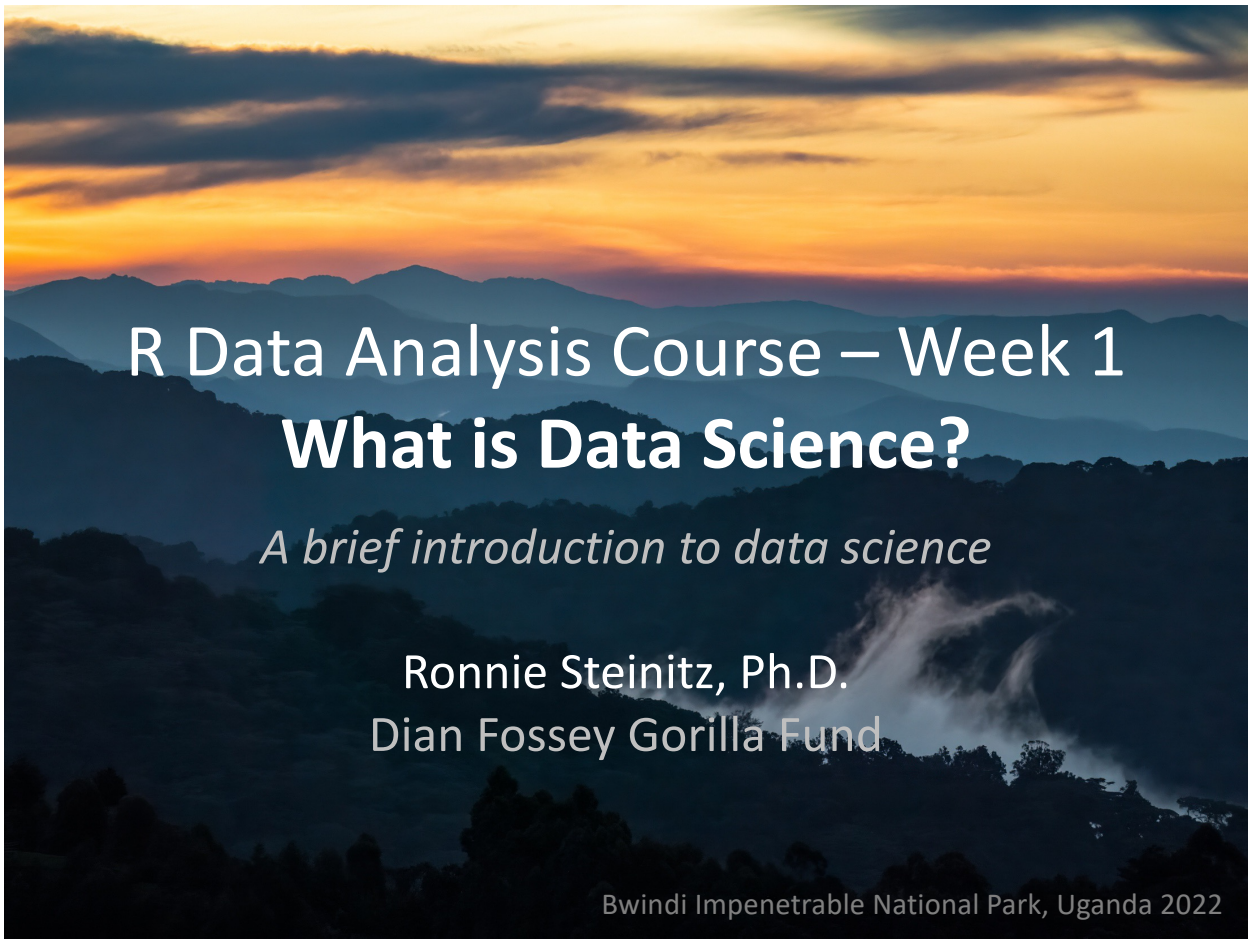
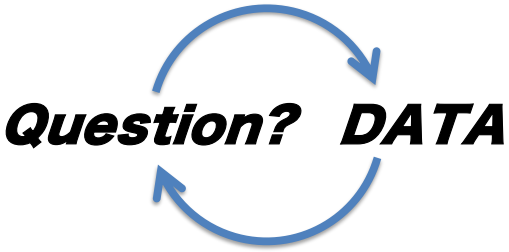


The Data Science Feedback Loop

- Data science begins with questions
- …questions that can be answered with data

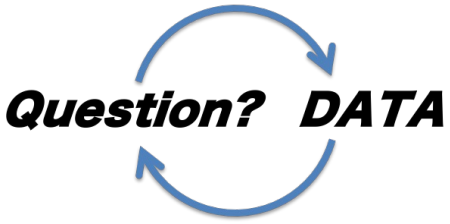


Collect, Curate, Clean

- Process is circular:
  - question → data → refine → repeat

What Do Data Scientists Actually Do?

1. Collect, curate, and clean data
2. Visualize the data
3. Analyze and model
4. Communicate findings and refine analysis

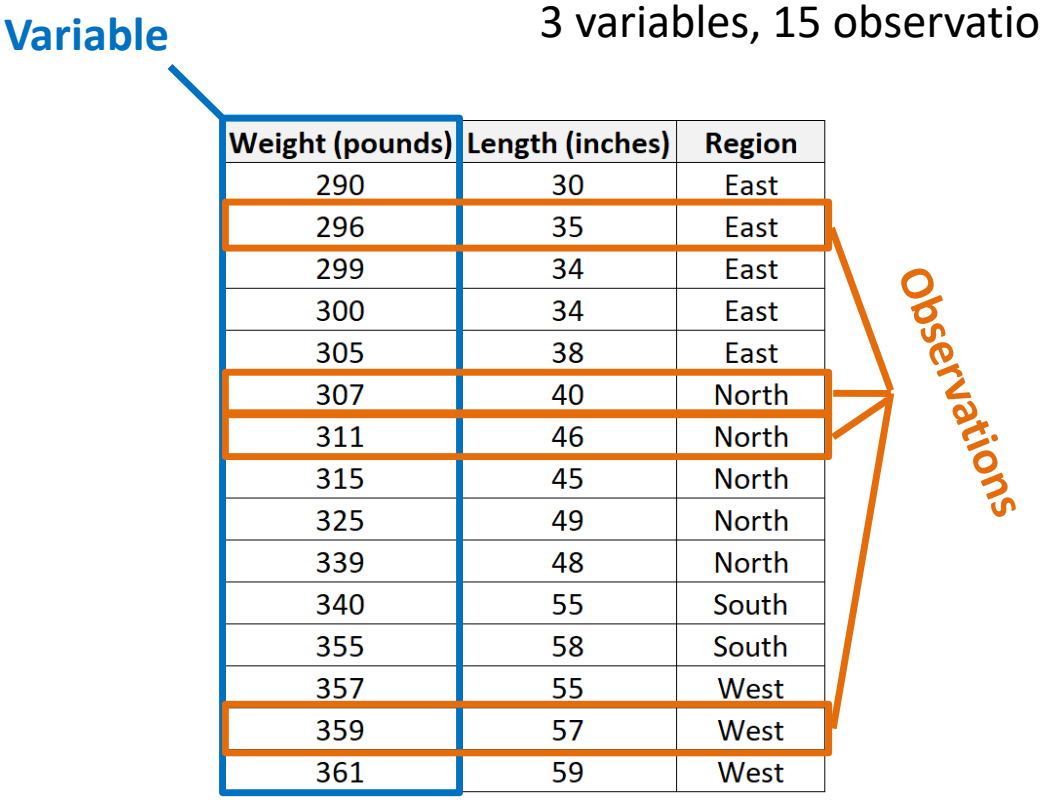


Step 1: Collect, Curate, Clean

Real-world data is messy

- Collection:
  - Get the data (field, sensors, records)
- Curation:
  - Choose what matters, organize it
- Cleaning:
  - Handle missing values, outliers, errors

Namabiro Aidah and Asaba Godfrey collecting monkey behavioral data on tablet
Kibale National Park, Uganda 2019



Weight (pounds)	Length (inches)	Region
290	30	East
296	35	East
299	34	East
300	34	East
305	38	East
307	40	North
311	46	North
315	45	North
325	49	North
339	48	North
340	55	South
355	58	South
357	55	West
359	57	West
361	59	West

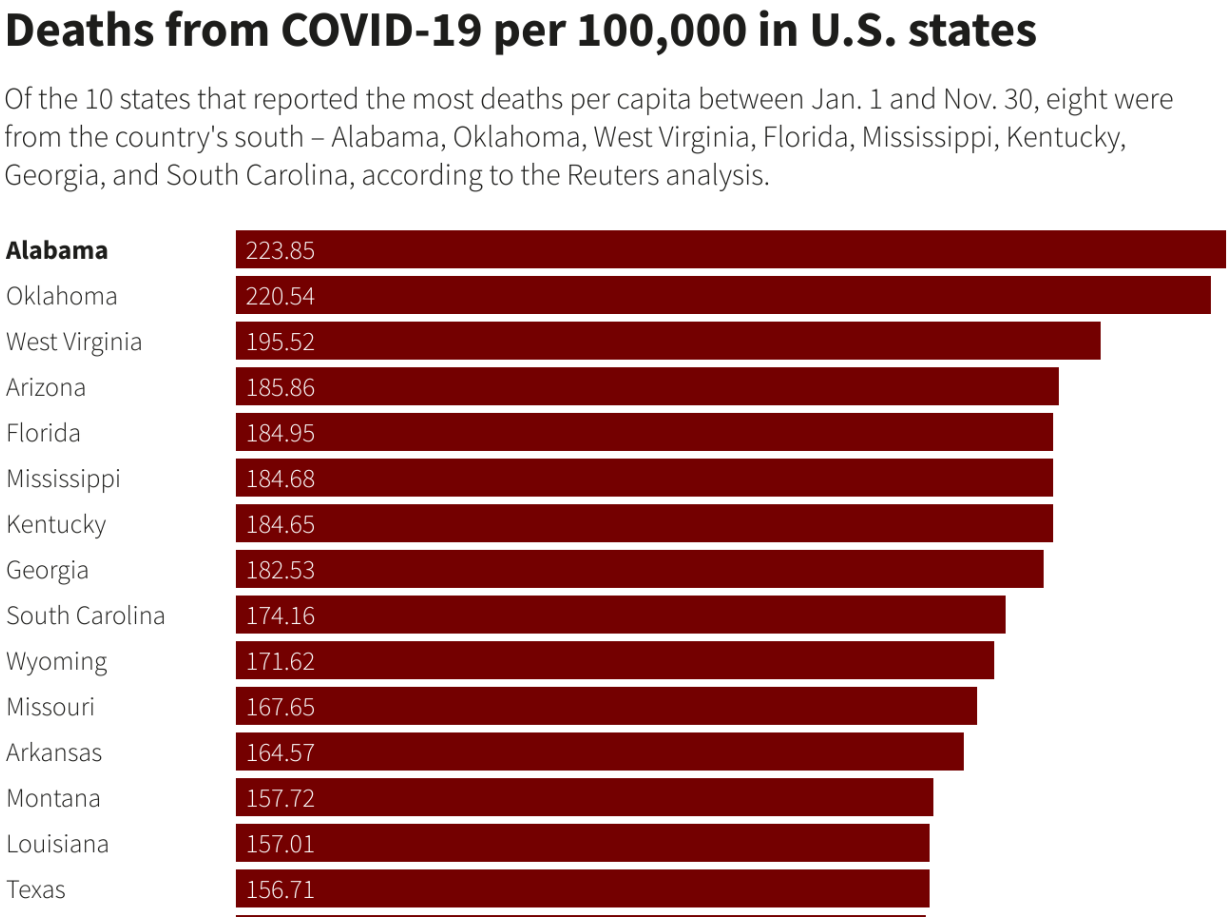
Step 2: Visualize the Data

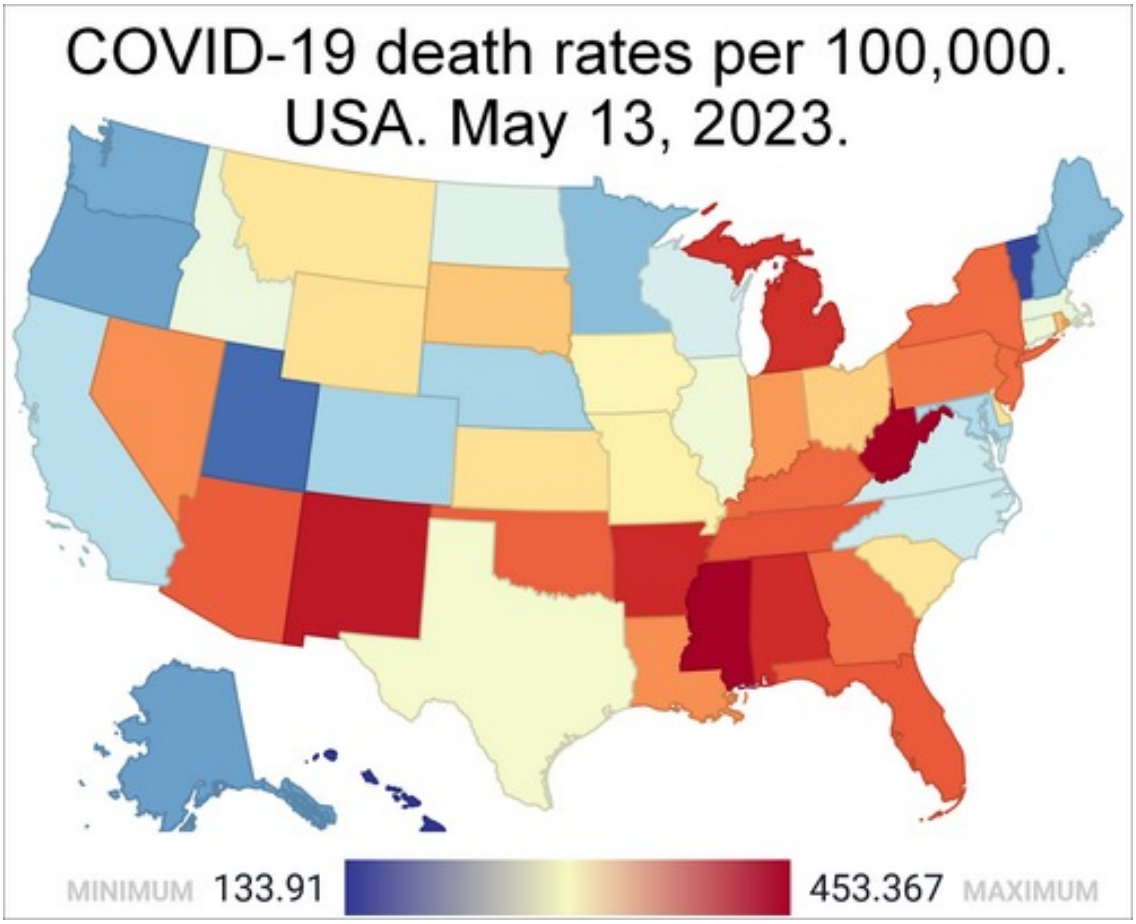
- Spot patterns, check distributions
- Communication is key
- Even the best data can be wasted without good communication

A great analysis means little if no one can understand it!

COVID-related deaths per capita in the U.S.

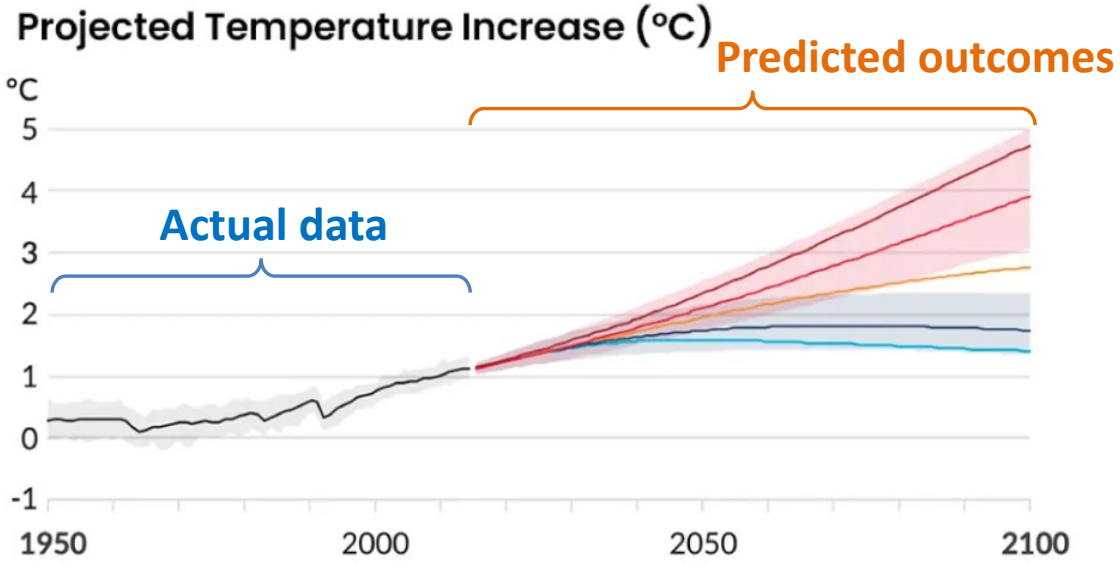
"Alabama had the highest deaths per capita, followed by Oklahoma and only slightly less in West Virginia. Then, falling lower, is Arizona and Florida, Mississippi and Kentucky, which have about the same rate, then Georgia, South Carolina, and finally Wyoming. All other states had far lower rates."





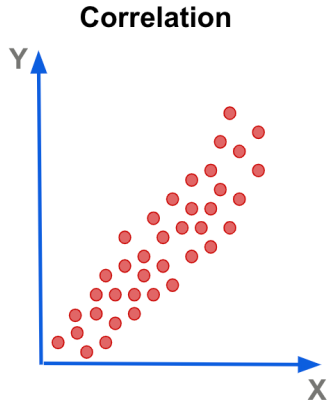
Step 3: Analyze and Model

- Use statistics and models to find meaning
- Build models for prediction or explanation



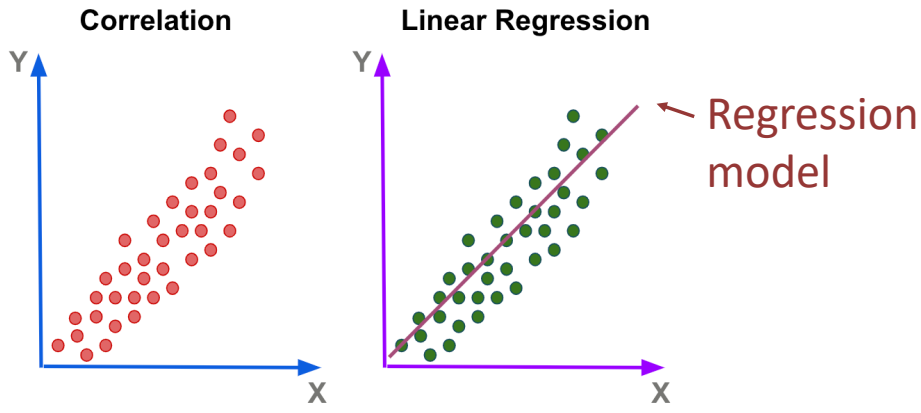
Step 4: Communicate and Refine

- Visuals and models help communicate insights
- May lead you to adjust your data or question
- Data science is a feedback process



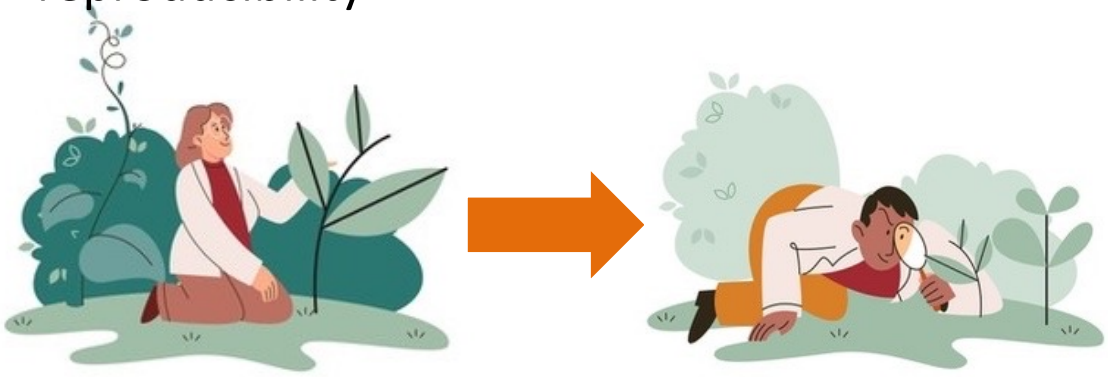
Step 4: Communicate and Refine

- Visuals and models help communicate insights
- May lead you to adjust your data or question
- Data science is a feedback process



Reproducibility in Data Science

- Science must be reproducible to be trusted
  - Many results fail to replicate — we can do better
- Use code, scripts, and standards to ensure reproducibility



Summary: What is Data Science?

- It's not about data or tools
  - it's about questions
- Data science = data-driven problem solving
- Requires
  - collaboration
  - curiosity
  - communication
- This course builds your practical data science skills

Skills Learning

code-along lecture

Juvenile red-tailed monkeys. Bigodi, Uganda 2022

Skills Application

laboratory exercise

Juvenile red-tailed monkeys. Bigodi, Uganda 2022

Week 1 - Skills Application

Working on **your own** dataset**:** create a Markdown file which you will be using continuously and adding to each week. Name it: ***Lastname_Firstname_Data.Rmd*** (in your working directory)

Working on **palmer_penguins_raw** dataset**:** create a Markdown file just for this week. Name it: ***Lastname_Firstname_Week1.Rmd*** (in Week 1)

0. Import the dataset into the environment and save as an object.
1. Identify which variables have **missing values**.
2. Check for **misclassified** columns (e.g., text stored as factors, numbers stores as text).
3. Convert at least one column to the correct type.
4. Get **summary statistics** for two variables of different classes.
5. Create one **bar plot** (categorical variable) and one **histogram** (numeric variable). Add a **title** and **axis labels** to each plot.
6. **Knit** the entire script into an HTML file; **save** to your working directory; **upload** to:

   Google Drive > *R Course Materials for Students* > *Submissions*