**Mary Ryan  
email:** [**mary.ryan@yale.edu**](mailto:mary.ryan@yale.edu)

**“STEM Meets Storytelling: Using Data Science to Tell Stories about Our Communities”  
*Pathways to Science Summer Scholars Workshop Syllabus 2022***

***Course Description:*** Data is all around us and can be important in understanding more about our communities. Often, though, we first need to summarize that data to understand the big picture. This course provides the opportunity to learn about how data is collected, summarized and visualized, and how we can use computer programming to perform these tasks and learn about our communities. Participants will work in small groups to help each other clean, transform, and visualize data that they will later use to write a blog post describing what is happening in the communities the data is about. Throughout the workshop, participants will be performing data analysis and coding just like real-world data scientists. Real-world applications of data science, jobs that utilize similar skills, and opportunities for further education in STEM will also be discussed.

**Course Objectives:**

1. Learn basic coding and programming skills
2. Learn the concept of “tidy data” and how to clean data.
3. Learn basic statistical summarization techniques.
4. Learn different types of data visualization, what they communicate, and identify with what types of data they are appropriate to use.
5. Understand how data can provide a different perspective on storytelling.
6. Use data and statistical programming to uncover and tell a community story through summarization measures and data visualization.

**Day 1: Monday  
Overview –** Introduction to data and the programming language R, including an introductory activity where students will get familiar with coding.

**Breakdown –**

|  |  |
| --- | --- |
| **Time** | **Activity Description** |
| *10 min* | Introductory discussion to data. Discussion points: What is data, how might it be collected? What is a variable and what is an observation? What are different types of data? How do we get from individual data points to overall observations about a group, area, or event? |
| *5-10 min* | **Activity 1:** Students will be split into small groups, which they will keep throughout they week. Students will practice gathering and identifying different types of data by asking a set icebreaker/get-to-know-you questions of their groupmates. |
| *10-15 min* | Continue introductory discussion of data, and the role of storytelling. Discussion points: Why might it be useful to make data available to the public? What are some real-world instances where data has been used to tell community stories? When did data get the story wrong, or lacked the context to tell the whole story? Followed by an introduction to statistical programming in R. |
| *30 min* | **Demonstration/code-along:** Instructor-led demonstration of the RStudio environment through RStudio Cloud, where students will code along with instructor. Students will learn about the console and command line, file structures and directories, and writing, opening, and closing scripts. Students will also learn important reproducibility concepts like using variables, writing comments in scripts to describe what is being done, and how to load and view data. |
| *30 min* | **Activity 2:** *Hello, world!* In small groups, students will practice coding in R by calculating how much it will cost to fill a car with gas. Using variables will demonstrate how this is easier to calculate as the price of gas and the size of the car tank changes.. They will also practice creating data tables using the data they gathered with classmates earlier. |

**EHS Notes:**

**Day 2: Tuesday  
Overview –** Introduction to tidy data. Students will perform their first data manipulations to transform data from a “raw” form to one that is easier to perform analyses on.

**Breakdown –**

|  |  |
| --- | --- |
| **Time** | **Activity Description** |
| *15 min* | Introductory discussion to tidy data. Discussion points: Review what is data, variables, and observations. Why does data need structure? What are the principles of tidy data? Why do we need a system for naming variables. Followed by a short activity. |
| *10 min* | **Activity 1:** Students will be given an example of a small dataset that is “messy”. They will be asked to identify what information is intended to be variables, and what are intended to be observations. How does the data violate tidy data principles? How would it need to change in order to be tidy? |
| *30 min* | **Demonstration/code-along:** Instructor-led demonstration in RStudio Cloud of how to transform messy data into tidy data. Students will learn important functions like clean\_names(), pivot\_longer(), pivot\_wider(), mutate(), arrange(), select(), distinct() and the concept of “piping” or chaining functions. We will follow a “cake recipe” analogy for the order data transformations need to be performed: some types of steps need to be performed before others (i.e, need to mix ingredients before you bake – should probably clean up variable names before you start creating new variables) but similar types of steps can be performed in any order together (i.e., doesn’t matter if you add eggs before water, or water before eggs – you can create new variable 1 before new variable 2, or new variable 2 before new variable 1). |
| *35 min* | **Activity 2:** *Time to clean up!* In small groups, students will be given a larger messy dataset than before and asked to make it “tidy” using the functions we learned in the demonstration. Hints and “cheat sheets” about processes will be provided. We will come back together as a class to compare what “messy” components each group identified, what order we cleaned them up in, and (potentially) different ways we cleaned them up. Were there multiple ways to perform the same task? Were some easier? Did some involve fewer lines of code |

**EHS Notes:**

**Day 3: Wednesday  
Overview –** Students will be introduced to the concept of data summarization through a short lecture. They will also learn how to calculate and interpret data summarization measures through an instructor-led demonstration. They will then summarize some data themselves.

**Breakdown –**

|  |  |
| --- | --- |
| **Time** | **Activity Description** |
| *20 min* | Introductory lecture on data summarization. Discussion points: what do we want to data to tell us about a population/group/event? How might we describe the most “typical” data point in the data? How do we measure the “middle” of the dataset? How do we determine whether our data points are very similar to one another or very different? How do our methods of summarization differ for numerical and categorical data? |
| *10-15 min* | **Activity 1:** Students will “measure” their height on the whiteboard by using a marker to mark where the top of their head touches the board. To the very righthand side of the board we will draw a vertical number line, and transfer the height measurements to that line; we will also measure those heights with a tape measure to place values next to the markings. Can students guess where the mean height will be on the line? The median? Can they guess what the standard deviation will be (less than 1 inch? More than 2 inches? More than 5 inches?)? We will calculate to confirm. |
| *25-30 min* | **Demonstration/code-along:** Instructor-led demonstration in RStudio Cloud of how to calculate summarization measures like counts, percentages/proportions, mean, median, minimum, maximum, standard deviation. Students will learn functions such as summarize(), n(), min(), max(), first(), last(), mean(), median(), sd(), group\_by() |
| *30 min* | **Activity 2:** In small groups, students will be given a (clean) dataset and asked to answer specific questions about summary measures in the data and how summary measures of subgroups compare to one another. Each group will get slightly different sets of summary measures to calculate. What conclusions do they come to about the data/population as a whole from these summary measures? Does each group come to the same conclusions, or do they come to different conclusions based on their different summary measures? |

**EHS Notes:**

**Day 4: Thursday  
Overview –** Students will be introduced to different types of data visualizations through a short lecture. They will learn how certain types of visualization are appropriate for certain types of data and not for others

**Breakdown –**

|  |  |
| --- | --- |
| **Time** | **Activity Description** |
| *30 min* | Introductory lecture on types of data visualization with examples from real-world journalism. Discussion points: When might it be easier to describe data visually instead of with just the values of data summary measures? When might displaying data visually bring your point across more strongly than just using the data summary measures? What are different types of data visualization? Why are some data visualizations appropriate for certain types of data and not for others? How can data visualizations be modified to mislead the reader about the data? |
| *30 min* | **Demonstration/code-along:** Instructor-led demonstration in RStudio Cloud of how to create and customize data visualizations. Students will learn functions like ggplot(), geom\_bar(), geom\_scatter(), geom\_line(), geom\_histogram(), geom\_map() |
| *30 min* | **Activity:** In small groups, students will be given a (clean) dataset and each be asked to create a different type of data visualization. They will also be asked questions to guide them in interpreting the visualizations. Groups will display their graph, without discussing it, while the rest of the students attempt to interpret it. Groups will then have the opportunity to modify their graphs to fix points of “misinterpretation” identified during the group discussion (i.e., adding graph/axis titles for context, adding legend for colors/symbols, etc.). If time allows, groups will present their edited graphs again to see if the “story” is clearer to the rest of the class and reveal what their prompt was. |

**EHS Notes:**

**Day 5: Friday  
Overview –** Students will take all that they’ve learned in the previous four days to perform analyses on some data we’ve been working with as well as new data, and write blog posts or record TikTok-style videos telling the story of what they’ve found.

**Breakdown –**

|  |  |
| --- | --- |
| **Time** | **Activity Description** |
| *10 min* | Very brief introduction to the goals of the day, introduction to what new data they have at their disposal. |
| *60 min* | **Activity:** *Tell me a story.* Students will work in small groups to clean, analyze, and visualize new and previously introduced data, and write a short blog post or record a short TikTok-style video explaining what they found and what observations they might be able to make about the community. All blogs/videos must contain at least one data summarization (i.e., mean, median, etc.) and one visualization. |
| *Remainder of time* | **Course wrap-up:** Whole-class discussion of what segments of the data the small groups focused on and what stories they decided to tell. Did some groups focus on similar topics but come to different conclusions? Was there data they wished they had access to that would’ve enriched what they explored? Would they have been able to tell that “story” through means other than data analysis (i.e., looking at documents, talking to community members)? Did the experience change their perspective on our community? |

**EHS Notes:**