DATA ANALYSIS

with Real-World Datasets

Building Statistical Literacy Through Authentic Data Exploration

GRADES 7 – 10

- 10 Real-World Dataset Activity Packs
- Comprehensive Graphing Worksheets
- Interpretation & Analysis Questions
- Class Discussion Guides
- Assessment with Complete Answer Key

UM Printables

TABLE OF CONTENTS

Teacher Introduction

Thank you for choosing **Data Analysis with Real-World Datasets**. This premium resource is designed to build genuine statistical literacy in students from Grades 7 through 10 by immersing them in authentic, meaningful data drawn from the world around them.

Why This Resource?

Traditional statistics units often rely on contrived datasets that fail to engage students. This resource uses **10 real-world datasets**—from NBA scoring and global climate data to social media usage and fast-food nutrition—so students see the immediate relevance of statistical thinking. Each activity pack is self-contained and progressively builds skills in measures of centre, measures of spread, and data displays.

What's Inside

- **10 Dataset Activity Packs** — Each with a real dataset, data table, guided analysis questions, charting task, and interpretation prompts.
- **4 Graphing Worksheets** — Focused practice on bar charts, line graphs, scatter plots, and box-and-whisker plots.
- **Interpretation Questions** — Higher-order thinking prompts that require students to read, interpret, and draw conclusions from data displays.
- **Class Discussion Guides** — Facilitation prompts for rich mathematical discourse around each dataset's real-world context.
- **Comprehensive Assessment** — A summative test with multiple-choice, short answer, and extended-response items, plus a full answer key.

How to Use This Resource

Step 1 — Teach the Concepts Use the Statistical Concepts Review section to introduce or refresh measures of centre, spread, and display types.

Step 2 — Dataset Exploration Assign activity packs individually or in rotation. Students work through the data table, compute statistics, create graphs, and answer questions.

Step 3 — Class Discussion Use the discussion guides to facilitate whole-class or small-group discourse about each dataset's context and implications.

Step 4 — Graphing Practice Assign the targeted worksheets for additional practice with specific display types.

Step 5 — Assess Administer the comprehensive assessment and use the answer key for efficient grading.

Differentiation Tips

- **Scaffolded:** Provide the pre-computed statistics and focus on interpretation.
- **On-Level:** Students compute all statistics and create graphs from scratch.
- **Extension:** Students collect their own related data and compare with the dataset, or explore the impact of outliers on different measures.

Statistical Concepts Review

Measures of Centre

Measures of centre describe the “typical” or “middle” value of a dataset. The three main measures are the **mean**, **median**, and **mode**.

Mean (Average)

The mean is the sum of all values divided by the number of values.

Formula: Mean = (Σ x) ÷ n

When to use: Best for symmetric data without extreme outliers.

Limitation: Strongly affected by outliers.

Median (Middle Value)

The median is the middle value when data are arranged in order.

If n is odd, the median is the centre value. If n is even, it is the mean of the two centre values.

When to use: Best for skewed data or data with outliers.

Advantage: Resistant to outliers.

Mode (Most Frequent Value)

The mode is the value that appears most often in the dataset.

A dataset can be unimodal (one mode), bimodal (two modes), or multimodal.

When to use: Best for categorical data or when identifying the most common value.

Advantage: The only measure of centre applicable to categorical data.

Measures of Spread

Measures of spread describe how much the values in a dataset vary. They tell us whether the data points cluster tightly around the centre or are widely dispersed.

Range

The range is the difference between the maximum and minimum values.

Formula: Range = Maximum – Minimum

Limitation: Only considers two values; sensitive to outliers.

Interquartile Range (IQR)

The IQR is the range of the middle 50% of the data.

Formula: IQR = Q3 – Q1

Q1 (first quartile) = 25th percentile; Q3 (third quartile) = 75th percentile.

Advantage: Resistant to outliers; used to identify them.

Mean Absolute Deviation (MAD)

The MAD is the average distance of each data point from the mean.

Formula: MAD = (Σ |x – mean|) ÷ n

When to use: Provides an intuitive measure of typical deviation from the mean.

Standard Deviation

The standard deviation measures the typical distance of data points from the mean, using squared differences.

Formula: $\sigma = \sqrt{[\Sigma(x - \text{mean})^2 \div n]}$

When to use: Preferred in advanced statistics; used with the mean for symmetric data.

A larger standard deviation indicates greater spread.

Data Displays

Different data displays highlight different aspects of a dataset. Choosing the right display is an essential statistical skill.

Bar Chart

Compares categorical data. Each bar represents a category's frequency or value. Best for comparing discrete groups.

Histogram

Displays the distribution of continuous numerical data in bins. Shows shape, centre, and spread at a glance.

Line Graph

Shows change over time (time series). Points connected by lines reveal trends, patterns, and rates of change.

Scatter Plot

Displays the relationship between two quantitative variables. Patterns reveal correlation (positive, negative, or none).

Box-and-Whisker Plot

Summarises the five-number summary (min, Q1, median, Q3, max). Excellent for comparing distributions and spotting outliers.

Pie Chart

Shows parts of a whole as proportional slices. Best for categorical data with few categories.

Outlier Detection Rule

A data point is considered an **outlier** if it falls:

• Below Q1 – 1.5 × IQR, OR

• Above Q3 + 1.5 × IQR

Outliers should be investigated—they may be errors or genuinely unusual values.

Dataset 1: NBA Player Scoring Statistics

The table below shows the Points Per Game (PPG) for 10 NBA players during a recent season. Analyzing sports statistics helps teams evaluate player performance and make strategic decisions.

Player	Points Per Game (PPG)
L. James	28.9
K. Durant	29.1
S. Curry	29.4
J. Harden	21.0
G. Antetokounmpo	31.1
L. Doncic	33.9
J. Tatum	30.1
D. Booker	27.8
D. Mitchell	28.3
T. Young	26.2

Sports Stats Context

PPG (Points Per Game) is a common measure of a basketball player's scoring output. A high PPG indicates a strong scorer, but it's important to look at the spread of those numbers across the league to see if a player is truly exceptional.

Activity 1: Compute & Analyse

1. Calculate the **mean** PPG for these 10 players. Show your working.

2. Find the **median** PPG. How does it compare to the mean?

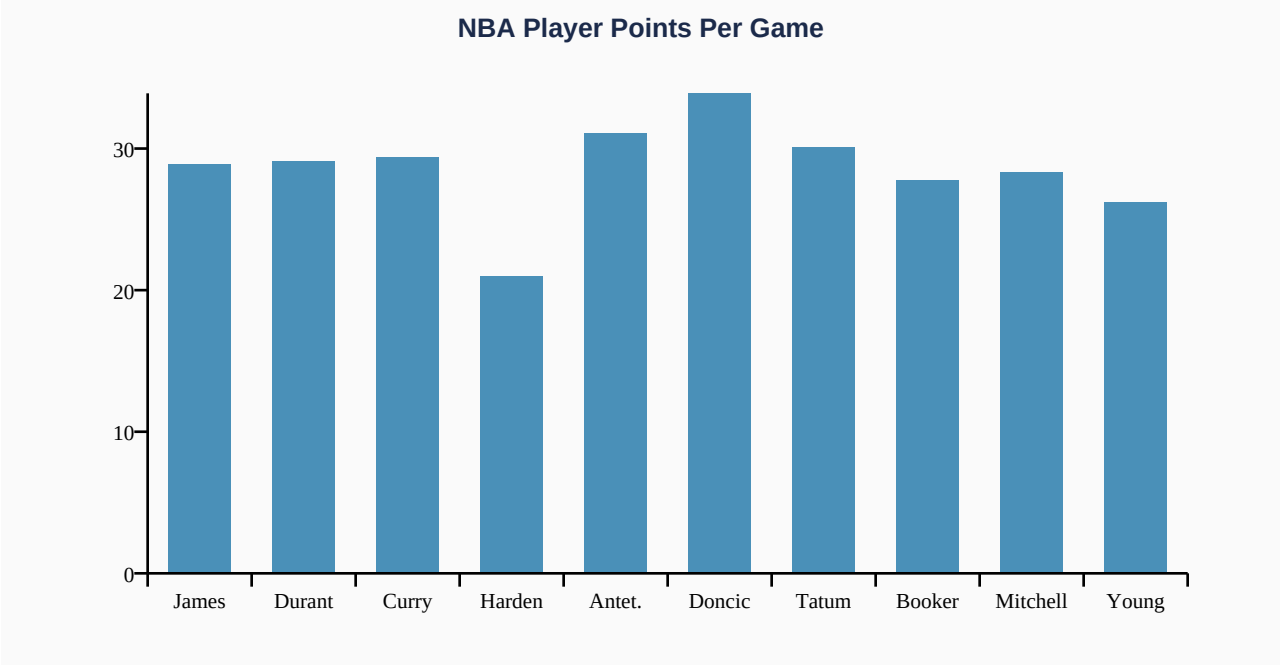
3. Is there a **mode** for this dataset? Explain why or why not.

4. Calculate the **range** of the PPG values.

5. Identify the **outlier(s)** using the IQR method (Q1 – 1.5×IQR and Q3 + 1.5×IQR). Does any player qualify as an outlier?

6. If a new player (Player K) with a PPG of 50.0 were added to the dataset, which measure of centre would be most affected: mean, median, or mode? Explain.

Activity 2: Graph It



7. Using the bar chart above, describe the distribution of PPG. Are the values clustered tightly, or is there significant variation?

8. Why is a bar chart appropriate for this data, but a histogram would NOT be?

Dataset 2: Global Temperature Anomalies

The table shows the global average surface temperature anomaly (in °C) relative to the 1951–1980 baseline. Positive values indicate warming above the baseline; negative values indicate cooling. Climate data like this is critical for understanding long-term environmental trends.

Year	Temperature Anomaly (°C)
1950	-0.16°C
1960	+0.03°C
1970	+0.01°C
1980	+0.26°C
1990	+0.45°C
2000	+0.42°C
2010	+0.72°C
2020	+1.02°C
2023	+1.17°C

Climate Data Context

Temperature **anomalies** are used instead of absolute temperatures because they provide a clearer picture of change over time. An anomaly of +1.0°C means the global average was 1 degree warmer than the baseline period.

Activity 1: Compute & Analyse

1. Calculate the **mean** temperature anomaly for the entire period (1950–2023).

2. Find the **median** anomaly. What does the difference between the mean and median tell you about the distribution?

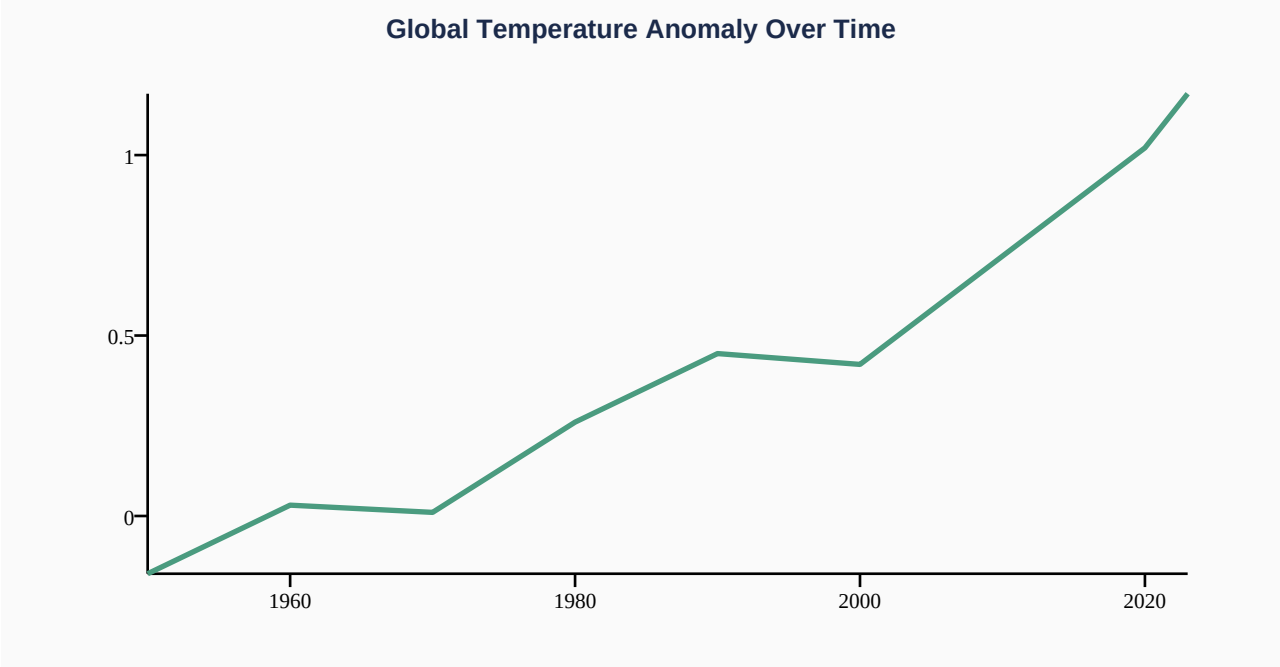
3. Calculate the **range** of the temperature anomalies.

4. Calculate the **mean anomaly** for just the years 1950–1980 and then for 1990–2023. Compare the two means. What conclusion can you draw?

5. Between which two consecutive decades did the **largest increase** in temperature anomaly occur?

6. If the year 1950 were removed from the dataset, how would the mean change? Explain your reasoning.

Activity 2: Graph It



7. Describe the overall trend shown in the line graph. Is it increasing, decreasing, or staying relatively flat?

8. Why is a line graph the most appropriate display for this type of time-series data?