|  |
| --- |
| **STEM Meets Storytelling:**  ***Using Data Science to Tell Stories about Our Communities***  Yale Pathways Summer Scholar Program 2022  Yale University |

Instructor: Mary Ryan

Contact: mary.ryan@yale.edu

Classroom: TBD

TA: TBD

***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.

***Activity Outline***

**Day 1: Introduction to data and computer programming with R**

There are many jobs out that that interact with data. For example, scientists collect data during experiments, TV stations collect data about what kinds of people are watching their shows, and . Sometimes data is collected about our communities and is accessible for everyone. But how do we go from looking at daily air quality observations in New Haven to understanding if the air is healthy overall? And how can you do this if you’re sometimes handling thousands of individual data points?

One method is through computer programming and data science. In our first activity, we will learn about the programming language R, and how we can communicate with the computer through code.

**Day 2: Creating tidy data**

Very rarely does data come to us in the perfect form. Have you ever tried writing the perfect letter with only one try? Sometimes there are mistakes we need to correct or we realize a sentence needs to be reworded or moved. The same thing happens with data! But why do we need tidy data and how can we clean data up?

We will learn about the principles of tidy data and why messy data can lead to mistakes. Then we will put our coding skills to the test and learn how to clean up existing data. Will *you* be able to find all the data dirt?

**Day 3: Summarizing data**

When you’re telling someone a story about something that happened to you, you don’t tell them every single detail about what color your shirt was and what the room smelled like, right? So much information at once is overwhelming and your friend might forget where the story was going! The reasons you *summarize* what happened to you are the same reasons we summarize data: they help us to keep track of the overall story the data is telling us.

We will learn about some basic data summary measures, what parts of the data they help us describe, and how to calculate them. Then we will summarize some actual data and try to figure out what its big picture is.

**Day 4: Getting the big picture with data visualization**

Data summarization is important but sometimes it’s easier to see patterns when we turn data points into a picture. What does it mean to visualize data? What types of data visualizations are there? How can we make sure we’re being honest about how we’re presenting the data?

We’ll learn all of this and make our very own data visualizations in R.

**Day 5: Tell us a story**

You’ve learned all the tools so now it’s time to put them to use. We will investigate some Connecticut public health data using our data cleaning, summarization, and visualization skills, and learn how to communicate what we find into a short blog post. This is exactly what data scientists, statisticians, and data journalists do in real life!