

INTRODUCTION TO DATA SCIENCE

Statistics & Data Science Unit

Grades 8–11 | No Coding Required | 3-Week Unit | 15 Lessons

15 Detailed Lesson Plans	3 Real-World Dataset Guides	Data Visualization Practice Set
10 Correlation vs. Causation Cards	Data Ethics Discussion Guide	Unit Project Rubric

A comprehensive, no-code introduction to the world of data science. Students explore data collection, cleaning, visualization, and analysis through real-world datasets and hands-on activities.

UM Printables
Premium Educational Resources

TABLE OF CONTENTS

Unit Overview

WEEK 1 — Foundations of Data Science

- Lesson 1: What Is Data Science?
- Lesson 2: Types of Data
- Lesson 3: Data Collection Methods
- Lesson 4: Surveys, Samples & Bias
- Lesson 5: Organizing & Storing Data

WEEK 2 — Working with Data

- Lesson 6: Data Cleaning Basics
- Lesson 7: Handling Missing & Inconsistent Data
- Lesson 8: Introduction to Data Visualization
- Lesson 9: Bar Charts & Pie Charts
- Lesson 10: Histograms & Scatter Plots

WEEK 3 — Analysis, Ethics & Project

- Lesson 11: Measures of Center
- Lesson 12: Correlation vs. Causation
- Lesson 13: Data Ethics & Privacy
- Lesson 14: Dataset Investigation (Project Day 1)
- Lesson 15: Presenting Findings (Project Day 2)

Real-World Dataset Investigation Guide

- Dataset 1: Student Screen Time & Grades
- Dataset 2: Fast Food Nutrition Comparison
- Dataset 3: City Climate & Population

Data Visualization Practice Set

Correlation vs. Causation Case Study Cards

Data Ethics Discussion Guide

Unit Project Rubric

UNIT OVERVIEW

Introduction to Data Science

This three-week unit introduces students in Grades 8–11 to the foundational concepts of data science — no coding required. Through real-world examples, hands-on activities, and guided investigations, students learn what data scientists do, how data is collected and cleaned, how to create and interpret data visualizations, and how to think critically about the claims data can (and cannot) support.

Grades	8–11
Duration	3 Weeks (15 Lessons × ~50 minutes)
Prerequisites	Basic arithmetic and graph-reading skills
Technology	No coding; spreadsheet software optional
Assessment	Ongoing formative + summative dataset investigation project

Essential Questions

- What does a data scientist do, and why is data science important in today's world?
- How do we collect, organize, and clean data so it is ready for analysis?
- What stories can data tell us — and what stories can it not tell?
- How do visualizations help us understand and communicate data effectively?
- Why is it critical to distinguish correlation from causation?
- What ethical responsibilities come with collecting and using people's data?

Learning Objectives

1. Define data science and describe the role of a data scientist.
2. Classify data as quantitative or qualitative and identify subtypes (discrete, continuous, nominal, ordinal).
3. Explain common data collection methods and evaluate their strengths and weaknesses.
4. Identify sources of bias in data collection and sampling.
5. Describe the purpose and process of data cleaning.
6. Create and interpret bar charts, pie charts, histograms, and scatter plots.
7. Calculate and interpret mean, median, and mode for datasets.
8. Distinguish between correlation and causation using real-world examples.
9. Discuss ethical issues related to data privacy, consent, and representation.

10. Conduct a guided real-world dataset investigation and present findings.

Materials Needed

- Printed student handouts (included in this resource)
- Graph paper or blank paper for sketching visualizations
- Colored pencils or markers
- Spreadsheet software (Google Sheets or Excel) — optional but recommended
- Calculator
- Correlation vs. Causation Case Study Cards (included)

Unit at a Glance

Week	Theme	Lessons	Key Focus
1	Foundations of Data Science	1–5	What is data science, types of data, collection methods, sampling & bias, organizing data
2	Working with Data	6–10	Data cleaning, missing data, visualization basics, bar & pie charts, histograms & scatter plots
3	Analysis, Ethics & Project	11–15	Measures of center, correlation vs. causation, data ethics, dataset investigation project

Differentiation Tips

Support: Provide pre-filled templates for visualizations; work through dataset examples as a class before independent practice; offer sentence starters for written responses.

Extension: Challenge advanced students to identify additional variables, create multi-variable scatter plots, or research real data science careers and present findings to the class.

ELL: Use visual vocabulary cards with images; pair with bilingual peers; allow responses in home language with translation support.

WEEK 1 — Foundations of Data Science

This week students build a foundational understanding of what data science is, the types of data they will encounter, how data is collected, the importance of sampling methods and bias awareness, and how to organize and store data effectively.

Objectives

- Define data science and explain what data scientists do.
- Identify real-world applications of data science.
- Describe the data science workflow: Ask, Collect, Clean, Explore, Communicate.

Materials

Chart paper or whiteboard; sticky notes (one per student); student notebooks

Key Vocabulary

Data: Facts, figures, and information collected for analysis.

Data Science: The interdisciplinary field of extracting knowledge and insights from data using scientific methods.

Data Scientist: A professional who collects, analyzes, and interprets data to solve problems.

Dataset: An organized collection of data, typically arranged in rows and columns.

Insight: A meaningful finding or understanding gained from analyzing data.

Warm-Up (5 min)

Display the question: *“How do you think Netflix knows what movies to recommend?”* Have students write their ideas on sticky notes and place them on the board. Briefly discuss responses, guiding students toward the idea that data (viewing history, ratings, similar users) drives recommendations. Explain that this is one example of data science in action.

Direct Instruction (15 min)

Explain that **data science** is the process of using data to answer questions and make decisions. Data scientists work at the intersection of statistics, technology, and domain expertise. They follow a structured workflow:

Step	Description
1. Ask	Identify a clear question or problem to solve.
2. Collect	Gather relevant data from appropriate sources.
3. Clean	Fix errors, remove duplicates, and handle missing values.
4. Explore	Analyze and visualize the data to find patterns and trends.
5. Communicate	Share findings through charts, reports, and presentations.

Share real-world examples: weather forecasting (analyzing historical patterns to predict future conditions), healthcare (identifying disease risk factors from patient records), sports analytics (evaluating player performance using game statistics), transportation (optimizing delivery routes with traffic data), and social media (personalizing content feeds based on user behavior). Ask students to suggest additional examples.

Guided Practice (10 min)

Present the scenario: “A school principal wants to know if students are getting enough sleep.” As a class, walk through each step of the workflow: What question would we ask? How would we collect data (surveys, sleep diaries)? What cleaning might be needed (inconsistent units, missing entries)? How would we explore and visualize findings? How would we communicate results? Record student ideas on the board in a five-column chart.

Independent Practice (10 min)

Students choose one scenario and write a brief plan addressing all five workflow steps:

- A city planner wants to reduce traffic accidents at dangerous intersections.
- A coach wants to determine the best training schedule for the track team.
- A library wants to find out which books are most popular among teens.

Closure (5 min)

Exit Ticket: Students write one sentence defining data science and one example of a question a data scientist might investigate.

Differentiation

Support: Provide a pre-labeled five-step graphic organizer for the independent practice.

Extension: Have students research a real data science application in an industry they choose and present a one-minute summary to the class.

Objectives

- Create and interpret histograms for quantitative data.
- Create and interpret scatter plots showing the relationship between two quantitative variables.
- Distinguish between a histogram and a bar chart.
- Identify positive, negative, and no-trend patterns in scatter plots.

Materials

Graph paper; colored pencils/markers; whiteboard; student notebooks

Key Vocabulary

Histogram: A chart that displays the distribution of quantitative data by grouping values into consecutive, non-overlapping ranges called bins.

Bin (Interval): A range of values on the x-axis of a histogram; each bin includes the lower bound and excludes the upper bound.

Scatter Plot: A chart that displays the relationship between two quantitative variables, with one variable on the x-axis and the other on the y-axis.

Positive Trend: As one variable increases, the other tends to increase.

Negative Trend: As one variable increases, the other tends to decrease.

No Trend: No clear pattern of increase or decrease between the two variables.

Warm-Up (5 min)

Ask: “How could we show the distribution of test scores in a class? We don’t just want to compare categories — we want to see how the scores spread out.” Introduce the idea that we need a chart for quantitative data ranges.

Direct Instruction (15 min)

Histograms: Explain that histograms look similar to bar charts but are fundamentally different. Histograms display the distribution of **quantitative** data by grouping values into bins (ranges). The bars in a histogram **touch** because the data is continuous. Demonstrate with test score data:

Score Range	Number of Students
50–59	2
60–69	5
70–79	12
80–89	9
90–100	4

Sketch the histogram on the board, emphasizing that bars touch and there are no gaps. Ask: “What is the most common score range?” (70–79) and “How does this differ from a bar chart?”

Feature	Bar Chart	Histogram
Purpose	Compare categories	Show distribution of quantitative data
Data Type	Qualitative (Categorical)	Quantitative (Continuous)
Bar Spacing	Bars are separated by gaps	Bars touch (no gaps)
X-Axis Order	Can be in any order	Must be in numerical order

Scatter Plots: Explain that scatter plots show the relationship between two **quantitative** variables. Each point represents one observation. Demonstrate with study hours vs. test scores. Plot points on the board. Identify the pattern: as study hours increase, test scores tend to increase — this is a **positive trend**. Contrast with a **negative trend** (e.g., absences vs. grades) and **no trend** (e.g., shoe size vs. test scores). Emphasize that trends describe general patterns; individual points may not follow the trend exactly.

Guided Practice (10 min)

Provide the following data. As a class, identify whether a histogram or scatter plot is appropriate, sketch the chart, and interpret it:

1. Heights of 30 students measured in inches. (*Histogram — one quantitative variable; shows distribution*)
2. Shoe size vs. height for 15 students. (*Scatter plot — two quantitative variables; expected positive trend*)

Independent Practice (10 min)

Students are given two small datasets:

1. A list of 20 daily temperatures (°F). Students create a histogram with appropriate bins and write one observation about the distribution.
2. A list of 10 (height, arm span) pairs in centimeters. Students create a scatter plot, identify the trend, and write one sentence describing the relationship.

Closure (5 min)

Quick Write: Explain the difference between a bar chart and a histogram in 2–3 sentences. Then, describe what a positive trend looks like in a scatter plot.

Differentiation

Support: Provide pre-labeled axes and bin ranges for the histogram; provide pre-labeled axes and data points plotted for the scatter plot (students trace and interpret).

Extension: Challenge students to create a scatter plot that shows a negative trend using their own invented but realistic data (e.g., miles driven vs. fuel remaining). They write a paragraph explaining why the trend is negative.

WEEK 3 — Analysis, Ethics & Project

This week students learn to summarize data using measures of center, critically evaluate the relationship between correlation and causation, explore the ethical responsibilities of working with data, and apply all their skills in a two-day real-world dataset investigation project.

Objectives

- Calculate the mean, median, and mode of a dataset.
- Explain how outliers affect the mean more than the median.
- Choose the most appropriate measure of center for a given dataset.

Materials

Calculators; whiteboard; student notebooks

Key Vocabulary

Mean: The sum of all values divided by the number of values; commonly called the average.

Median: The middle value when the data is arranged in order; if there is an even number of values, it is the average of the two middle values.

Mode: The value that appears most frequently in a dataset; a dataset can have one mode, more than one mode, or no mode.

Outlier: A data point that is significantly different from the other values in the dataset.

Measure of Center: A single value that summarizes or represents the center of a dataset.

Warm-Up (5 min)

Display five numbers on the board: 2, 4, 5, 5, 7. Ask students to find the average (mean) and the middle number (median). Mean = $(2+4+5+5+7)/5 = 23/5 = 4.6$. Median = 5. Tell students these are two different ways to describe the "center" of the data.

Direct Instruction (15 min)

Define and demonstrate each measure of center using a clear dataset:

Measure	How to Find	Example	Strength	Weakness
Mean	Sum all values and divide by the count.	$(2+4+5+5+7)/5 = 4.6$	Uses every data point; easy to calculate.	Pulled toward extreme values (outliers).
Median	Arrange in order; find the middle value.	Ordered: 2, 4, 5, 5, 7 → Median = 5	Not affected by outliers; good for skewed data.	Does not use every data point in the calculation.
Mode	Find the most frequent value.	5 appears twice; Mode = 5	The only measure that works for categorical data.	May not exist or may not be unique.

Now introduce an **outlier**. Add the value 100 to the dataset: 2, 4, 5, 5, 7, 100. Recalculate together:

- **New Mean:** $(2+4+5+5+7+100)/6 = 123/6 = 20.5$ (jumped from 4.6 to 20.5)
- **New Median:** $(5+5)/2 = 5$ (barely changed)
- **New Mode:** 5 (unchanged)

Conclude: The mean is **sensitive to outliers**; the median and mode are **resistant to outliers**. When data has extreme outliers, the median is usually a better representation of the "typical" value.

Guided Practice (10 min)

Students calculate mean, median, and mode for two datasets, then decide which measure is most appropriate:

1. Test scores: 72, 85, 90, 88, 91, 78. (*Mean=84, Median=86.5, Mode=none. Mean is fine — no extreme outliers.*)
2. Salaries at a small company: \$30K, \$32K, \$35K, \$38K, \$250K (the owner). (*Mean=\$77K, Median=\$35K, Mode=none. Median is better — the \$250K outlier inflates the mean, which does not represent a typical salary.*)

Independent Practice (10 min)

Students create their own small dataset (6–8 values) that includes one outlier. They calculate the mean and median, then write one sentence explaining which measure better represents the data and why.

Closure (5 min)

Exit Ticket: A real estate agent says the average home price in a neighborhood is \$750,000, but the median is \$400,000. What does this tell you about the distribution of home prices? (*A few very expensive homes are pulling the mean up; the median gives a more accurate picture of a typical home price.*)

Differentiation

Support: Provide a step-by-step calculation guide: (1) List numbers in order, (2) Add them up, (3) Divide by count, (4) Find the middle, (5) Find the most frequent.

Extension: Ask students to research the concept of "trimmed mean" (removing a percentage of extreme values before averaging) and explain when it might be useful.

Objectives

- Define correlation and causation and explain the difference.
- Identify the role of confounding (third) variables in creating spurious correlations.
- Evaluate claims to determine whether correlation implies causation.

Materials

Correlation vs. Causation Case Study Cards (included later in this resource); whiteboard; student notebooks

Key Vocabulary

Correlation: A statistical relationship between two variables — when one changes, the other tends to change as well.

Causation: A relationship where one variable directly causes a change in another variable.

Confounding Variable (Third Variable): An unseen variable that influences both of the variables being studied, creating a false impression of a direct cause-and-effect relationship.

Spurious Correlation: A correlation between two variables that is not due to a direct causal link but to coincidence or a confounding variable.

Warm-Up (5 min)

Display: *“Ice cream sales and drowning incidents both increase in the summer. Does eating ice cream cause drowning?”* Students discuss with a partner. Guide them to see that a third variable — **hot weather** — causes both. This is the key difference between correlation and causation.

Direct Instruction (15 min)

Explain: **Correlation** means two variables move together. **Causation** means one variable directly makes the other change. The phrase *“correlation does not imply causation”* is one of the most important principles in data science. Just because two things happen together does not mean one causes the other.

Present three reasons why two variables might be correlated without one causing the other:

Reason	Explanation	Example
Confounding Variable	A third variable causes both.	Ice cream sales and drowning both increase because of hot weather (the confounding variable).
Coincidence	The correlation is due to random chance.	The number of people who drowned by falling into a pool correlates with films starring Nicolas Cage (a famous spurious correlation from real data).
Reverse Causation	The presumed effect actually causes the presumed cause.	Does stress cause poor sleep, or does poor sleep cause stress? The direction of causation may be reversed, or it may be cyclical.

Key Principle

To establish causation, data scientists need controlled experiments, not just observational data. A correlation is a clue worth investigating, but it is not proof of a cause-and-effect relationship.

CORRELATION VS. CAUSATION CASE STUDY CARDS

Correlation vs. Causation Case Study Cards

These 10 case study cards present real-world scenarios where two variables are correlated. For each card, discuss with a partner: Does one variable cause the other, or is a confounding (third) variable responsible? Cut out the cards for small-group activities.

Card 1: Ice Cream & Sunburns

A study finds that on days when ice cream sales are high, the number of people getting sunburns is also high.

Question: Does eating ice cream cause sunburns?

Answer: No. The confounding variable is sunny/hot weather. People eat more ice cream and spend more time in the sun on hot days, causing both to increase together.

Card 2: Shoe Size & Reading

A study of elementary school students shows that children with larger shoe sizes tend to have higher reading scores.

Question: Does having bigger feet make you a better reader?

Answer: No. The confounding variable is age. Older children have larger feet and are also in higher grade levels with more advanced reading skills.

Card 3: Firefighters & Damage

Data shows that the more firefighters sent to put out a fire, the greater the financial damage to the property.

Question: Do firefighters cause more damage to the property?

Answer: No. The confounding variable is the size and severity of the fire. Larger, more destructive fires require more firefighters and naturally cause more damage.

Card 4: Storks & Births

In some European countries, areas with higher stork populations also have higher human birth rates.

Question: Do storks deliver babies?

Answer: No. The confounding variable is rural vs. urban development. Rural areas tend to have more storks and traditionally higher birth rates than urban areas.

Card 5: Winter Coats & Accidents

During months when winter coat sales are highest, the number of car accidents also increases significantly.

Question: Do winter coats cause car accidents?

Answer: *No. The confounding variable is winter weather (snow and ice). Cold weather causes people to buy coats, and icy roads cause more accidents.*

Card 6: Night Lights & Vision

A study finds that children who sleep with a night light are more likely to develop nearsightedness (poor distance vision).

Question: Do night lights cause nearsightedness?

Answer: *No. The confounding variable is genetics. Nearsighted parents are more likely to use night lights (to see at night) and also pass the nearsightedness trait to their children.*

Card 7: Churches & Crime

Cities with a higher number of churches tend to have higher crime rates.

Question: Do churches cause crime?

Answer: *No. The confounding variable is population size. Larger cities have more churches and more crime simply because there are more people.*

Card 8: Breakfast & Grades

Students who eat breakfast tend to have higher grades than students who skip breakfast.

Question: Does eating breakfast cause higher grades?

Answer: *Not necessarily. Confounding variables include stress, lack of sleep, and socioeconomic factors. It could also be reverse causation: stressed students may rush and skip breakfast.*

Card 9: Car Weight & Fuel Use

Heavier vehicles tend to use more fuel per mile than lighter vehicles.

Question: Does a car's weight cause it to use more fuel?

Answer: *Yes (Causation). It takes more energy to move a heavier object. Therefore, car weight directly causes higher fuel consumption.*

Card 10: Fertilizer & Crop Yield

Farmers who apply more fertilizer to their fields tend to harvest larger crop yields.

Question: Does fertilizer cause crops to grow better?

Answer: *Yes (Causation). Fertilizer provides essential nutrients directly to the plants, causing them to grow larger and produce more, assuming other conditions are also met.*

Total Score: ____ / 24

Thank You for Your Purchase!

Your support means the world to me! If you found this resource helpful in your classroom, I would be incredibly grateful if you could take a moment to leave a review.

How to Leave a Review:

1. Go to your **My Purchases** page on Teachers Pay Teachers.
2. Find this resource and click **Leave a Review**.
3. Share your honest feedback — it helps other teachers and earns you TPT credits!

UM Printables

Premium Educational Resources
