USCOTS Workshop June 2023
Engaging Students in Data Science with Authentic Data
link to slides

6/2/23, S. Johnson, M. Kursav, S. Pauls, and C. Franklin.

Contents

Introduction

Welcome to *Engaging Students in Data Science with Authentic Data* - if you haven't already, please join the Google Jamboard using the QR code. On the first page, use the sticky to record whatever words come to mind when you think about data science.

Good afternoon! I'm Sheri Johnson from The Mount Vernon School and I'm here with my colleagues Merve Kursav and Scott Pauls, both from Dartmouth College, and Christine Franklin of the ASA.

This presentation will have four parts. Scott will tell us about the DIFUSE Project at Dartmouth College, then he and Merve will guide us through working with one of their many modules. I will share about our high school class implementation and Merve will help me summarize our exploratory analysis of pre and post surveys and student interviews. Lastly, Chris and Scott will wrap up a discussion of frameworks and some reflection. We aim to have time for questions at the end, but also welcome questions you may have throughout.

Part 1: DIFUSE at Dartmouth College

The NSF supported (IUSE 1917002) Data Science Infusion into Undergraduate STEM Education (DIFUSE) project at Dartmouth College aims to infuse data science into introductory courses in STEM and the social sciences. In this section, we'll briefly introduce you to the project. To get us started, we want to see what you all think about Data Science. As an emerging field, Data Science doesn't yet have a single definition so use this Jamboard to put down words that come to mind when you try to define it for yourself.

[Give 2-3 minutes for people to do the activity]

In the DIFUSE project, we adopted a definition that covers many of your ideas.

> Data science, drawing from mathematics, statistics, and computer science, is a multidisciplinary field using scientific methods to gain knowledge and insights from structured and unstructured data.

In the DIFUSE project, our primary goal is to expose students to data science who might not otherwise see it. We want to generate interest, have them see tools in action in another field, and see how Data Science may interact with other fields they are interested in.
We identified some Data Science competencies: acquiring, managing, analyzing, and visualizing data; drawing conclusions, communicating data. These are drawn from EDISON Data Science Framework ("EDISON Data Science Framework (EDSF) | Edison Project", 2017), and some Data Science guidelines endorsed by ASA ("Curriculum Guidelines for

Undergraduate Programs in Statistical Science", De Veaux et. al., 2017), which in part is based on the [ASA's 2014 Curriculum Guidelines and white papers] (American Statistical Association, n.d.). We wanted to create flexible tools and resources that can be reused and spread across different courses. In the interest of time, we will not dive into what motivated us in writing the grant or about some of our extra tools. But if you are interested, take a look at some supplemental materials at https://dartgo.org/difuse-info.

Data Science Modules

Our main product is a curricular intervention we call modules – self-contained, usually short course components that feature data science and that can be easily incorporated into existing courses. We'll briefly talk about 4 example modules from the 15 we have created so far. They are from a diverse set of subject areas and cover different aspects of data science and modeling.

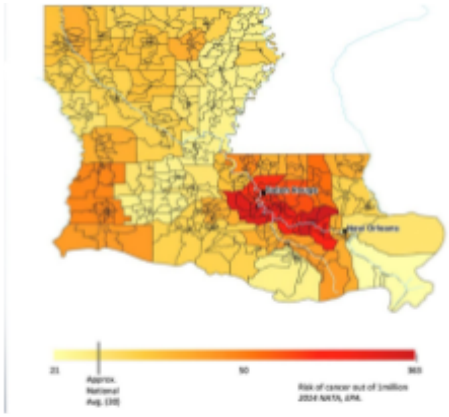


*Figure SEQ Figure * ARABIC 2: A map of Louisiana with a color-scale representing the risk of cancer per million* ("Tulane Study: Louisiana's Severe Air Pollution Linked to Dozens of Cancer Cases Each Year | Tulane Law School" n.d.).

The first module looks at air quality data in Germany and uses visualizations of air quality measures and wind speed and direction to investigate the impact of siting different types of building in different places in and around Berlin and other cities in Germany. Students put all of this together into a policy recommendation for proposed government projects.

The second module considers the impact of different statistics on health outcomes in Texas counties. The data is a mixture of demographic and other statistics including access to healthcare, pollution, physical inactivity, etc. that could impact health. Students use visualizations to generate hypotheses about relationships and then study correlations and the results of linear regressions to evaluate them.

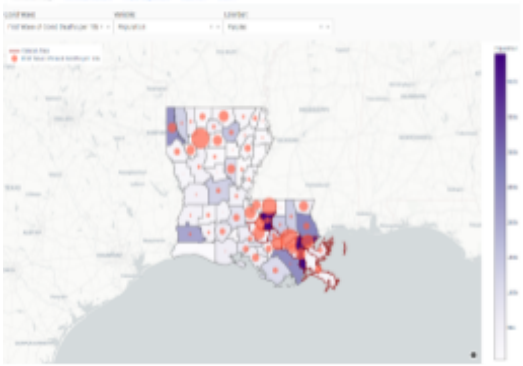


*Figure SEQ Figure * ARABIC 2: A screenshot from the Environmental Studies module showing a comparison between COVID-19 deaths and population.*

In the third module, students first learn about a model for the glucose-insulin interaction in humans and then use glucose and insulin level data to fit parameters of the model. They then use the fitted model to make predictions about the outcomes of different therapeutic interventions.

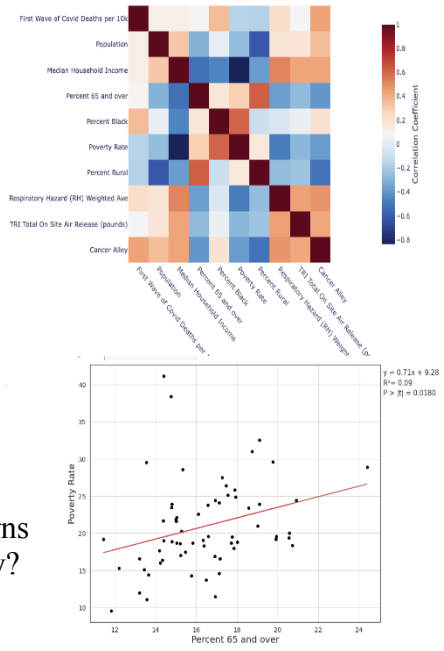
Part 2: Working with a DIFUSE module

The fourth module is one we will explore together and is the module used by students at The Mount Vernon School. To give you a better sense of what the modules are and how they work, we ask you to try one of them out. This module was developed for an Environmental Science course at Dartmouth, called Environment and Society, that looks at the interplay and interdependencies between environmental and societal factors. The module has students (and you!) look at the potential relationships between demographic and environmental measures and COVID deaths over time early in the pandemic.

The DIFUSE module team selected this geographic region because of an area called cancer alley - a string of parishes (a parish to the Louisiana analogue of a county) running from New Orleans towards the northwest along the Mississippi river. This shows an EPA visualization of cancer rates across Louisiana, which cancer alley shown in bright red - this corridor is highly polluted, and some communities have much higher cancer rates than average. Much of this pollution is airborne, creating higher respiratory health risk among inhabitants. Given the respiratory symptoms associated with COVID, the team saw this as a rich area of investigation and a nexus of potential causal relationships.

For a few minutes, we'll have you work with one aspect of the module to give you some familiarity and work with the visualization tool. The image on the right is a screenshot from the web application we created for this module. There are several tabs on the top - interactive map, correlation matrix, linear regression, data info, and credits. We will only be using the interactive map, but we'll talk about the others a little bit later. This shows a version of the interactive map, showing the parishes in Louisiana with two different annotations - color representing one of the statistics picked from the drop-down menu labeled "variable" and a circle whose size corresponds to the number of COVID deaths in the wave selected from the "COVID Wave" dropdown.

As we mentioned earlier, cancer alley has a high degree of pollution that contributes to poorer air quality. We might conjecture that higher health risk due to air quality would also place people at higher risk for severe COVID due to higher baseline stress on their respiratory systems. We can get a sense of whether this is true by visualizing the COVID deaths and respiratory hazard measures on the same map. The GIF to the right shows an animation of executing this process.



Now it is your turn! Follow the link at https://envs3-app-yaex.onrender.com/ and think about two questions:

1. When you look at the Interactive Map, what patterns emerge between race, class, and COVID-19 mortality?

2. What are some hypotheses about what drove COVID-19 mortality in Louisiana?

Spend a few minutes (~5) looking for relationships and generating hypotheses.

You've just (quickly!) completed the initial part of the assignment that students would use with this app. Let's take a moment to discuss your reactions:

1. What did you think of the interface and its use? How will students respond?
2. What types of conclusions can you draw from using the interactive map?

After this step, students would then proceed to use different methods to evaluate those hypotheses, revise them, and iterate. The app includes two additional tools to help with this. The Correlation Matrix tab (Figure 3) provides a visualization of the correlation matrix across all the variables allowing students to see associations across many variables at once. The Linear Regression tab (Figure 4) provides bivariate scatter plots determined by the variables in the drop-down menus and can also display a linear fit coupled with its defining equation and R- squared value. The latter opens the door to learn about lots of statistical methods and ideas in mathematical modeling - the extent of these discussions in the class depends on student experience and available time.

Part 3: DIFUSE in High School

With successful modules developed at the college level, we were curious about their efficacy at the high school level. So, we used the DIFUSE module you just worked with in a high school setting. We selected this because it has accessible mathematics for high school students.

The Mount Vernon School is a private, independent school in the metro-Atlanta area. We are a school of inquiry, innovation, and impact where we use a competency-based approach where students can reassess any summative assignments as long as their formative work has been completed. We piloted this module in both an AP Statistics (19 students) and a non-AP Statistical Reasoning class (11 students).

We wanted to learn about students' knowledge, attitudes, and beliefs so we surveyed and interviewed students. The pre-survey was given prior to any instruction on two quantitative variables. It is important to note that this survey was given after instruction on one quantitative variable, because that may have influenced some of the students' responses. The post survey was given after the module was completed and we used these survey results as well as student work to select three students to interview.

Linkage to GAISE II

GAISE II ("Guidelines for Assessment and Instruction in Statistics Education (GAISE) Reports" 2020) gives us a framework for statistics and data science at the high school level. One of the changes from GAISE I to GAISE II was that "Collect the Data" changed to "Collect/Consider the Data", so the statistical problem-solving process can start there. This DIFUSE module gave the

students an opportunity to start with the data and use some visualization tools to complete some analyses, continually question, and formulate a statistical investigative question.

GAISE identifies Levels A, B, and C which are generally aligned with elementary, middle, and high school - maybe level D is college, but most importantly they are sequential as statistical concepts take time to develop. So, for some concepts even a college student might need to start at level A.

An alternative Google CoLab version of the web-app tool you just used includes an option to report all the R-squared values simultaneously. For a class that starts at level A the color-only version might be a nice option, however a more advanced student, maybe at level C or D, might appreciate the number-filled matrix.

Classroom Implementation

We made some modifications to how the ENVS3 module was used at the college level to adjust for high school students (lesson slides). Some of these pedagogical choices included giving a simpler article to understand the inequities that exist in Cancer Alley. And since rates are more difficult for us humans to work with (as Regina nicely pointed out in this morning's keynote) we explicitly discussed the difference between cases, case rates, deaths and death rates. We also had the students use a tool from the CDC to discover and communicate interesting observations.

Since many students struggle to master some important mathematics concepts like proportional reasoning and geometry, we took an opportunity to address these concepts through a data activity. We used a NYT's "What's going on in this graph" to introduce a third quantitative variable and discussed scaling the bubbles - should that be done by area or diameter?

Students were tasked to determine if the bubbles in the map you just worked with were appropriately scaled. This proved to be a VERY challenging exercise for high school students. Data Scientists range in their technical expertise and most high school students are not adept at computer programming, so we decided to let the students select an option to either review and modify the python code or create a change request form to scale the bubbles appropriately.

To complete the statistical problem-solving process and communicate their results, the final product for each group was to complete a poster and submit it to ASA's high school data visualization contest. Prior to working with the module, we discussed what a good poster would look like. To help scaffold their work and think more deeply about communicating the data in a way that answers the statistical investigating question they created, the students reviewed winning ASA posters.

Survey Results

So, what did these high school students think about this? First, we will look at the survey results and then the interviews. Overall students were rather favorable about their attitudes and beliefs. A

majority of students consistently value data science - they are interested in working with data, don't get too frustrated, and see the importance of data science skills for their future.

Some initial statistics:

- 33% of responses are from post-surveys, and 67% of them are from pre-surveys.
- Whereas 69% of students are from AP class, 28% of students are from SR class (3% of them did not identify the class).

An interesting change from pre- to post was that overall students' desire to communicate with the data increased but their desire to analyze the data decreased. Our survey also included some open-ended questions and from a qualitative perspective, students were glad to have multivariate datasets and work through the entire statistical problem-solving process. After the project, their attitudes and confidence in their statistical knowledge improved. This is consistent with results from evaluations of the modules in higher education settings, where ratings of self-efficacy rose after deployment of the modules. The pre-survey responses had a substantial focus on one quantitative variable rather than 2, which is likely due to a recency effect.

Interview Results

To supplement and contextualize the quantitative data, three students participated in semi-structured interviews.

- Evangeline - a junior who is a disciplined elite athlete, both conscientious and advanced academically, rarely having to reassess. Her group was collaborative and capable.

- Amelia - A junior who is conscientious, but generally needs some time to understand instructions, struggles to remember past concepts, and often reassesses assessments. Her group was a bit disjointed, and in the poster presentations, Amelia corrected some incorrect information from other group members.

- Jay - A senior who is a mathlete and an athlete. He has an interest and aptitude for mathematics, but with only a few days left in school, he mentioned that he didn't need a good grade on this project to keep a good grade in the class and initially was not engaged. After his group members struggled to complete a Novice draft poster, Jay worked to improve their efforts.

In the table below, we share some quotes taken from when each interviewee was asked about their knowledge, skillset, and Beliefs and Attitudes towards data science before and after completing the module.

	Evangeline	Amelia	Jay
Knowledge	"I learned a lot but still need to have a better grasp of R-squared values and interpreting them. I think that using them in context helped me to understand the regression and what a high correlation looks like."	"I learned what to look at the data… it was interesting to be able to use correlations for multiple variables"	"I thought that it helped me understand a lot more about, p-values… R-squared"
Skillset	"I had never worked with a map or a matrix, so my skillset on how to interpret those definitely improved."	"I can think about data more critically"	"I learned some coding. I would like to look at code to be able to create similar graphs and regression models, stuff like that"
Beliefs/Attitudes	"I felt very engaged. I really enjoyed doing the work. I was very bought into it. I liked learning about it…I feel like an intern"	" I like working with data science. I think this project was helpful obviously."	"overall, I feel like it was like a very interesting project…. seeing all of the code is kinda like what made me like a little nervous…I liked the project. I thought that it gave like more of a real-world application."

Evangeline said that she had not worked with heat maps before this project and this was an interesting experience for her. Evangeline felt engaged, enjoyed working with the data, and had fun exploring different variables and assumptions. The most important thing she said was that "I felt like an intern."

Amelia thought it was interesting to be able to use correlations for multiple variables. She had some initial wonders about the dates for the different Covid waves and some difficulty realizing the bubbles didn't rescale when the other variables were changed. She learned how to look at the data more critically. She enjoyed the analysis more than communicating the results and discussed gains in knowledge more than attitudes.

For Jay, although the initial display of code on Google Colab was scary, he expressed a desire to improve their coding skills to better manipulate and analyze data. He loved communicating the results. He wants to have a career in finance and data science. He liked the module and thought this helped him to understand what he needs to improve to be successful in his career in the future. He is more interested in learning coding. In the module, he thought the map was the most helpful tool for him to understand a lot more of the concepts they were talking about in class. He thought that he got a better understanding of statistics and how data science is used in the real world.

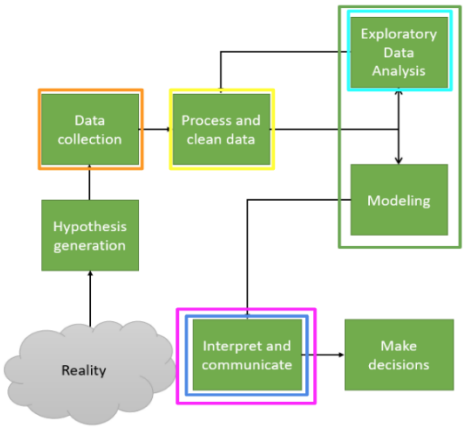
Overall, their interest in and knowledge of data science increased after the module deployment. In

our interviews, we asked students, if they were given a spectrum from 0 to 10, how they would rate their interest, resources, and experiences in data science before and after their module experience in their class. We clarified that resources include both instructional resources and personal-psychological resources (e.g., confidence, self-efficacy). Every student noted an increase for each of these aspects.

Part 4: Reflections and Connections

In writing the DIFUSE grant, we drew from the content components of the EDISON Data Science Framework ("EDISON Data Science Framework (EDSF) | Edison Project", 2017), and some Data Science guidelines endorsed by ASA ("Curriculum Guidelines for Undergraduate Programs in Statistical Science", De Veaux et. al., 2017), which in part is based on the ASA's 2014 Curriculum Guidelines and white papers (American Statistical Association, n.d.).

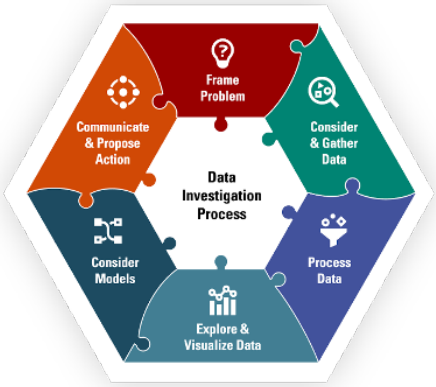
We also thought about a standard workflow for data science projects, as illustrated in the figure to the right. We start with observing a system or process and making hypotheses. Then, we collect data that can help us evaluate our hypotheses and then clean and/or process it for analysis. Next, we enter an iterative process of exploratory data analysis and modeling which may prompt us to reprocess the data, or even collect new data. As this iteration ends, we finish the analysis of our data and associated models and turn to interpreting and communicating the results in terms of our initial hypotheses. Last, we make decisions about the questions we initially posed.



From our use of the two frameworks, we identified six content areas with which we could classify the modules we create - acquiring data (orange), managing data (yellow), analyzing data (green), visualizing data (cyan), drawing conclusions (blue), and communicating data (purple). These map into the workflow - we show this in the diagram by outlining the workflow component with the color associated to the content area. For example, acquiring data maps to data collection and visualizing data maps to EDA.

Connections to other frameworks: Data Investigation Process

We see connections with recent work of Lee, Mojica, Thrasher, and Baumgartner (Lee et al., 2022) whose "Data Investigation Process" (see Figure 1 of that paper) matches our framework closely. The data investigation process has 6 interlocking pieces: Frame the Problem, Consider and Gather Data, Process Data, Explore and Visualize Data, Consider Models, Communicate and Propose Action. While one can move through these sequentially, there is also lots of room for iteration, backtracking, and revision. Our process maps neatly onto these categories and, consequently, our competencies do as well.



Connections to other frameworks: GAISE II

We can also map aspects of the DIFUSE framework to GAISE II. The top-level GAISE II Statistical Problem-Solving Process (SP) is a general organizing principle that serves as an umbrella framework for many more narrowly oriented data science frameworks and DIFUSE is no exception. The SPSP echoes the scientific method in its four elements – Formulate a Statistical Questions, Consider/Collect Data, Analyze the Data, and Interpret the Results – with many interconnections between them, as shown in Figure 6. Mapping the DIFUSE data science workflow into the SPSP simply granularizes some of these broader categories. Hypothesis Generation fits into "Formulate Statistical Investigative Questions", while both Collect and Process Data map into "Collect/Consider Data." Visualization and Modeling fall under "Analyze the data," and Communicate Results and Make Decisions link to "Interpret the Results."

Figure 6: A schematic of the Statistical Problem-Solving Process.

In its revision, GAISE II calls for seven actions for the field to implement. In the table below, we summarize the extent to which DIFUSE modules can answer these calls for action.

GAISE II calls for …	**DIFUSE fills the gap by…**
1. Asking questions throughout the SPSP	Scaffolding of the assignments within each module prompt students to question and reflect at multiple points.
2. Considering different data as a starting point and using data throughout the SPSP	Modules present students with data and tools to easily manipulate, visualize, and explore the data sets and connections within them.
3. Inclusion of multivariate thinking	Modules are designed to allow students to consider multiple possible avenues of explanation of phenomena within a data set. Assessments prompt students to identify impacts of multiple variables in the questions they address.
4. Using probabilistic thinking versus deterministic thinking	Modules are built on probabilistic models which engage students in thinking about problems probabilistically.
5. Incorporating technology	Almost all the modules incorporate new technological components into a course, including python notebooks, custom applications, and MATLAB analyses.
6. An enhanced importance of clearly and accurately communicating statistical information	Almost all modules emphasize deliverables that communicate conclusions and recommendations to a broader audience.
7. Assessment that requires conceptual understanding and the SPSP	All modules contain assessment components that probe both conceptual understanding and aspects of the SPSP.

To elaborate on this table, we will discuss each item in turn.

1. *Asking questions throughout the SPSP.*

In all of the modules, we paid a lot of attention to scaffolding the assignments within the module to prompt students to question and reflect at multiple points. This ties right into the statistical problem-solving process in getting students to generate ideas and hypotheses and questions, and then follow them through the data analysis process.

2. *Considering different data as a starting point and using data throughout the SPSP.*

In the modules, a lot of the initial work in the SPSP is actually behind the scenes. They're done by the project teams in collaboration with the instructor for the course. Usually the data is curated beforehand, so we don't really address considering different data types.

This is a choice made in the module design to emphasize different aspects of the process that are more relevant to the course or the instructor's goals, as well as to the student preparation. Remember

that the courses are introductory courses in fields that are not statistics or mathematics. Students don't have any particular background knowledge and training that we can rely upon.

So that leads us to pick and choose between the parts of the SPSP to, again, emphasize the things that are most important in that context. One thing I will emphasize is that the modules are created to help students very quickly engage with the data. So, we build in tools for visualization and analysis, as well as that data curation step, so that students can jump into the analysis and interpretation part.

3. *Inclusion of multivariate thinking*
Some of the modules are organized around univariate statistics, some are bivariate statistics, but often what we have is what you saw in the Environmental Studies module, where there are lots of variables and students have to pick and choose between them and think about how they might tell various stories or various sort of causal explanations for what's going on, which leads them towards multivariate thinking. There are a few other modules that go even further and allow you to build, for example, multivariate models and then test them with the data that you have against different hypotheses.

4. *Using probabilistic thinking versus deterministic thinking*
Even for the modules that use differential equations that are deterministic in nature, all of these have probabilistic components and almost all the models are probabilistic in nature. One caveat here is that most of the students that engage with these are not going to have much if any statistical training. And approaching the probabilistic model is a different activity for them than you might see in a statistics course.
This gets back to something that we talked about a lot in the last couple of weeks, which is sort of on the ladder from A, B, C to D in the GAISE framework. For these students we might have to drop back because they simply don't have the level of training that a more sophisticated statistics student at the undergraduate level might have.

5. *Incorporating technology*
Almost all the modules use technology to help bring the work to the students easily. Most of them use Python, some use R, and a couple use MATLAB. The one that uses almost no technology is an anthropology module where they're looking at hominid footprints, so we built a big sandbox where they walk and then measure their own footprints and compare them to ones from fossil records. We use Excel to capture and then look at that data.

6. *An enhanced importance of clearly and accurately communicating statistical information*
One of the common pieces of most modules, again, is this last step of solving some problem and then communicating it to a general audience. For example, in the module on air quality, students are thinking in terms of a policy question about different types of governmental projects, tying in real thinking, decision making, and communication with the underlying computational effort.

7. *Assessment that requires conceptual understanding and the SPSP*

One of the things we felt was important was a learning design component – we paid a lot of attention to the assessments and making sure that the assessments were probing deeper things, rather than more surface questions about application output or top-level answers to questions. We wanted assessment that found out whether they were engaging with the conceptual frameworks, thinking deeply about, for example, policy problems like I mentioned before, and how the process that they used, whatever part of the SPSP that they used, came into their thinking.

References

American Statistical Association, Curriculum Guidelines for Undergraduate Programs in Statistical Science. n.d., Retrieved from: https://www.amstat.org/education/curriculum-guidelines-for-undergraduate-programs-in-statistical-science-.

De Veaux, Richard D., Mahesh Agarwal, Maia Averett, Benjamin S. Baumer, Andrew Bray, Thomas C. Bressoud, Lance Bryant et al. "Curriculum guidelines for undergraduate programs in data science." *Annual Review of Statistics and Its Application* 4 (2017): 15-30. https://doi.org/10.1146/annurev-statistics-060116-053930

EDISON Data Science Framework. 2017. Retrieved from https://edison-project.eu/edison/edison-data-science-framework-edsf

Guidelines for Assessment and Instruction in Statistics Education (GAISE) Reports." n.d.. Retrieved from: https://www.amstat.org/education/guidelines-for-assessment-and-instruction-in-statistics-education-(gaise)-reports.

Lee, Hollylynne S., Gemma F. Mojica, Emily P. Thrasher, and Peter Baumgartner. 2022. "Investigating Data Like a Data Scientist: Key Practices and Processes." *Statistics Education Research Journal* 21 (2): 3–3. https://doi.org/10.52041/SERJ.V21I2.41.

Tulane Study: Louisiana's Severe Air Pollution Linked to Dozens of Cancer Cases Each Year | Tulane Law School." 2022. Accessed June 2, 2023. https://law.tulane.edu/news/tulane-study-louisianas-severe-air-pollution-linked-dozens-cancer-cases-each-year.