Using RStudio to Engage School Students in Data Science

Ahmad M. Alhammouri, *Jacksonville State University*

Rami Al-Ouran, *Al-Hussein Technical University*

Abstract: *In this article, we provide an activity on how to engage high school students in data science projects. We discuss the data science framework before we present and solve a data science activity using the RStudio development environment. The code and data used in the activity are provided through easy-to-use downloadable online links.*

Keywords: *data science, statistics, quantitative reasoning, RStudio, statistical problem-solving*

Introduction

In the last decade, data science has become increasingly important in our daily lives and careers. Data are being generated at a rapid pace (i.e., big data), and now, more than ever, we must be equipped with the right tools and analytical abilities to digest and analyze the data to make sound decisions and conclusions. Large amounts of data are offered by countless online platforms, such as social media, computer systems, and networks (van der Aalst, 2016). In data science, formal and informal data are used to make decisions and predictions concerning real-world scenarios, such as the COVID-19 pandemic and climate change. Individuals must be equipped with the right skills to read such data and make data-driven decisions.

School education is the cornerstone in preparing the upcoming generations to be data literate, which encompasses collecting, reading, analyzing, and understanding data to make real-world decisions and predictions. However, school education in general and mathematics education specifically have not been providing enough opportunities to make students data literate (The Ohio Department of Education [ODE], 2020). At the high school level, data science is usually addressed in advanced placement (AP) statistics, which primarily focuses on explaining the data rather than using (or applying) the data for meaningful connections to real-world applications (Chen, 2020).

To address this issue, the ODE is developing a year-long course called Data Science Foundations (DSF), which can be considered equivalent to an Algebra 2 course. According to the ODE, the course aims to teach "students to reason with and think critically about data in all forms." The course will be piloted during the 2021-2022 school year. This course will be launched and evaluated in two phases (i.e., 2022-2023 and 2023-2024 school years). Professional development programs will be offered to teachers who will teach the DSF course starting in the summer of 2022.

In this article, we aim to contribute to teacher preparation efforts by presenting and solving a data science activity using the R programming language and RStudio to advance teacher knowledge and classroom enactment of the data science process.

What is Data Science?

When students engage in a data science activity, they collect and analyze data to make decisions concerning real-world phenomena. The ODE indicates that data science encompasses three major themes: quantitative reasoning, computer science, and probability and statistics (see Figure 1).

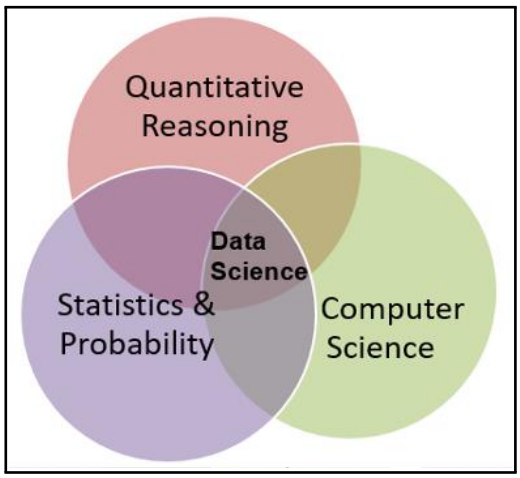

Figure 1: *Data Science themes (ODE, 2020).*

In 2020, Pre-K–12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II): A Framework for Statistics and Data Science Education was released by the National Council of Teachers of Mathematics [NCTM] (Bargagliotti, Franklin, Arnold, Gould, Johnson, Perez, and Spangler, 2020). Bargagliotti et al. (2020) place statistics at the heart of data science. They presented a framework for the statistical problem-solving process. The framework "supports all students as they learn to appreciate the vital role of statistical reasoning and data science and acquire the essential life skill of data literacy" (Bargagliotti et al., 2020, p. 3).

Figure 2 shows the statistical problem-solving process framework presented in the GAISE II report. The process starts with students developing a problem statement that is written in the form of a statistical question about a real-world phenomenon. This question anticipates variability, indicates the group/population of the study, and requires descriptive data (Bargagliotti et al., 2020).

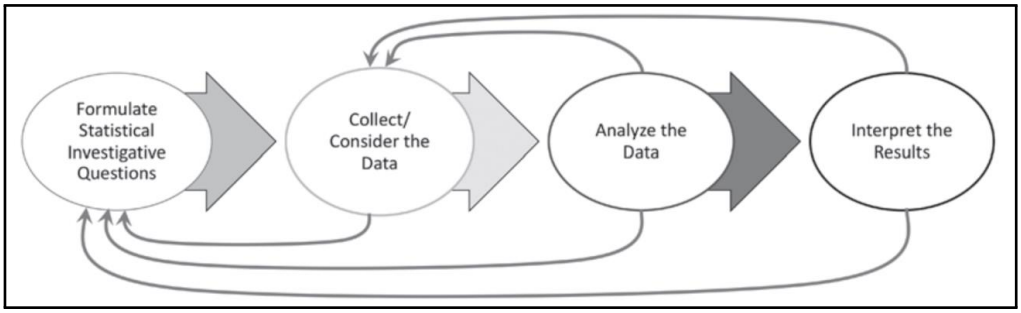

Figure 2: *Statistical problem-solving process framework (Bargagliotti et al., 2020, p. 13).*

Then, students collect or consider data. At this stage, students may develop an instrument to collect descriptive data required by the research question, collect the data from the group/population of the study, and then organize and categorize the data. Next, students analyze the collected data to understand its variability. To analyze the data, statistical tests might be conducted, such as calculating the central tendency and representing the data using charts. Finally, students use statistical evidence to answer the statistical question that was defined at the beginning of the process

(Bargagliotti et al., 2020). At this stage, students may write a report to conclude, interpret, and present their findings regarding the research question.

The statistical problem-solving process framework presented in Figure 2 forms the backbone of the data science process, which addresses probability and statistics. The other two themes (i.e., quantitative reasoning and computer science) are built around the statistical problem-solving process. According to Bargagliotti et al. (2020), the statistical problem-solving framework can be implemented within three levels (i.e., A, B, C). Students should experience Level A, Level B, and then Level C. As they move from Level A to Level C, students are expected to engage in more reasoning and critical thinking, which leads to the second theme of the data science framework (i.e., quantitative reasoning). The National Governors Association Center for Best Practices and the Council of Chief State School Officers (NGA Center & CCSSO, 2010) indicated: Quantitative reasoning entails habits of creating a coherent representation of the problem at hand; considering the units involved; attending to the meaning of quantities, not just how to compute them; and knowing and flexibly using different properties of operations and objects. (p. 6)

While the students are involved in the statistical problem-solving process, they can be engaged in open-ended and real-world problems, critical thinking, reasoning, and justification that address the quantitative reasoning theme of data science.

The third theme in the data science framework is computer science. Similar to quantitative reasoning, computer science can be used to enhance students' engagement in the statistical problem-solving framework. For example, students can use computer science programs to sort, analyze, and represent the collected data during the statistical problem-solving process. In this article, we will use R and RStudio.

What is R and RStudio?

R is a programming language mainly used for statistical analysis. R provides valuable utilities to solve statistical problems and flexible approaches to visualize data and results. RStudio is an Integrated Development Environment (IDE) to use and run the R language. As shown in Figure 3, RStudio has four main panels: (1) the **programming panel** where the code is inserted; (2) the **environment panel** that displays the output of running the R code; (3) the **console panel** that lists the variables defined in the code block; and (4) the **plot panel** that displays plots produced by the R code.

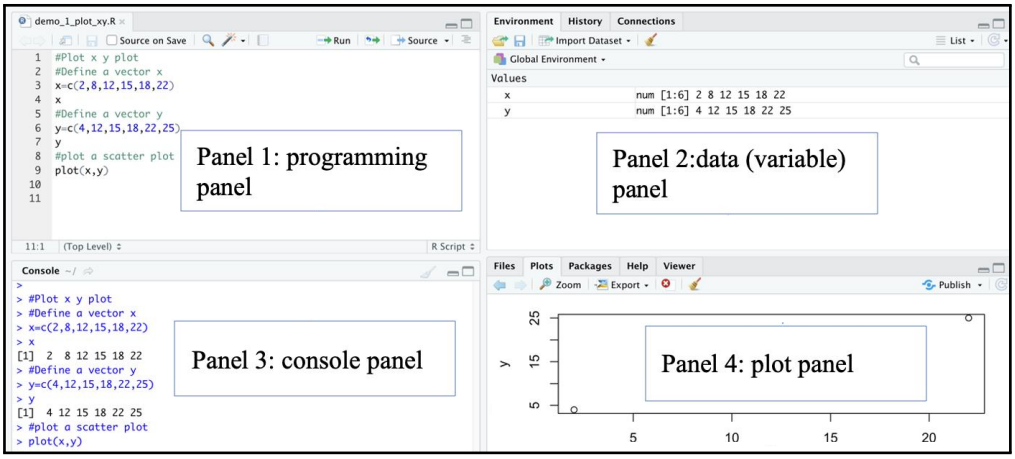

Figure 3: *The RStudio Integrated Development Environment (IDE).*

RStudio enables users to write, run, and display code results all in one platform. With the increasing importance of data science, R has emerged as a popular programming language to efficiently analyze data.

R Markdown

Integrated within RStudio is R Markdown. R Markdown is a markup language that enables users to organize code in a report style to easily write and read R code and plot intermediate results (Baumer, Cetinkaya-Rundel, Bray, Loi, & Horton, 2014). The code is organized in blocks or chunks, and each block can be executed individually. Below each code block, data and plots can be displayed, and users can create HTML and PDF reports. This feature allows students to easily submit their reports and share and compare analyses with other students in the class. Figure 4 shows an example of using R Markdown in RStudio.

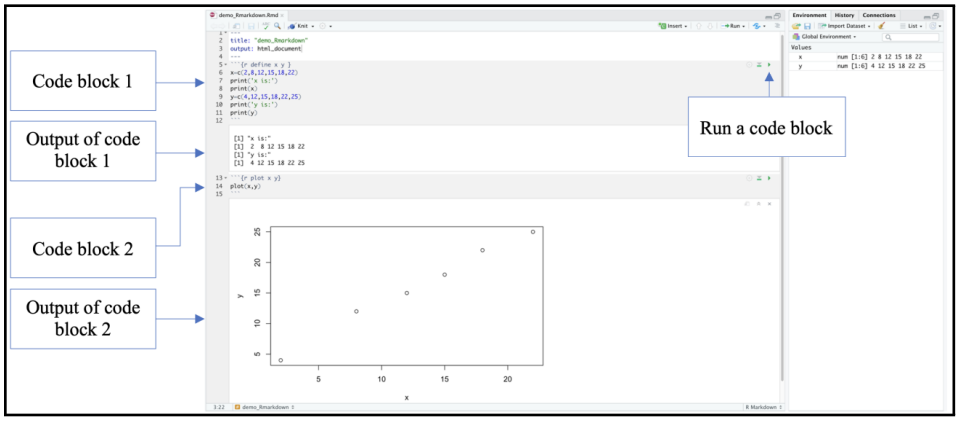

Figure 4: *R markdown example in RStudio.*

Installation of the R programming language and RStudio

To utilize RStudio, we need to first install the R programming language available at https://cran.rstudio.com/. Next, we install the RStudio IDE available at https://www.rstudio.com/products/rstudio/download. Both R and RStudio can be installed on different operating systems (Windows, Mac, Linux) with ease.

Commuting to Work: Box Diagrams, Central Direction, and Outliers

The commuting to work activity was adopted from the United States Census Bureau website (https://www.census.gov/schools). The activity aims to address the following learning concepts: outliers and their effect on the mean, median, data distribution, and accuracy of measures of the central tendency.

The activity is presented using 26 questions to be answered by the students (see https://www.census.gov/programs-surveys/sis/activities/math/commuting.html). We extracted a set of questions that focus on the main concepts to incorporate data science gradually and effectively. We reduced the length of the activity such that students could finish the activity in a timely manner while still focusing on the main concepts, which, as a result, would maintain the cognitive demand of the task (Stein, Smith, Henningsen, & Silver, 2009). We provide the activity as a student handout and we provide handouts for teachers that include the solutions. Next, we will present the activity with suggested answers.

The Activity: *Commuting to Work*

After ensuring that the R programming language and RStudio are installed as described above, proceed with the following steps:

1. Download the following files to your computer to the same folder:
   - Bike_Commute_data.csv: This is the input data for the activity. The data consist of two columns: U.S. state and number of bicycle commuters per state. A comma-separated value (csv) file is a text file in which file elements are separated by commas. We used a .csv file instead of an Excel file so that no extra R libraries were required to be installed. The file is available at https://www.dropbox.com/s/2npz2y5orkd9qjw/Bike_Commute_data.csv
   - Biking_example.Rmd: This is the R Markdown file that includes the R code. The file is divided into code blocks and is available at https://www.dropbox.com/s/sd0l4hfub54tpy0/Biking_example.Rmd
2. Open the RStudio program by clicking on the program's icon and the RStudio panel will be displayed as shown below:

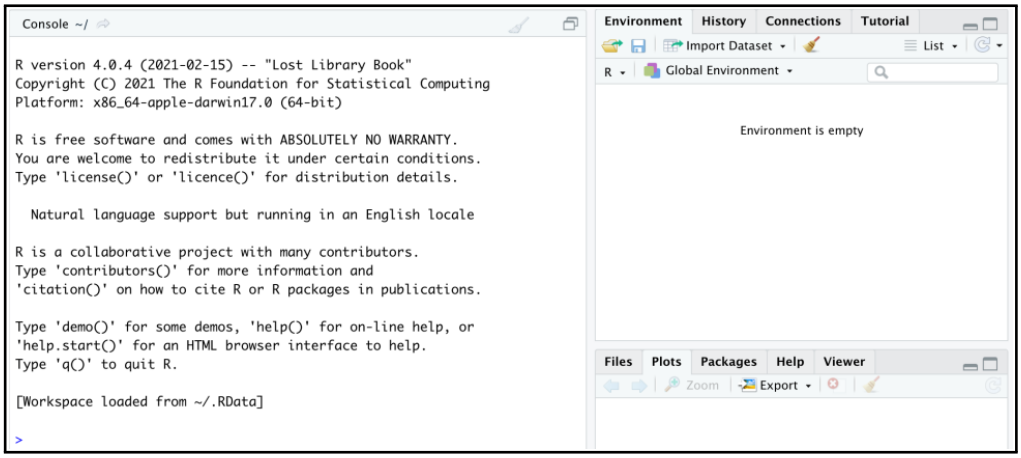

Figure 5: *RStudio panel.*

3. Open the R Markdown file using "Open File" to open the file Biking_example.Rmd.

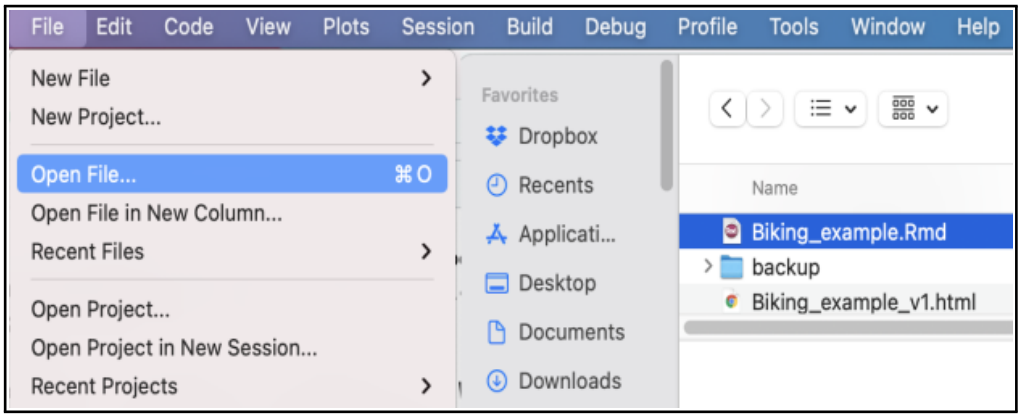

Figure 6: *Opening the biking example R file.*

4. From the session menu in RStudio, select “Set Working Directory” to “Source File Location.’

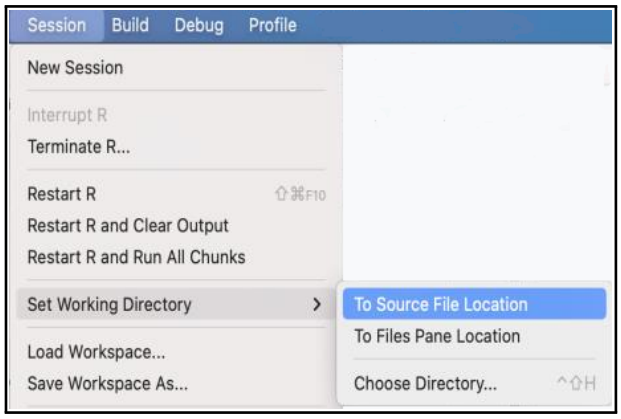

Figure 7: *Sourcing the file location.*

Activity, Part 1

1. Each code block has a play button that runs the code. Start with the ”read table” block and run the code by clicking on the ”play/run” button. What output is shown?"

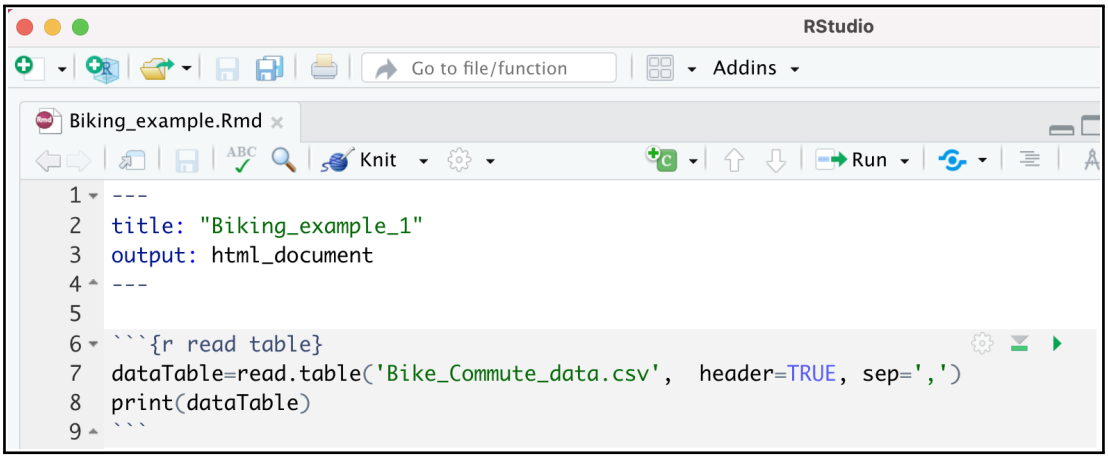

Figure 8: *Reading a data table using R.*

Answer: A table with 10 rows and 2 columns. The first column represents the states, and the second column is the number of bicycle commuters.

2. How many observations are there (run the code block called, ‘Check table information 1’)? How many variables are there (run the code block titled, ‘Check table information 2’)?

   Answer: There are 10 observations and 2 variables (Note: RStudio lists these as “State” and “Number_ of_Bicycle_Commuters”).

3. List variable 1 values (run the code block called, ‘Print values in variable 1’)

   Answer: "Alabama", "Alaska", "Connecticut", "Delaware”, "Florida", "Georgia", "Idaho”, "Kansas", "Maryland", "Virginia".

4. List variable 2 values (run the code block titled, ‘Print values in variable 2’)
   - What is the minimum and maximum value of variable 2?

     Answer: The minimum value is 1329, and the maximum value is 54652

- Run the code block 'Make bar plots.' Was it easier to find the states with minimum and maximum values now? Why?

  Answer: Yes, visualization helps in summarizing data and identifying patterns.

- Which state has the maximum number of bicycle commuters? Which state has the minimum?

  Answer: Florida has the maximum number of bicycle commuters, and Delaware has the minimum number of bicycle commuters.

5. Find the mean of variable 2 (number of bicycle commuters) using a calculator. Interpret your finding.

   Answer: $\frac{2505+3553+5320+1329+54652+9735+6746+4814+8955+15654}{10} = 11326.3$. The mean represents the average value across all measured observations.

6. Run the code block titled, 'Get the mean.' Does the answer match your finding in the previous step? Which way was faster?

   Answer: Yes. Using R is faster.

7. Find the median of variable 2 (number of bicycle commuters) using paper, pencil, and calculator. Interpret your finding.

   Answer: First order the values of variable 2: 1329 2505 3553 4814 5320 6746 8955 9735 15654 54652. Find the mean of the middle two values: $\frac{5320+6746}{2} = 6033$ commuters. The median represents the center of data that falls in the middle of the sorted observations, whether ascending or descending.

8. Run the code block called, 'Get the median.' Does the answer match what you found in the previous step? Which way was faster?

   Answer: Yes. Using R is faster.

9. Find the 5-number summary using the code block titled, 'Get 5-number summary'

   Answer:

min	1st Qu.	Median	Mean	3rd Qu.	Max.
1329	3553	6033	11326	3735	54652

10. Visualize the 5-number summary using a box plot by running the code block called, 'Plot boxplot'

    - What are the median, 1st quartile, 3rd quartile, minimum, maximum values?

      Answer: Median = 6033, 1st quartile = 3553, 3rd quartile = 9735

    - By examining the boxplot, are there any outliers? If yes, how did you reach that conclusion? Which state is the outlier, if any?

      Answer: Yes, one outlier exists, which is the state of Florida. In the boxplot, the point representing Florida is more than $1.5 \times$ Interquartile Range (IQR) above Q3 and thus is considered an outlier.

Activity, Part 2

The goal of the second part of the activity is to reanalyze the biking commute data excluding the outlier to examine the effects of the outlier on the summary statistics. Utilize the code provided in Part 1 throughout Part 2. You will have the opportunity to write R code to conduct the analysis similar to Part 1.

1. Download the following files to your computer to the same folder:
   - Bike_Commute_data_without_outlier.csv: This is the input data without the outlier. The file is available at https://www.dropbox.com/s/hhw401vee0j22vs/Bike_Commute_data_without_outlier.csv?dl=0
   - Activity_2_template.Rmd: This is the R Markdown template for Activity 2. The file is available at https://www.dropbox.com/s/keupsd2u92e6ilq/Activity_2_template.Rmd
2. Open the file using RStudio.
3. Read the data provided in the file titled, 'Bike_Commute_data_without_outlier.csv' and store it in a table. To do so, print the following code (Note: We recommend that you type the following code rather than copy and pasting):

```
```{r read table}
dataTable=read.table('Bike_Commute_data_without_outlier.csv',
   header=TRUE, sep=',')
print(dataTable)
```
```

4. Find the mean on Rstudio using the following code:

```
```{r get the mean}
print('The mean is:')
mean(dataTable$Number_of_Bicycle_Commuters)
```
```

5. Compare the mean value that you get in Activity, Part 2 to the value obtained in Activity, Part 1. Which mean value better reflects the data collected and why?

   Answer: The mean found in Part 2 excluding the outlier is roughly 6512 commuters which is smaller than the mean found in Part 1. The mean found in Part 2 better reflects the average of the observations measured. The mean is biased by extreme values.

6. Plot the boxplot using Rstudio:

```
```{r plot boxplot}
boxplot(dataTable$Number_of_Bicycle_Commuters, horizontal = TRUE,
   xlab='Number_of_Bicycle_Commuters')
text(x=boxplot.stats(dataTable$Number_of_Bicycle_Commuters)$stats,
labels=boxplot.stats(dataTable$Number_of_Bicycle_Commuters)$stats,
y = 1.3, srt=90)
```
```

7. Compare the boxplot to the boxplot obtained in activity Part 1. What do you notice? Are there any outliers? Justify your answers.

Answer: There are no outliers in the box plot. The median is less biased by the outliers than the mean. This shows that the outlier can heavily influence the mean.

To facilitate working with the activities, we provided activity handouts (for both Windows and Mac) which could be distributed to the students. The handouts are available for download as detailed below:

- Activity handout for students (Windows): https://www.dropbox.com/s/bmy7i1fisl88b66/Bike_Commute_Activity_Windows.docx?dl=0
- Activity solution for teachers (Windows): https://www.dropbox.com/s/jtz9xe2mqllaomc/Bike_Commute_Activity_solution_Windows.docx?dl=0
- Activity handout for students (Mac): https://www.dropbox.com/s/bwaj6ate30uubev/Bike_Commute_Activity_mac.docx?dl=014
- Activity solution for teachers (Mac): https://www.dropbox.com/s/3ersd665v1hjgf5/Bike_Commute_Activity_solution_mac.docx?dl=0

Summary and Extensions

Data science involves three major components: (1) probability and statistics, (2) quantitative reasoning, and (3) computer science. As our world becomes increasingly reliant on big data for day-to-day decision making, data science has become an important area of study for school students.

In this paper, we provided an example that helps ease students (and their teachers) into an exploration of data science using a model adapted from Bargagliotti et al. (2020) (See Figure 2). Specifically, we addressed three elements of the model—namely, (1) considering the data, (2) analyzing the data, and (3) interpreting the results. To address the quantitative reasoning component, we utilized level C of the GAISE II report by asking students to explain, interpret, and justify their findings. We address the last component from ODE's *Data Science* themes (i.e., computer science) by using the R programming language to represent and analyze the data (See Figure 1).

To enhance student engagement in the *Commuting to Work* activity, teachers may consider addressing the first step of the statistical problem-solving framework, which calls on students to formulate statistically investigative questions. Teachers may consider using the form of: "Here is a situation; think about it" (adapted from Pollak, 1966) to create an open-ended activity for the students that increases the cognitive demand of the task. As always, teachers should anticipate students' responses and monitor their engagement to ensure that they are on the right track as they complete the activity (Smith, Steele, & Sherin, 2020).

Conclusion

In this paper, we introduced how to use the R programming language and RStudio to analyze data science problems within a classroom setting. Data science is an emerging field, and students should be equipped with the right tools to analyze and interpret data originating from multiple real-world sources and scenarios. The RStudio framework provides one unified environment where students can read the data, present the data, perform analysis, and visualize the analyzed results.

All data and R code used in the activities are provided through Dropbox, in which teachers can easily download and share the activities with the students. In the presented activities, we showcased the ease of using the R programming language with RStudio to analyze data science problems.

References

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. (2020). *Pre-K-12 Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report II*. American Statistical Association and National Council of Teachers of Mathematics. https://www.amstat.org/asa/files/pdfs/GAISE/GAISEIIPreK-12_Full.pdf

Baumer, B., Cetinkaya-Rundel, M., Bray, A., Loi, L., & Horton, N. J. (2014). R Markdown: Integrating a reproducible analysis tool into introductory statistics. *Technology Innovations in Statistics Education, 8*(1).

Chen, A. (2020). High school data science review: Why data science education should be reformed. *Harvard Data Science Review, 2*(4).

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common core state standards for mathematics*. Washington, DC: Author. Retrieved from http://corestandards.org/assets/CCSSI_Math%20Standards.pdf.

Pollak, H. O. (1966). On individual exploration in mathematics education. In E. G. Begle (Ed.), *The role of axiomatic and problem solving in mathematics* (pp. 117–122). Washington, DC: Conference Board of the Mathematical Sciences (published by Ginn).

Smith, M. (Peg), Steele, M. D., & Sherin, M. G. (2020). *The 5 practices in practice [high school]: Successfully orchestrating mathematics discussions in your high school classroom*. Thousand Oaks, CA: Corwin.

Ohio Department of Education. (2021). *Data Science Foundations Course Pilot*. Retrieved from: http://education.ohio.gov/Topics/Learning-in-Ohio/Mathematics/Resources-for-Mathematics/Math-Pathways/Data-Science-Foundations.

Stein, M. K., Smith, M. S., Henningsen, M. A., & Silver, E. A. (2009). *Implementing standards-based mathematics instruction: A casebook for professional development* (2nd ed.). Reston, VA: National Council of Teachers of Mathematics, and New York, NY: Teacher College Press.

van der Aalst, W. (2016). *Process Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Ahmad M. Alhammouri is an Assistant Professor of Mathematics Education in the College of Education and Professional Studies at Jacksonville State University. His interests include the teaching and learning of mathematical modeling, quantitative reasoning, and data science. In addition, Dr. Alhammouri's interest includes designing and implementing professional development activities for mathematics teachers.

Rami Al-Ouran is an Assistant Professor in the School of Computing and Informatics at Al-Hussein Technical University. Previously he was a bioinformatics and data analyst in the Laboratory for Integrative Functional Genomics and the Bioinformatics core lab at Baylor College of Medicine, Houston, Texas. His research interests include Bioinformatics, Machine learning, Data mining, and Computational regulatory genomics.