Ed 7(16). https://doi.org/10.1186/s40594-020-00219-2

LaMar T. (2023). The importance and emergence of K-12 data science - kappanonline.org. https://kappanonline.org/math-importance-emergence-k12-data-science-lamar-boaler/

Lytle, N., Cateté, V., Boulden, D., Dong, Y., Houchins, J., Milliken, A., Isvik, A., Bounajim, D., Wiebe, E., & Barnes, T. (2019). Use, Modify, Create: Comparing Computational

Thinking Lesson Progressions for STEM Classes. Innovation and Technology in Computer Science Education, 395–401.https://doi.org/10.1145/3304221.3319786

Massachusetts Department of Elementary and Secondary Education. (2016). Digital Literacy and Computer Science, Grades Kindergarten to 12, Massachusetts Curriculum Framework - 2016. https://www.doe.mass.edu/frameworks/dlcs.pdf

Massachusetts Department of Elementary and Secondary Education. (2017). Mathematics, Grades Kindergarten to 12, Massachusetts Curriculum Framework - 2017. https://www.doe.mass.edu/frameworks/math/2017-06.pdf

Massachusetts Department of Elementary and Secondary Education. (2016). Science and Technology / Engineering, Grades Pre-Kindergarten to 12, Massachusetts Curriculum Framework - 2016. https://www.doe.mass.edu/frameworks/scitech/2016-04.pdf

Palid O, Cashdollar S, Deangelo S, Chu C, Bates M. (2023). Inclusion in practice: a systematic review of diversity-focused STEM programming in the United States. International Journal of STEM Education, 10(2). https://doi.org/10.1186/s40594-022-00387-3

Standard Deviation. Bootstrapworld.org. https://bootstrapworld.org/materials/fall2022/en-us/lessons/standard-deviation/index.shtml

Strode P, Brokaw A. (2015). Using BioInteractive Resources to Teach Mathematics and Statistics in Biology Using BioInteractive Resources to Teach Mathematics and Statistics in Biology. HHMI / Biointeractive. https://www.biointeractive.org/sites/default/files/media/file/2019-05/Statistics-Teacher-Guide.pdf

The Bio-CS Bridge. (2023). WPI. https://biocsbridge.wpi.edu/website/home

UCLA. (2022). Introduction to Data Science Curriculum | Introduction to Data Science. www.ucladatascienceed.org/introduction-to-data-science-curriculum

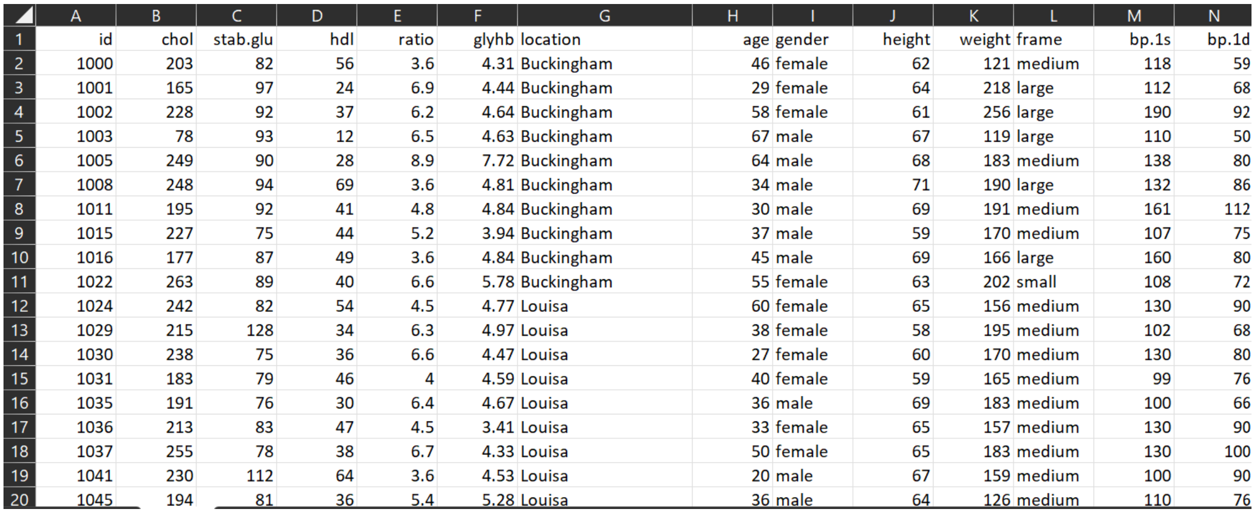
U.S. Department of Education. (2021). Catalyzing a New Field: Data Science in K-12 Education. https://ies.ed.gov/ncer/whatsnew/techworkinggroup/pdf/DataScienceTWG.pdf

Willems J. P., Saunders J. T., Hunt D. E., Schorling J. B. (1997). Prevalence of coronary heart disease risk factors among rural blacks: a community-based study. Southern medical journal, 90(8), 814–820. https://doi.org/10.1097/00007611-199708000-00008

Willems S. M., Abeln S, Feenstra K. A., de Bree R, van der Poel E. F., Baatenburg de Jong R. J., Heringa J, van den Brekel M. W. M. (2019). The potential use of big data in oncology. Oral Oncology, 98, 8–12. https://doi.org/10.1016/j.oraloncology.2019.09.003

Appendix

Appendix A: Example of A Dataset



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	id	chol	stab.glu	hdl	ratio	glyhb	location	age	gender	height	weight	frame	bp.1s	bp.1d
2	1000	203	82	56	3.6	4.31	Buckingham	46	female	62	121	medium	118	59
3	1001	165	97	24	6.9	4.44	Buckingham	29	female	64	218	large	112	68
4	1002	228	92	37	6.2	4.64	Buckingham	58	female	61	256	large	190	92
5	1003	78	93	12	6.5	4.63	Buckingham	67	male	67	119	large	110	50
6	1005	249	90	28	8.9	7.72	Buckingham	64	male	68	183	medium	138	80
7	1008	248	94	69	3.6	4.81	Buckingham	34	male	71	190	large	132	86
8	1011	195	92	41	4.8	4.84	Buckingham	30	male	69	191	medium	161	112
9	1015	227	75	44	5.2	3.94	Buckingham	37	male	59	170	medium	107	75
10	1016	177	87	49	3.6	4.84	Buckingham	45	male	69	166	large	160	80
11	1022	263	89	40	6.6	5.78	Buckingham	55	female	63	202	small	108	72
12	1024	242	82	54	4.5	4.77	Louisa	60	female	65	156	medium	130	90
13	1029	215	128	34	6.3	4.97	Louisa	38	female	58	195	medium	102	68
14	1030	238	75	36	6.6	4.47	Louisa	27	female	60	170	medium	130	80
15	1031	183	79	46	4	4.59	Louisa	40	female	59	165	medium	99	76
16	1035	191	76	30	6.4	4.67	Louisa	36	male	69	183	medium	100	66
17	1036	213	83	47	4.5	3.41	Louisa	33	female	65	157	medium	130	90
18	1037	255	78	38	6.7	4.33	Louisa	50	female	65	183	medium	130	100
19	1041	230	112	64	3.6	4.53	Louisa	20	male	67	159	medium	100	90
20	1045	194	81	36	5.4	5.28	Louisa	36	male	64	126	medium	110	76

Appendix A shows the first 20 lines of the Diabetes Dataset.

Appendix B: Lesson #5

Analyzing Heart Disease Dataset

Name ________________________________ Date: ___________ Period: ______

For each activity below:

✅ means to complete this task

✏️ means to write an answer here

Investigating Heart Disease

With access to a comprehensive heart disease dataset, you are part of a team that is investigating the relationships between various risk factors and the likelihood of a heart attack. Your team has created a dataset and the variables it includes are:

- age - Age of individual in years
- sex - Gender of the individual
  - 1 = Male
  - 0 = Female
- cp - Chest pain type from 4 different types
  - 1 = typical angina
  - 2 = atypical angina
  - 3 = non-anginal pain
  - 4 = asymptomatic
- trestbps - Resting blood pressure in mm Hg on admission to the hospital
- chol - Serum cholesterol measured in mg/dl
- fbs - Fasting systolic blood pressure > 120 mg/d.
  - 0 = False
  - 1 = True.
- restecg - Resting electrocardiographic results
  - 0 = normal
  - 1 = abnormal

  - 2 = probable or definite hypertrophy
- thalach - Maximum heart rate achieved measured in beats per minute
- exang - Exercise-induced angina
  - 1 = yes
  - 0 = no
- oldpeak - ST depression induced by exercise relative to rest. Does not have any levels.
- slope - Slope of peak exercise ST segment
  - 1 = upsloping
  - 2 = flat
  - 3 = downsloping.
- ca - Number of major vessels colored by flouroscopy, measured from 0 to 3.
- thal - Type of Thalassemia, a inherited blood disorder
  - 0 = normal
  - 1 = fixed defect
  - 2 = reversable defect.
- target - Likelihood of heart attack
  - 0 = low risk of heart attack
  - 1 = high risk of heart attack.

You are tasked with studying 2 relationships:

1. **The relationship between serum cholesterol level and resting blood pressure.**
2. **The relationship between serum cholesterol level and heart disease risk.**

Activity #1: Outline your analysis

a. ✏️ Think about the task of studying if there is an association **between serum cholesterol level and resting blood pressure**. Using the knowledge from previous lessons, break down the task. Write down the steps you will take.

Activity #2: Use R to Perform Your Analysis

1. ✅ Save and load the Heart Disease dataset provided by your teacher into Posit Cloud.

2. ✅ Plot serum cholesterol level and resting blood pressure.

    a. ✏️ Write down your observations.

3. ✅ Based on what you saw in step 2, test to see if there exists a meaningful association between serum cholesterol level and resting blood pressure. Use the techniques you learned in prior lessons.

    b. ✏️ Write down your observations.

Activity #3: Analyze and interpret your tests

a. ✏️ Analyze the results of the test(s) you performed. What did you conclude? Justify your conclusion(s).

Activity #4: Outline your analysis

a. ✏️ Think about the task of studying if there is a relationship **between serum cholesterol level and heart disease risk**. Using the knowledge from previous lessons, break down the task. Write down the steps you will take.
Hint: Consider the differences between those who have a high risk of heart attack and those who have a low risk.

Activity #5: Use R to Perform Your Analysis

1. ✅ Plot and compare the serum cholesterol level of those with heart disease risk and those without.
    - Hint: You may need to create a new categorical variable based on the numerical “target” variable. If you are having trouble doing this, ask the teacher for help

    a. ✏️ Write down your observations.

2. ✅ Based on what you saw in step 1, test to see if there exists a meaningful difference in serum cholesterol levels between those with heart disease risk and those without. Use the techniques you learned in prior lessons.
    - State your null and alternative hypothesis before you start.

    b. ✏️ Write down your observations.

3. ✅ Redo Activities #4-5 but use other variables that may be related to heart disease risk.

Activity #6: Analyze and interpret your tests

a. ✏️ Analyze the results of the test(s) you performed. What did you conclude? Justify your conclusion(s).

b. ✏️ How do you think the knowledge and skills gained from this activity can be applied to real-world scenarios in the fields of medicine, research, or data analysis?

c. ✏️ Overall, how has this activity helped you develop a deeper appreciation for the importance of analyzing and interpreting data in the context of heart disease research? What other areas do you think data analysis could be valuable in the field of biology and medicine?