# WEEK 03

INSTRUCTOR: YANAN WU

TA: KHADIJA NISAR

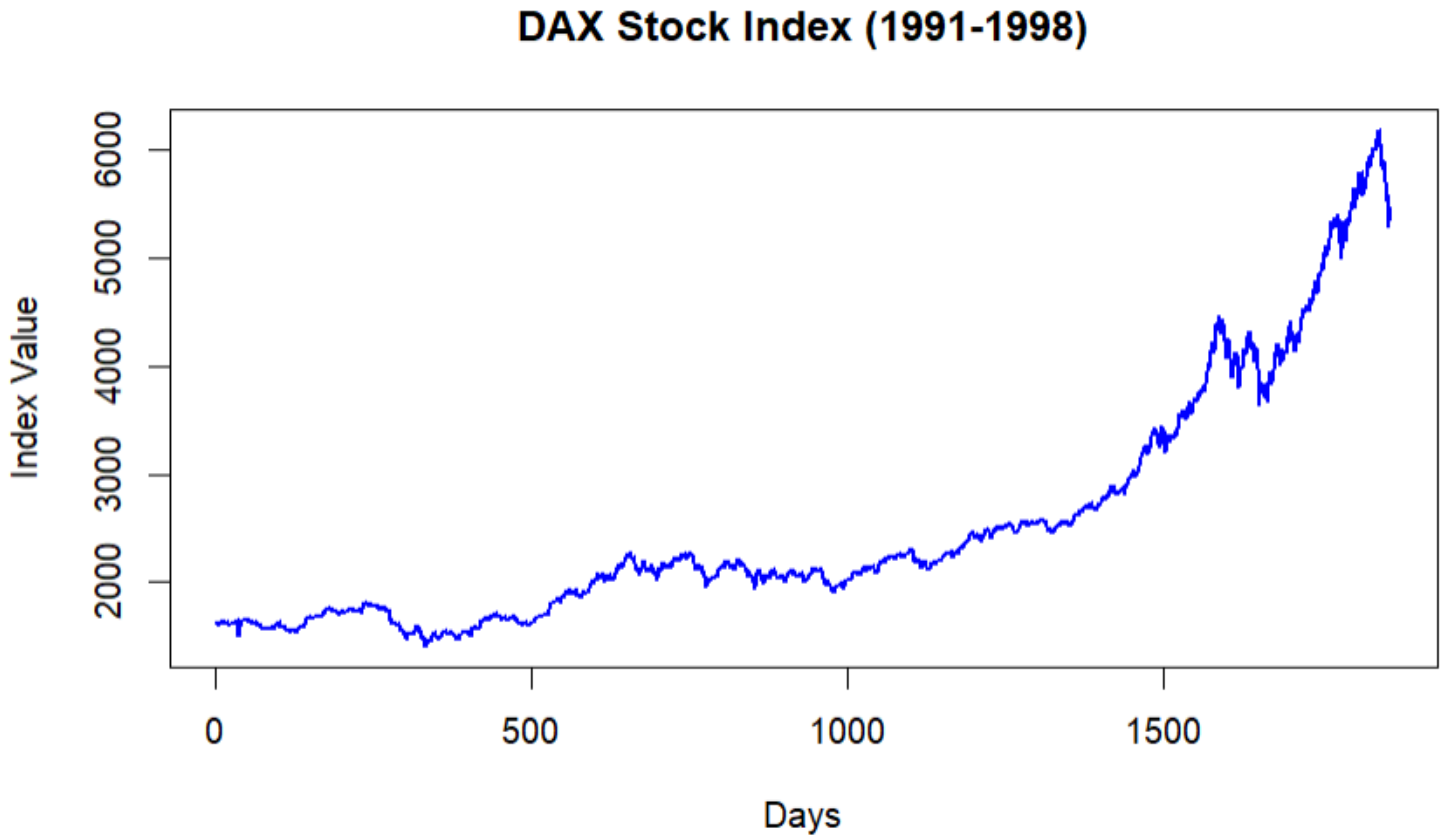
SPRING 2025

# 2.1.1
# DATA VISUALIZATION

# LINE GRAPH

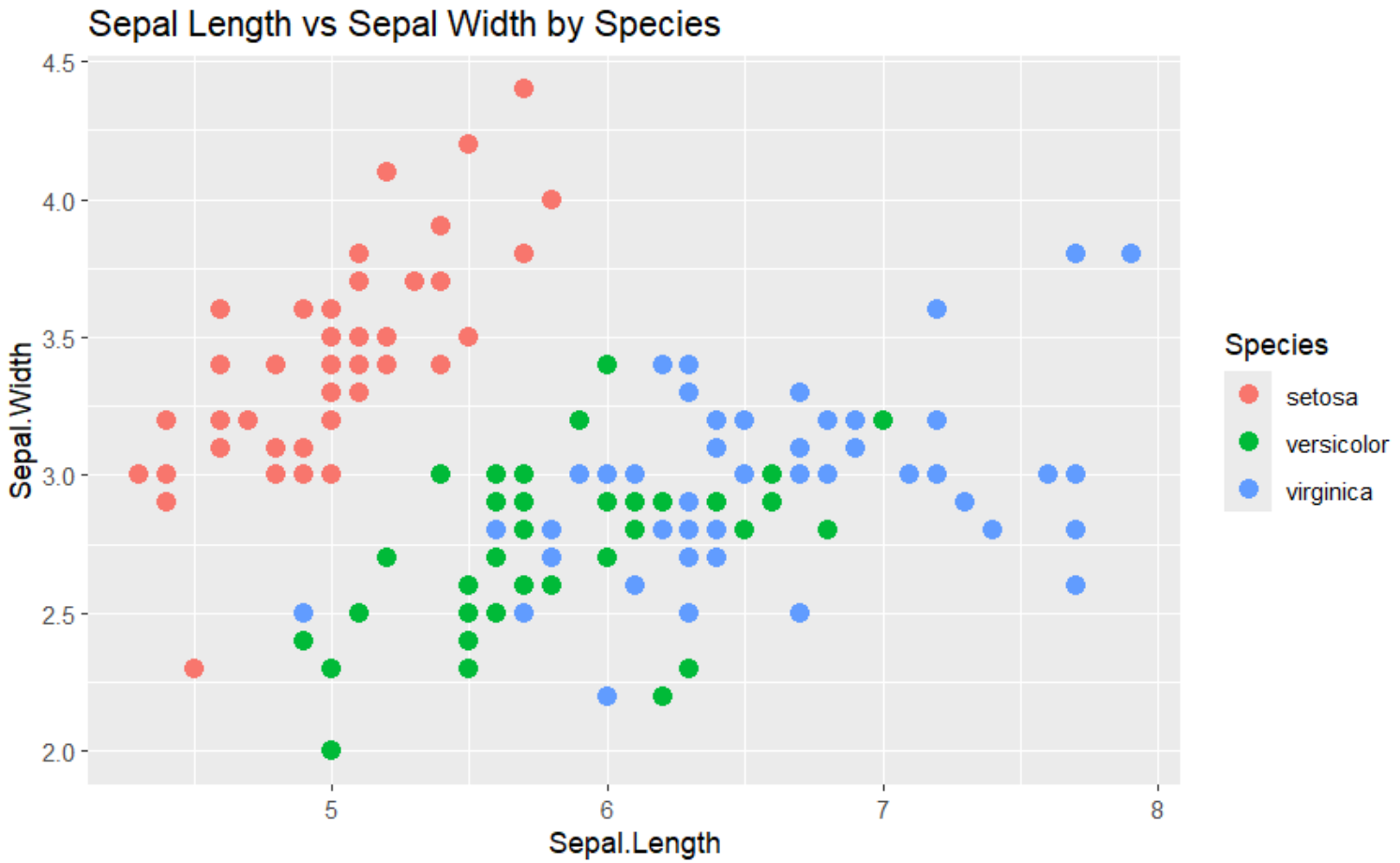
- At its core, **data visualization is about turning numbers into stories** — stories that people can understand, remember, and act upon. It's the bridge between raw data and human perception.

| DAX | SMI | CAC | FTSE |
|---|---|---|---|
| 1628.75 | 1678.1 | 1772.8 | 2443.6 |
| 1613.63 | 1688.5 | 1750.5 | 2460.2 |
| 1606.51 | 1678.6 | 1718.0 | 2448.2 |
| 1621.04 | 1684.1 | 1708.1 | 2470.4 |
| 1618.16 | 1686.6 | 1723.1 | 2484.7 |
| 1610.61 | 1671.6 | 1714.3 | 2466.8 |
| 1630.75 | 1682.9 | 1734.5 | 2487.9 |
| 1640.17 | 1703.6 | 1757.4 | 2508.4 |
| 1635.47 | 1697.5 | 1754.0 | 2510.5 |
| 1645.89 | 1716.3 | 1754.3 | 2497.4 |
| 1647.84 | 1723.8 | 1759.8 | 2532.5 |
| 1638.35 | 1730.5 | 1755.5 | 2556.8 |
| 1629.93 | 1727.4 | 1758.1 | 2561.0 |
| 1621.49 | 1733.3 | 1757.5 | 2547.3 |
| 1624.74 | 1734.0 | 1763.5 | 2541.5 |
| 1627.63 | 1728.3 | 1762.8 | 2558.5 |
| 1631.99 | 1737.1 | 1768.9 | 2587.9 |
| 1621.18 | 1723.1 | 1778.1 | 2580.5 |
| 1613.42 | 1723.6 | 1780.1 | 2579.6 |
| 1604.95 | 1719.0 | 1767.7 | 2589.3 |
| 1605.75 | 1721.2 | 1757.9 | 2595.0 |



# SCATTERPLOT

- Scatterplots are useful for visualizing the relationship between two numerical variables. They can also incorporate a third variable through additional visual elements, such as color, size, or shape of points



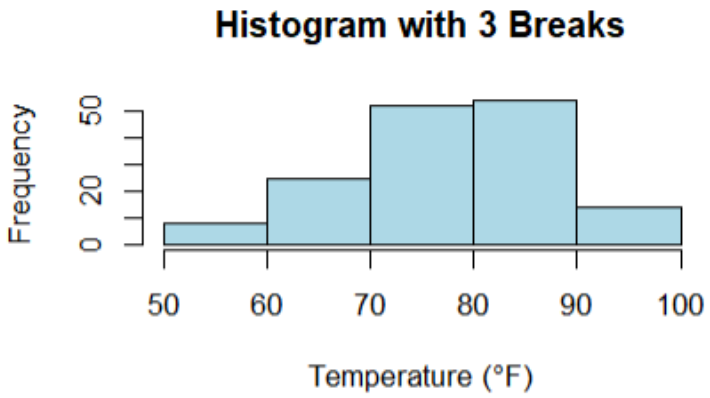

# HISTOGRAM

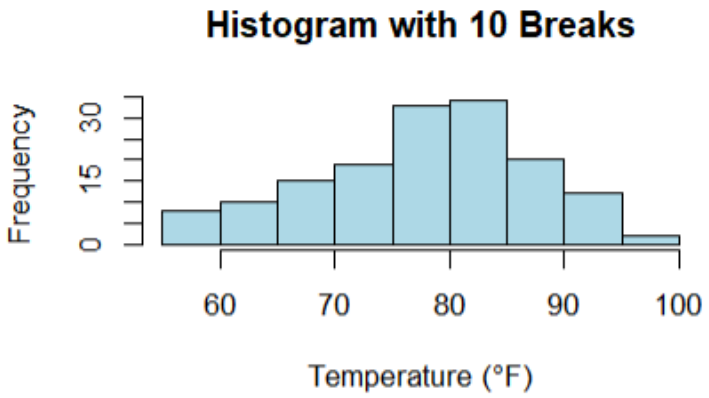
- A histogram displays data by grouping values into bins (ranges) and shows the frequency of values falling within each bin

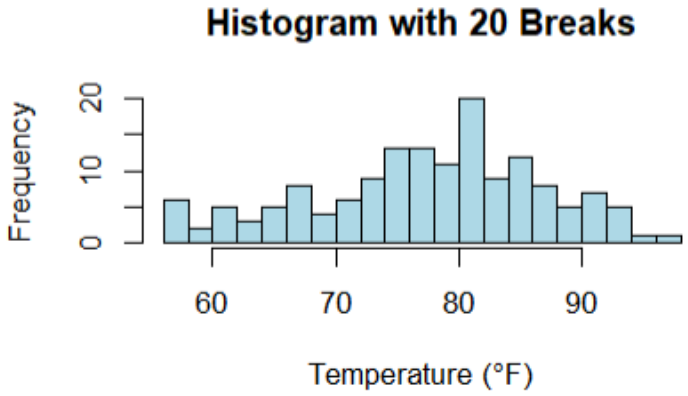
- Why Use a Histogram?
  - Quickly understand the overall distribution of a dataset.
  - Identify important characteristics, such as:
    a) Whether the data is **normally distributed**.
    b) If the data is **skewed** (left or right).
    c) **Gaps** or **clusters** in the data.
    d) The presence of **outliers**.

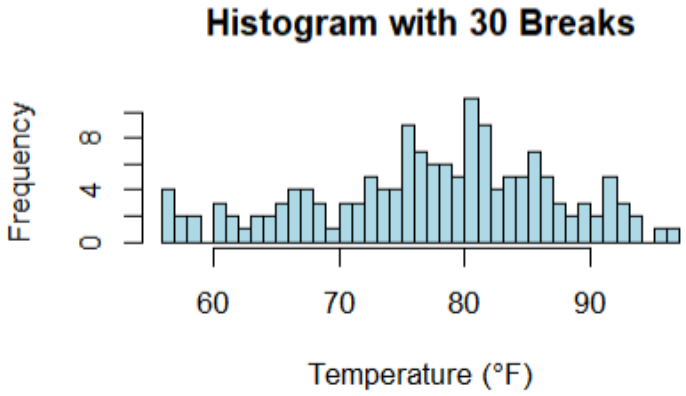
- What does the distribution of variable tell us?
  - Identify the reasonable range of values?
  - Highlights the most frequency occurring values?

Which one(s) of these histograms are informative? 🤔



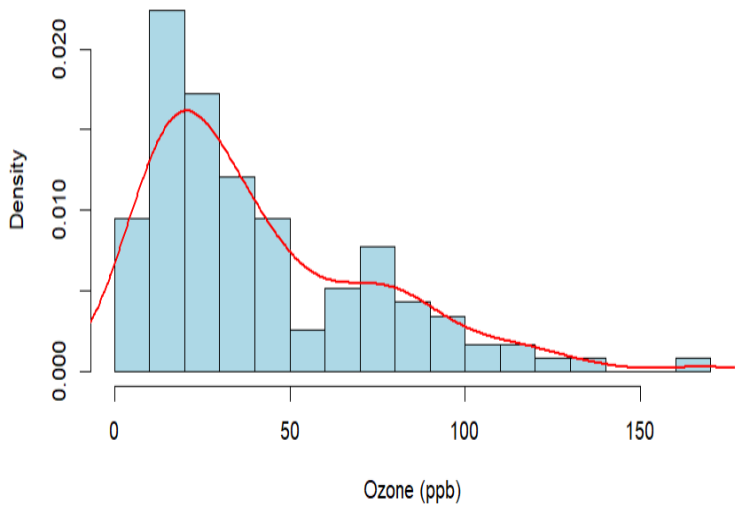



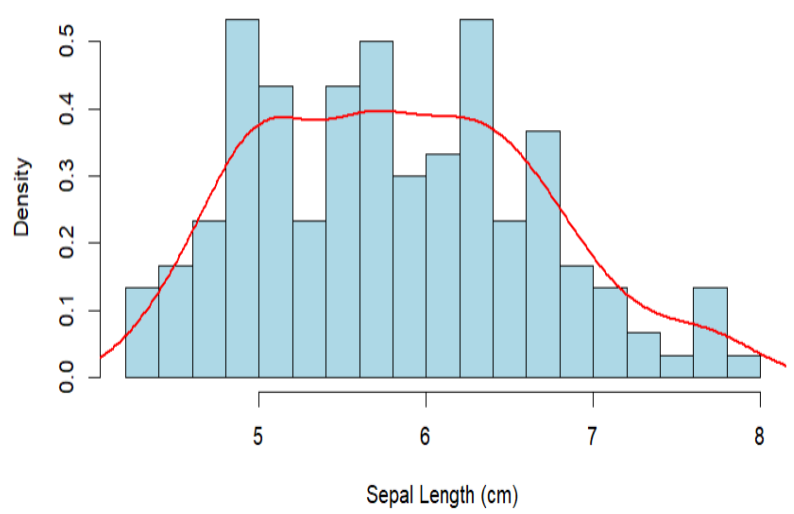



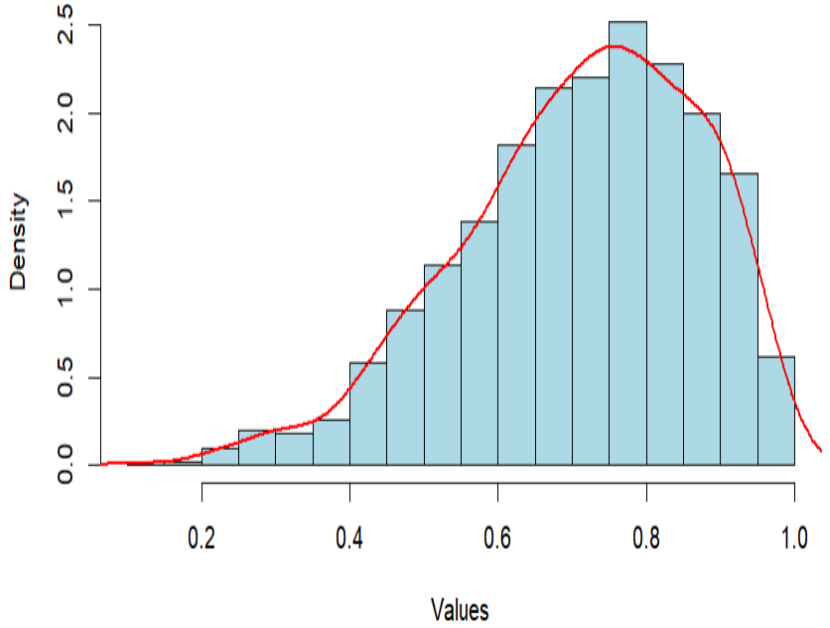



# SHAPE OF DISTRIBUTION: SKEWNESS

- What does the distribution of variable tell us?
  - The spread and central tendency of the distribution
  - Is the histogram right skewed, left skewed, or symmetric?

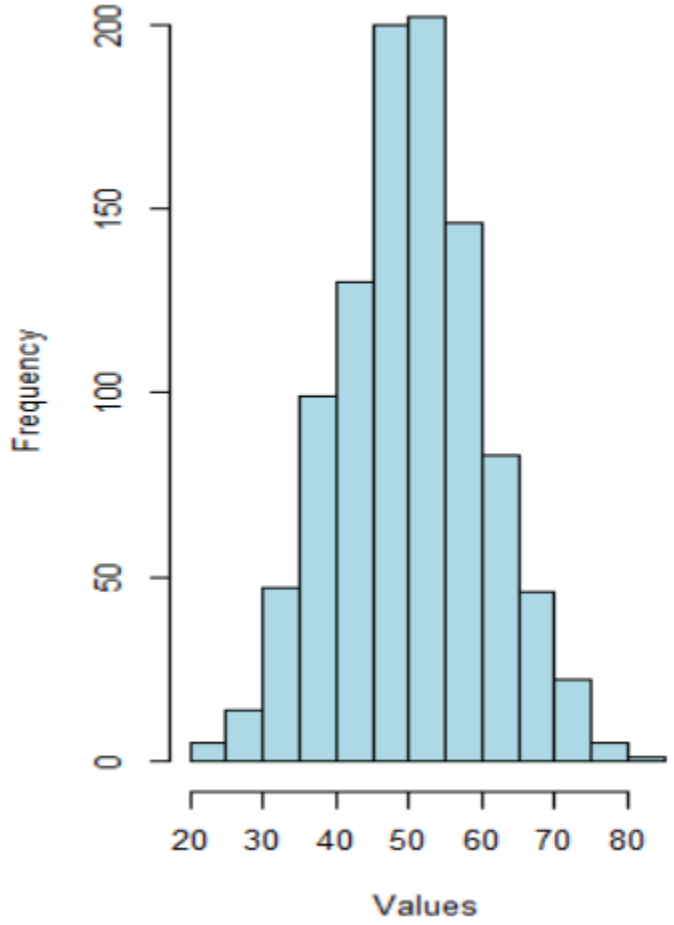


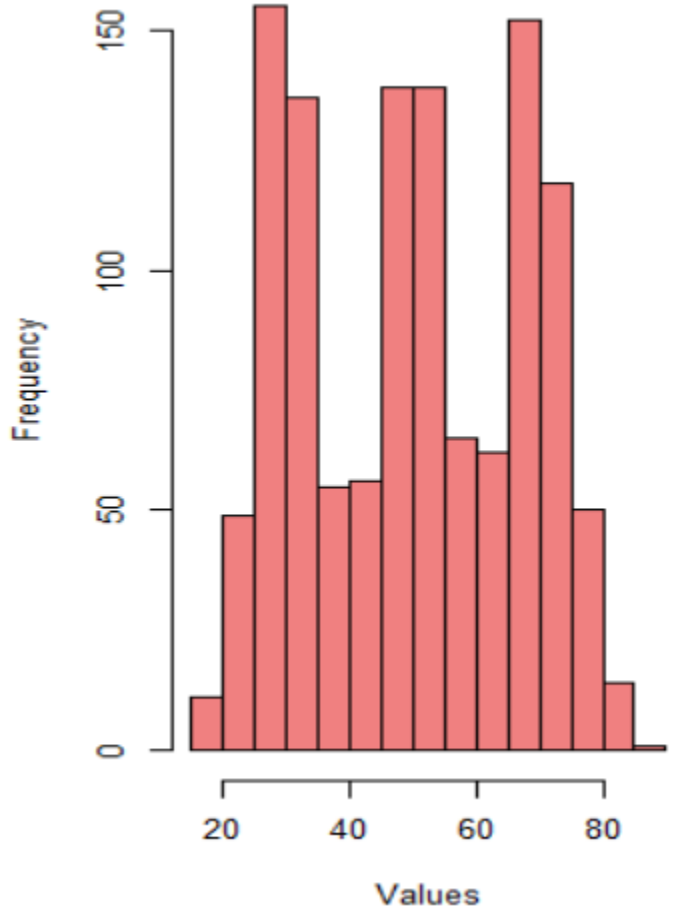



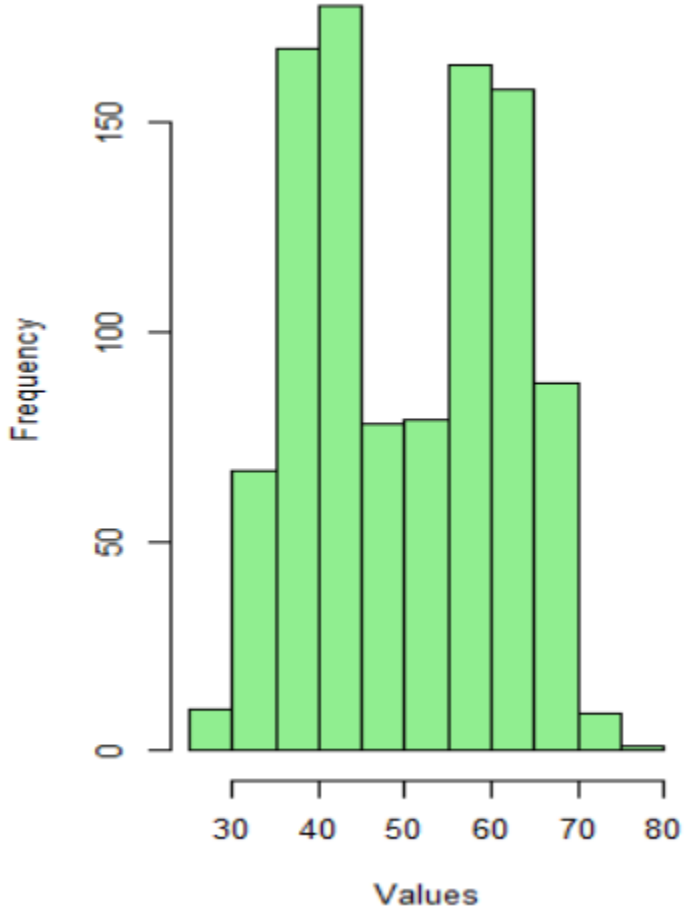



# SHAPE OF DISTRIBUTION: MODALITY

- Modality describe the number of meaningful cluster of observation

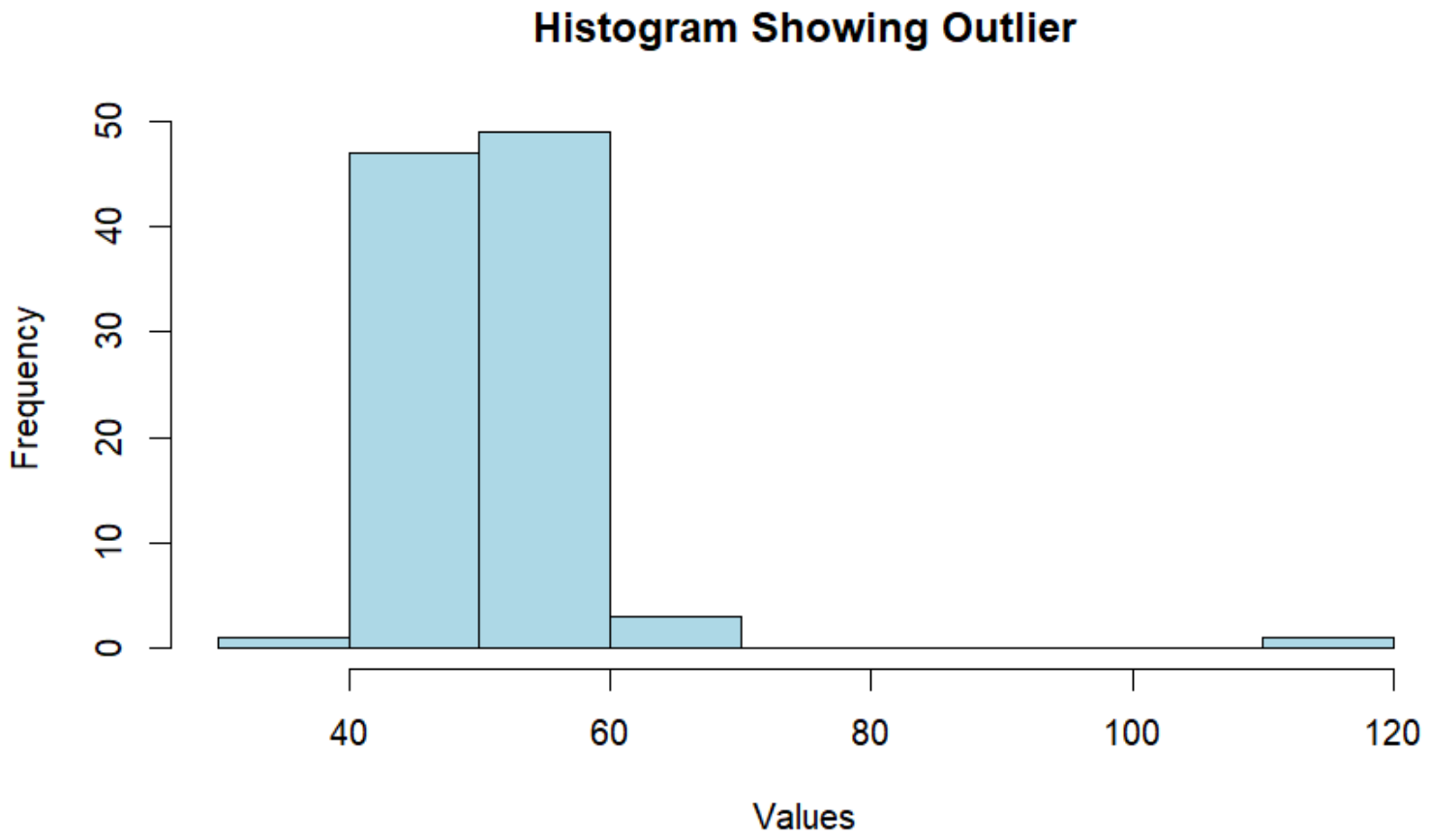


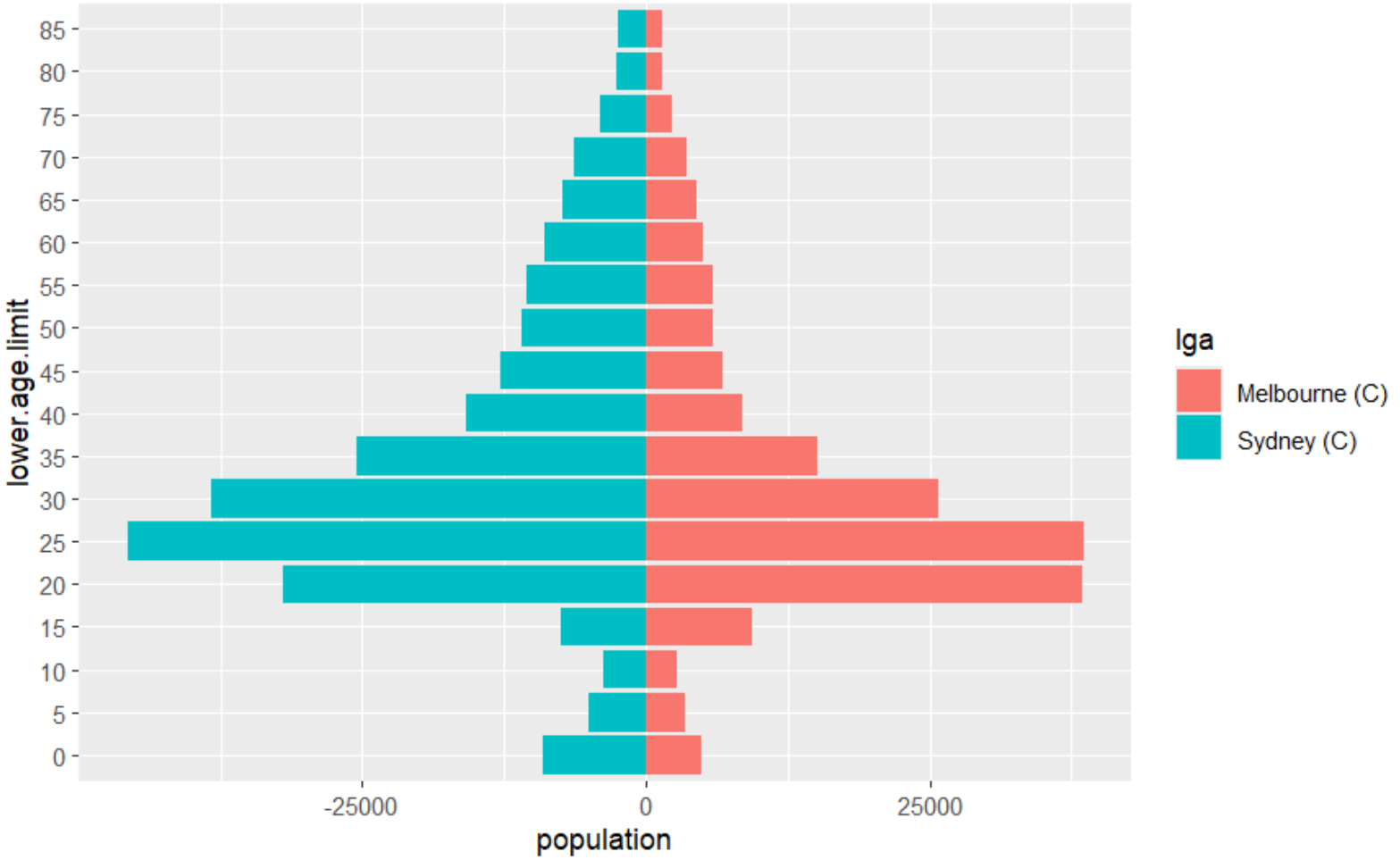
# SHAPE OF DISTRIBUTION: OUTLIER

- What does the distribution of variable tell us?
  - Detect outliers or unusual observations compared to the rest of the sample



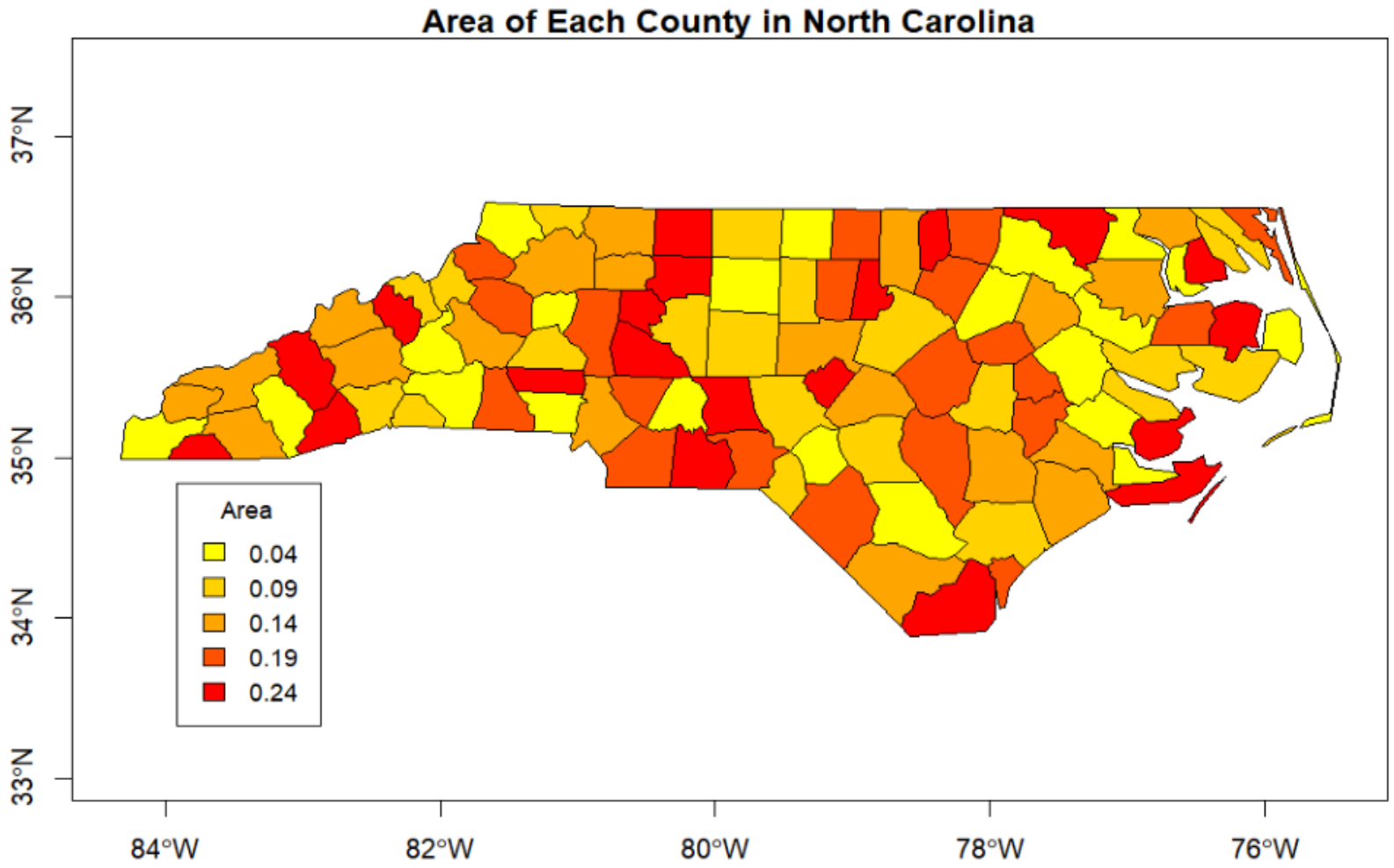
# SIDE-BY-SIDE HISTOGRAM

- A **side-by-side graph** is a comparative visualization technique used to display **two related data sets** next to each other, making it easier to **compare values across different categories or groups**



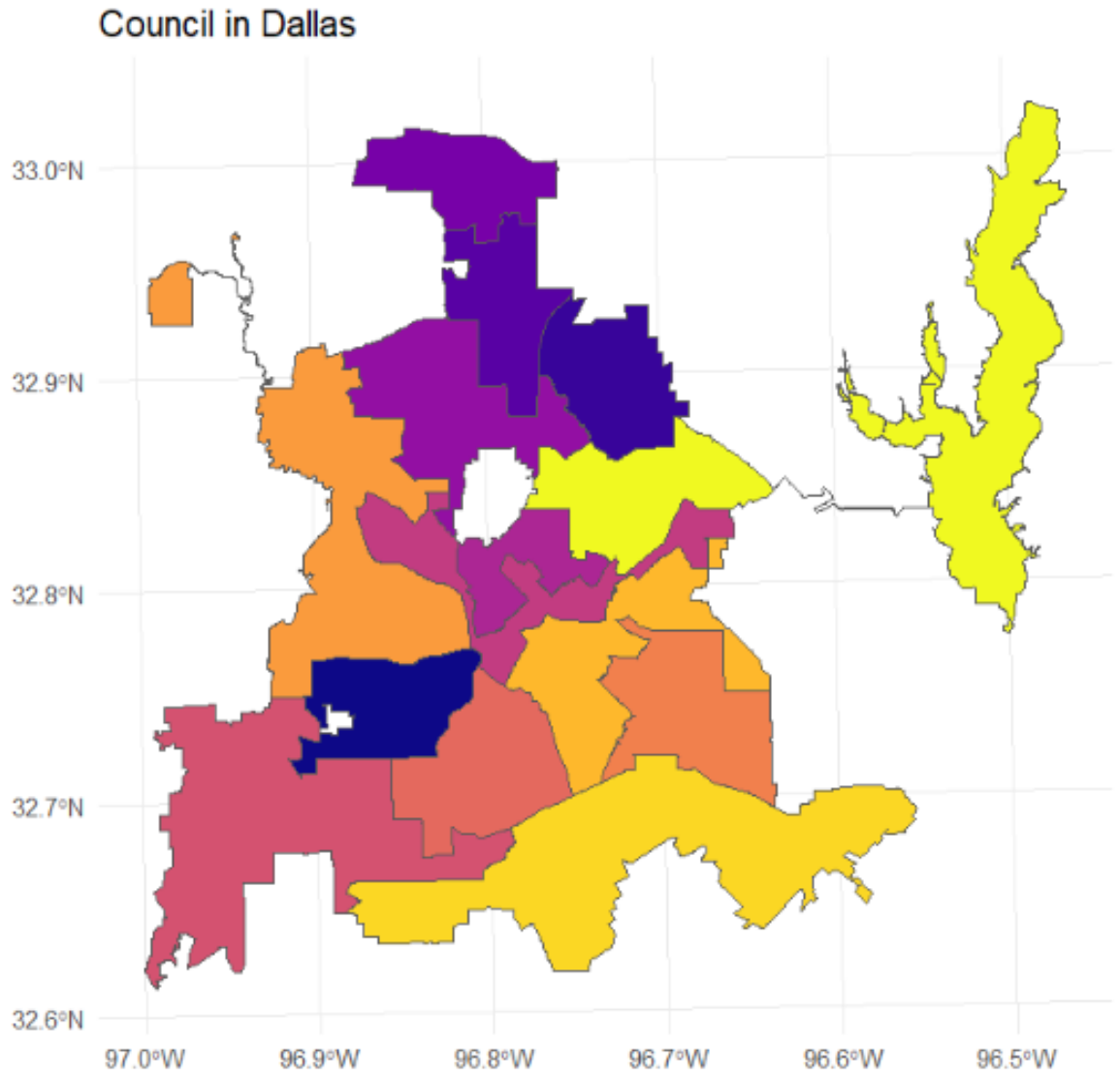
# SPATIAL DATA

- Gradient continuous map theme



# SPATIAL DATA

- Categorical map theme



# 2.1.2 UNIVARIATE VARIABLE DISTRIBUTIONS

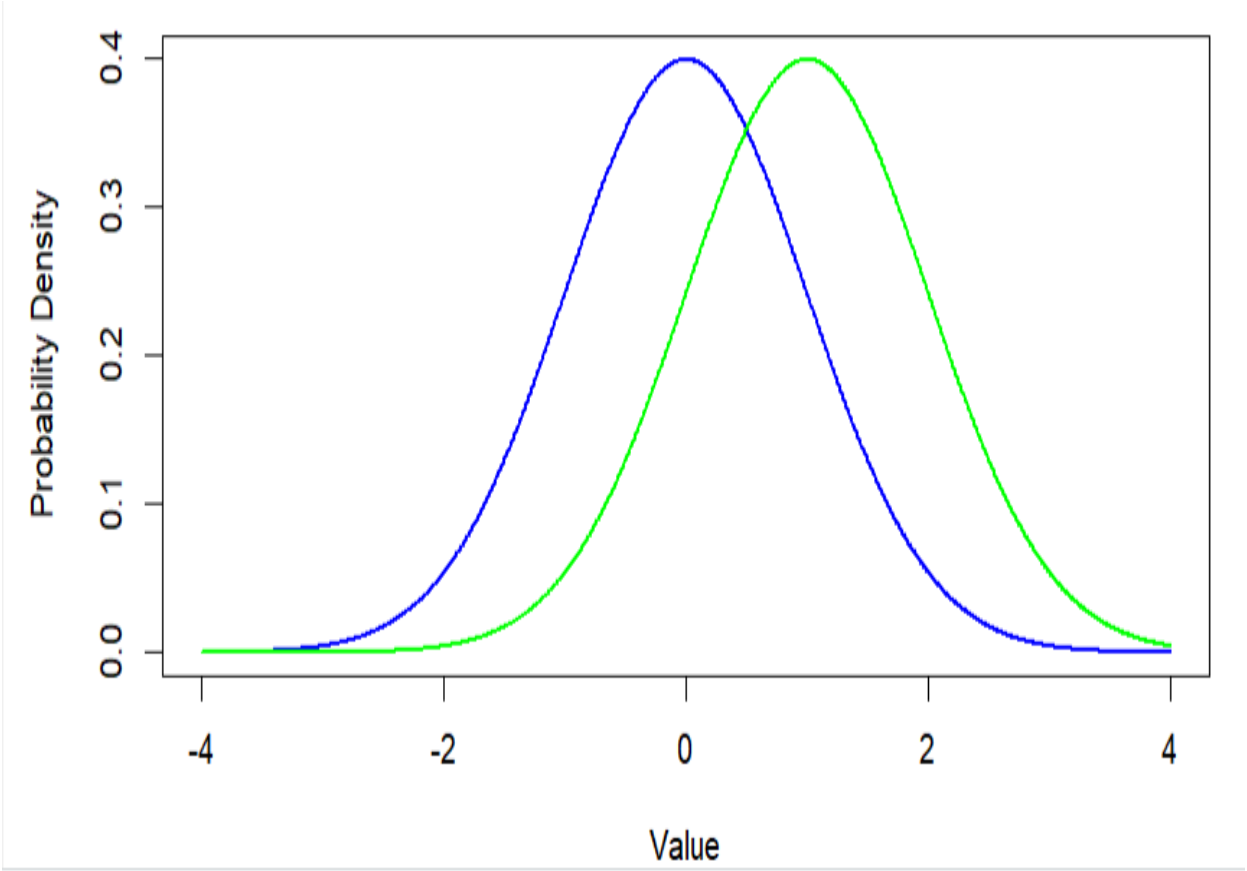
# DATA DESCRIPTION

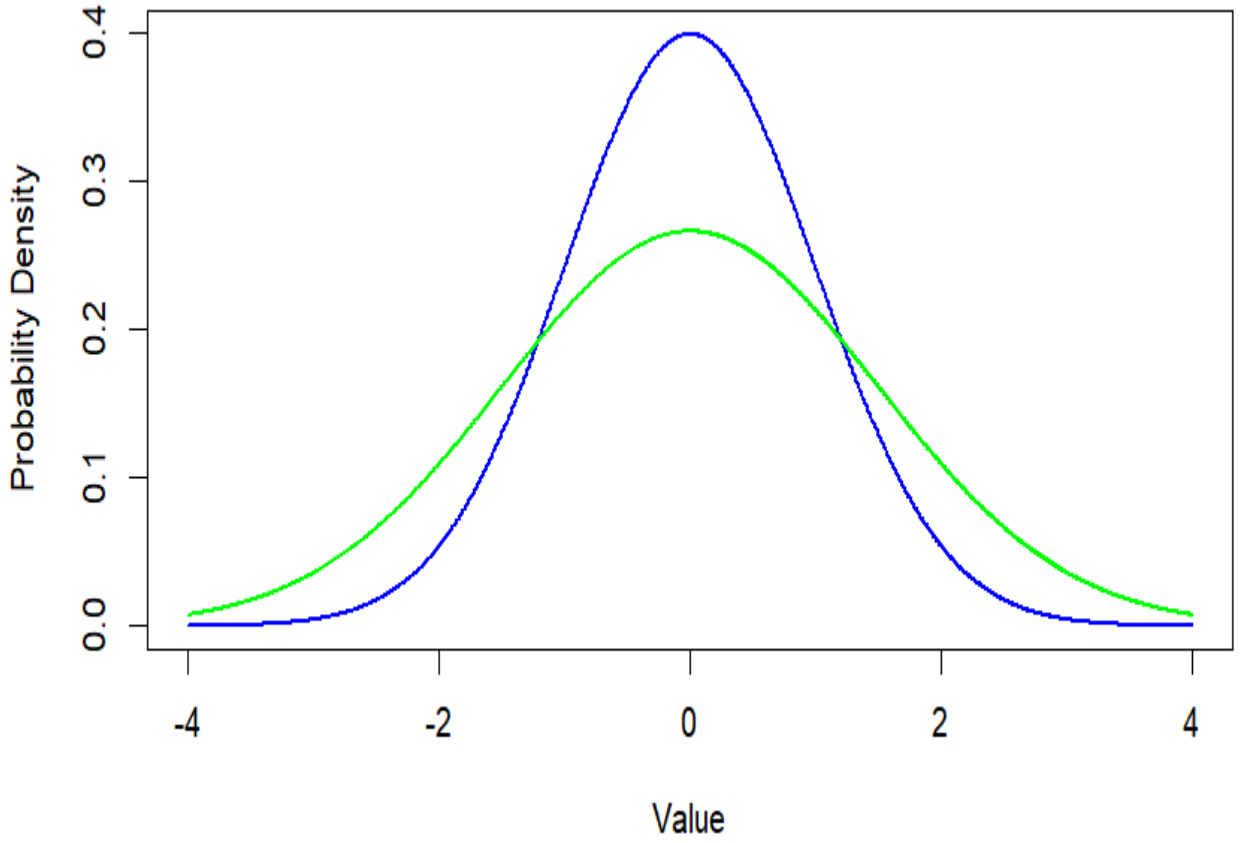
- Describing the distributions of a variable is a very import steps of any data analysis.
- The distribution or shape of a univariate distribution can have substantial impact on the outcome of statistical analysis.
- For example, skewness data may influence the outcome of statistical analysis and parameter estimations.
- Most methods assume the variables are symmetric and normally distributed.
- Why do normal distribution matter?
  - Parameter estimation: Most parametric models estimate parameters (mean, variance) based on the assumption that the data follows a normal distribution.
  - Hypothesis testing: Tests like t-tests and ANOVA require normal distribution to calculate valid p-values.
  - Inference: For methods like linear regression, normally distributed residuals are important for making accurate predictions.

# NORMALITY

- The normal distribution is a bell-shaped curve that represents a continuous probability distribution with the highest density of observations clustered around the mean.
- As you move farther from the mean in either direction, the frequency of observations gradually decreases, resulting in symmetry about the center.
- The distribution is fully characterized by two parameters: the mean (μ), which determines the central location of the curve, and the standard deviation (σ), which controls the spread or dispersion of the data.

- Which graph shows curves with the same mean but different standard deviations?
- Which graph shows curves with the same standard deviation but different means? 😕

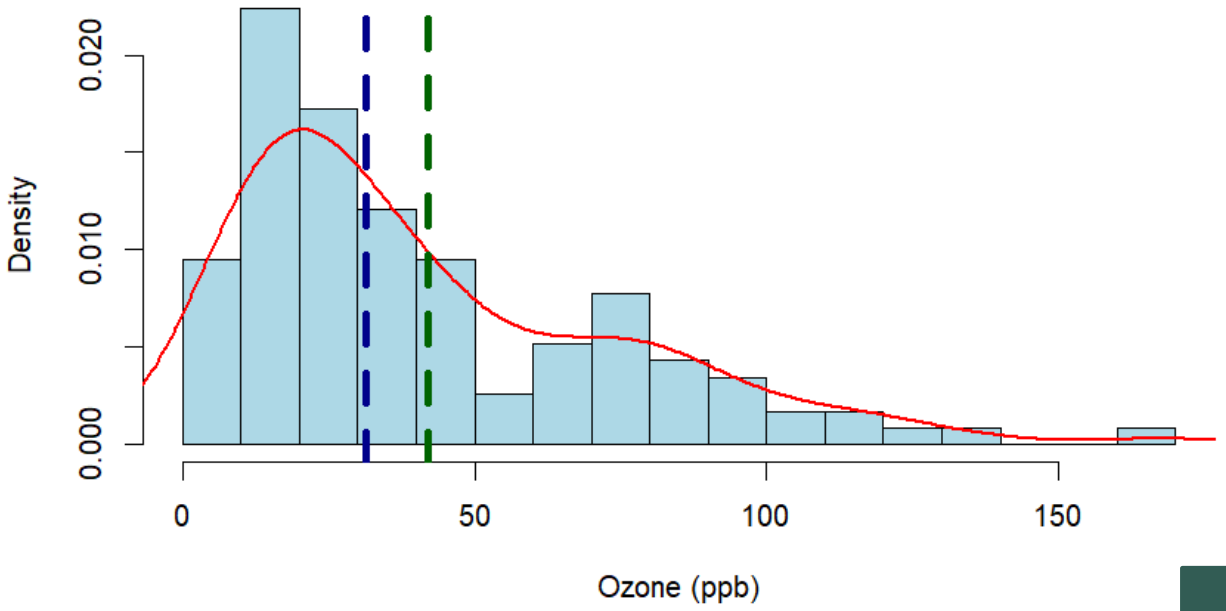


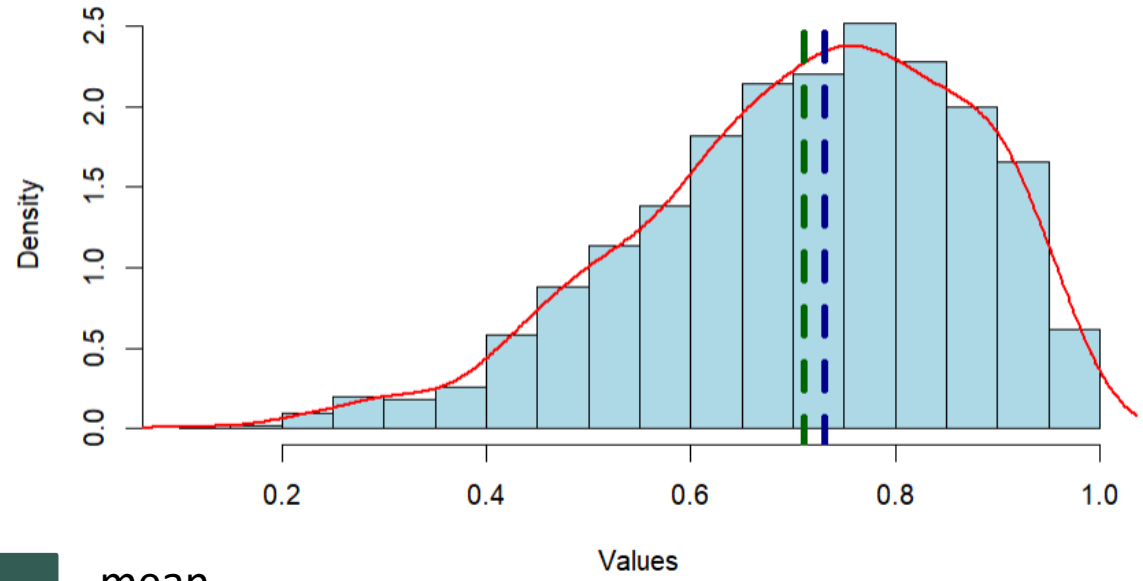



# NON-NORMALITY

- The concept of the **mean** as a measure of **central tendency** becomes **less reliable** due to the disproportionate influence of **extreme values** (outliers)d distribution

- Consequently, the **median**, being less sensitive to extreme values, often provides a **more robust and meaningful representation** of the **center of the distribution**



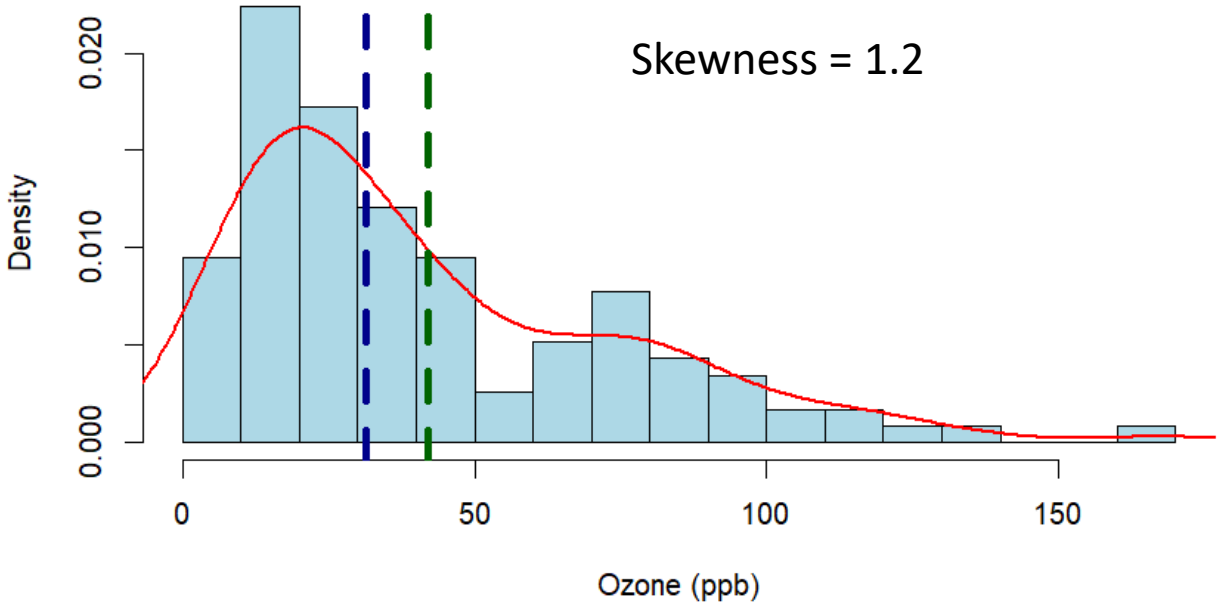



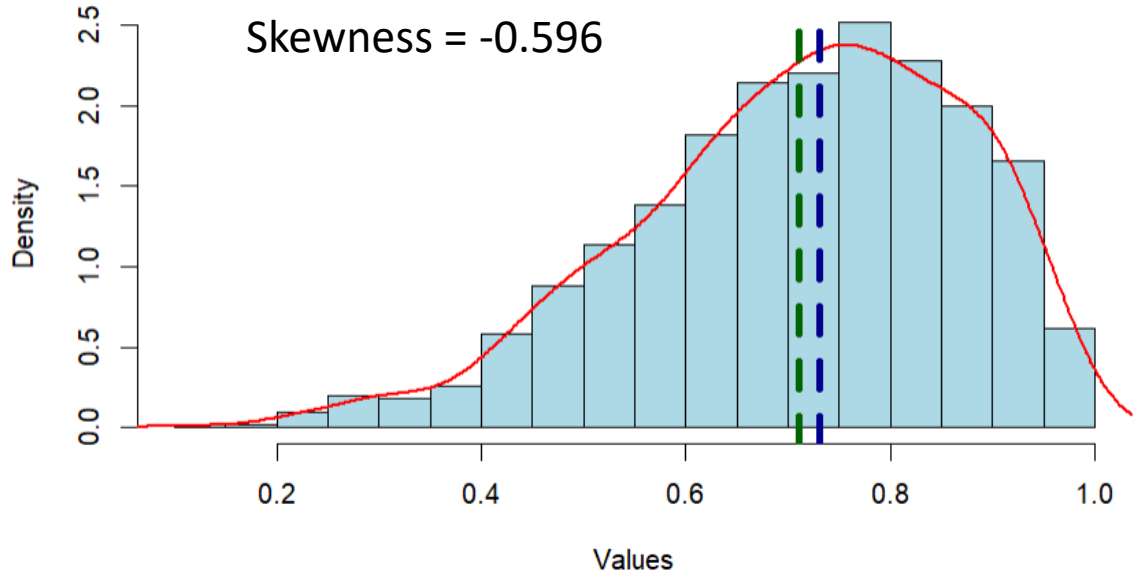

mean

median

# SKEWNESS

- The skewness is defined as by :
  - $\text{skewness(X)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n * s_X^3}$




What is the skewness for a normal distribution? 🤔

# OUTLIER IN UNIVARIATE

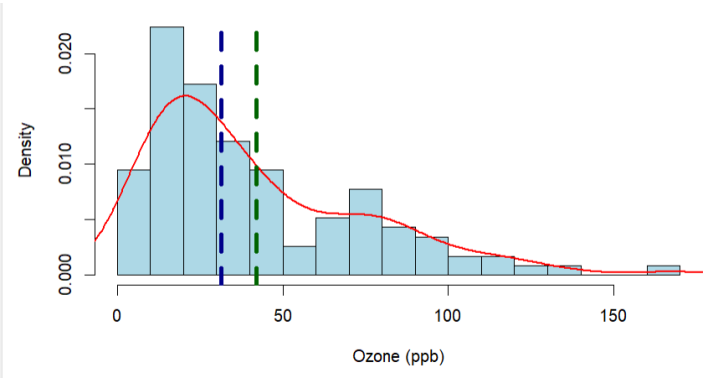
- Outliers are observations with a unique combination of characteristics indicating they are distinctly different from the other observations.

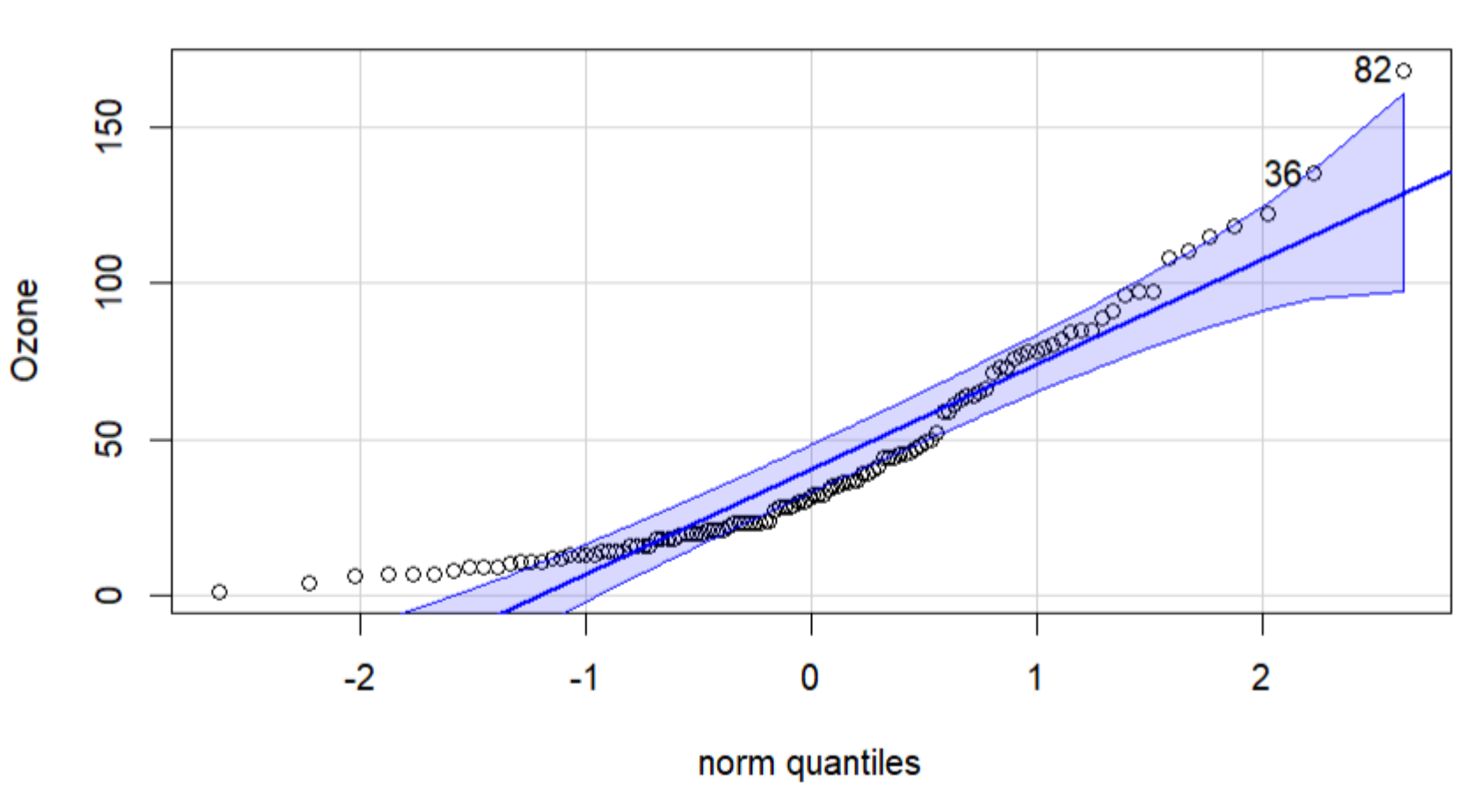
- These differences can be on a single variable (univariate outlier), a relationship between two variables (bivariate outlier), or across an entire set of variables (multivariate outlier).

- The univariate identification of outliers examines the distribution of observations for each variable in the analysis and selects as outliers those cases falling at the outer ranges (high or low) of the distribution.

# QUANTILE-QUANTILE(QQ)-PLOT

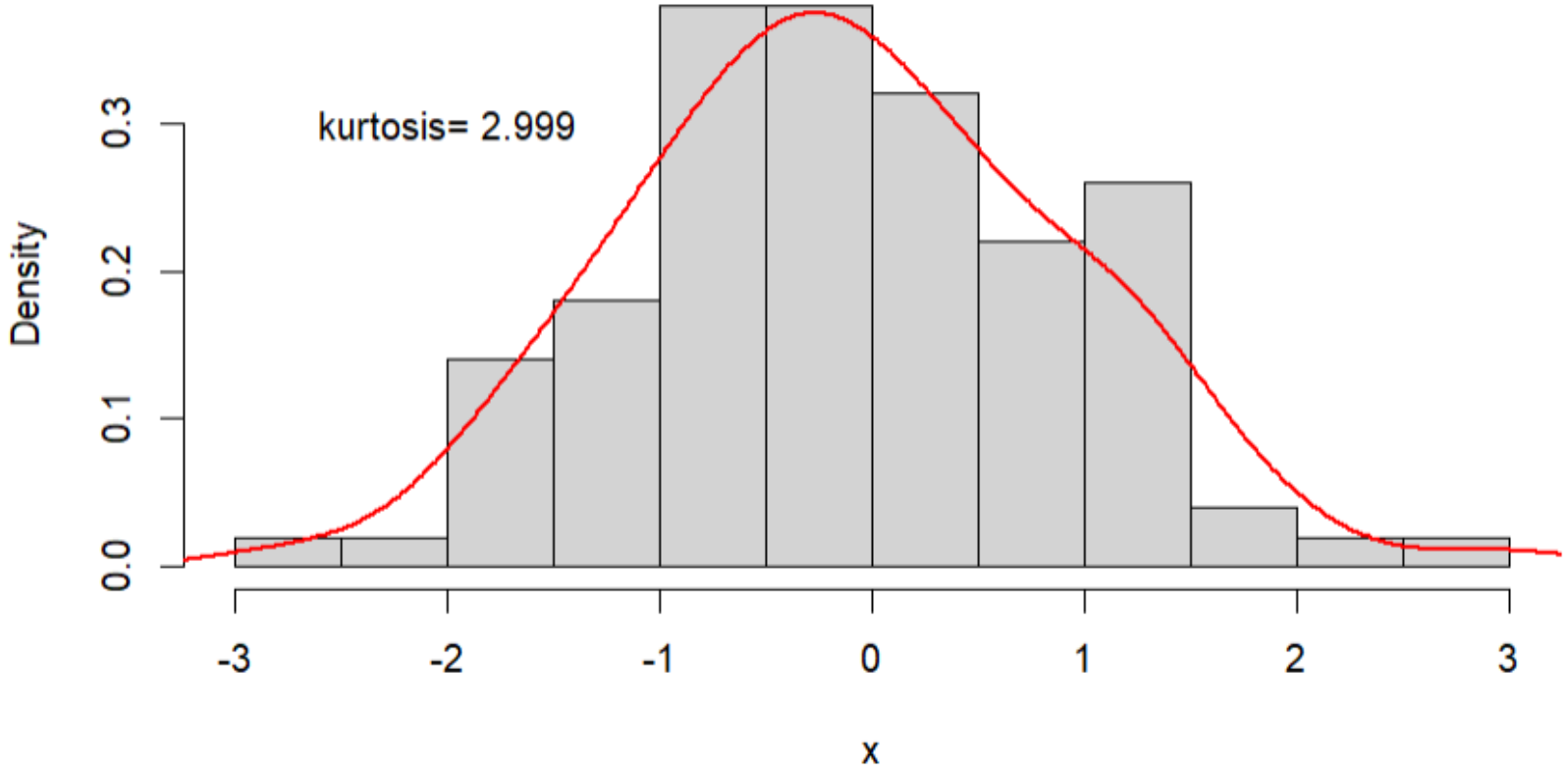
- A **QQ plot (Quantile-Quantile plot)** is a graphical tool used to **compare the distribution of a dataset** to a **theoretical distribution** (usually a normal distribution). It helps to **identify deviations** from normality and **spot potential outliers**.



# KURTOSIS

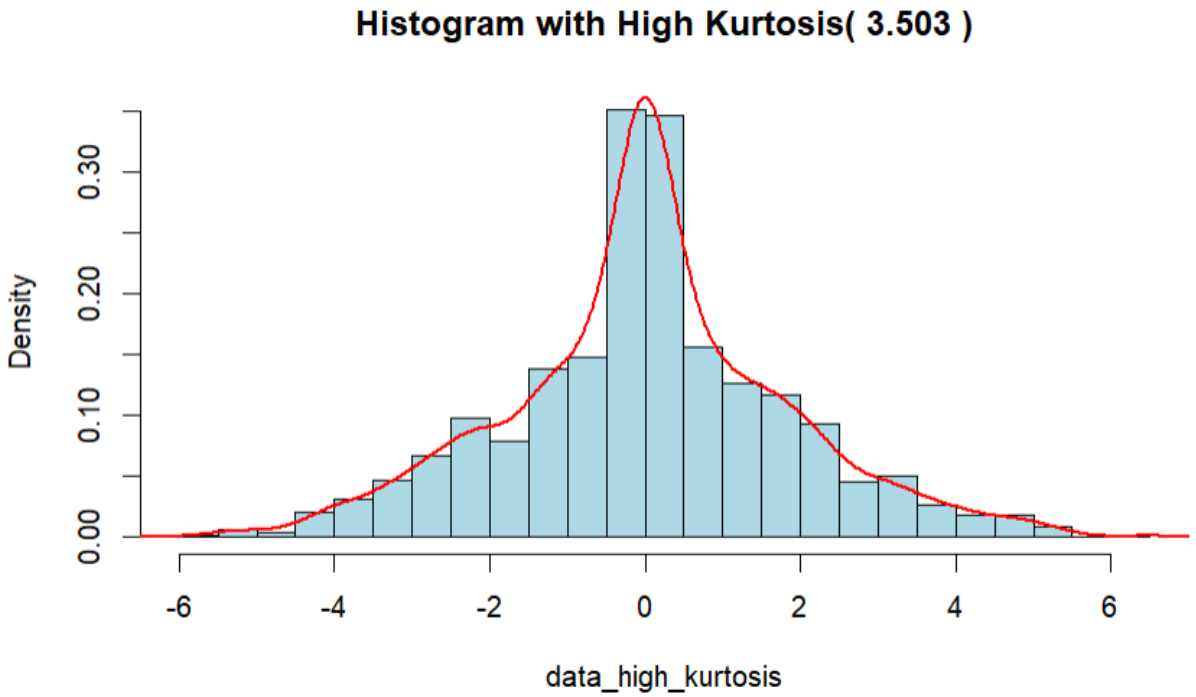
- Kurtosis is a measure of the “tailedness” of the probability distribution. A standard normal distribution has kurtosis of 3 and is recognized as mesokurtic [1].
- Kurtosis is a measure of whether or not a distribution is heavy-tailed or light-tailed relative to a normal distribution.

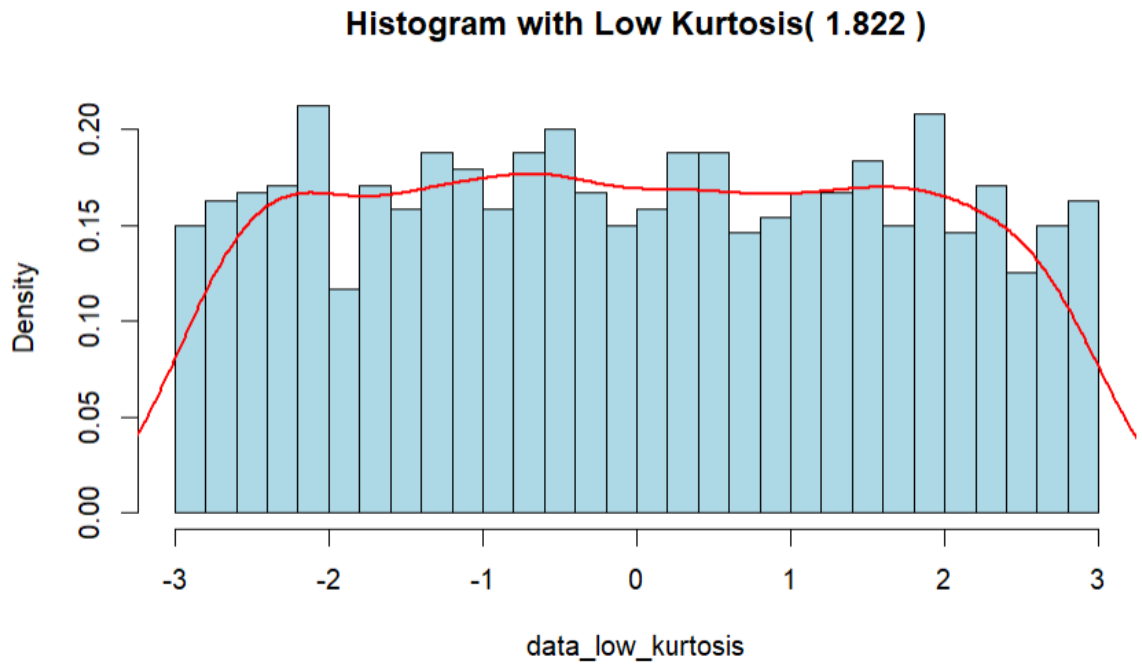



[1] LABORATORY STATISTICS, HANDBOOK OF FORMULAS AND TERMS. CHEMISTRY INTERNATIONAL, 36(4), 23-23.

# LEPTOKURTIC & PLATKURTIC

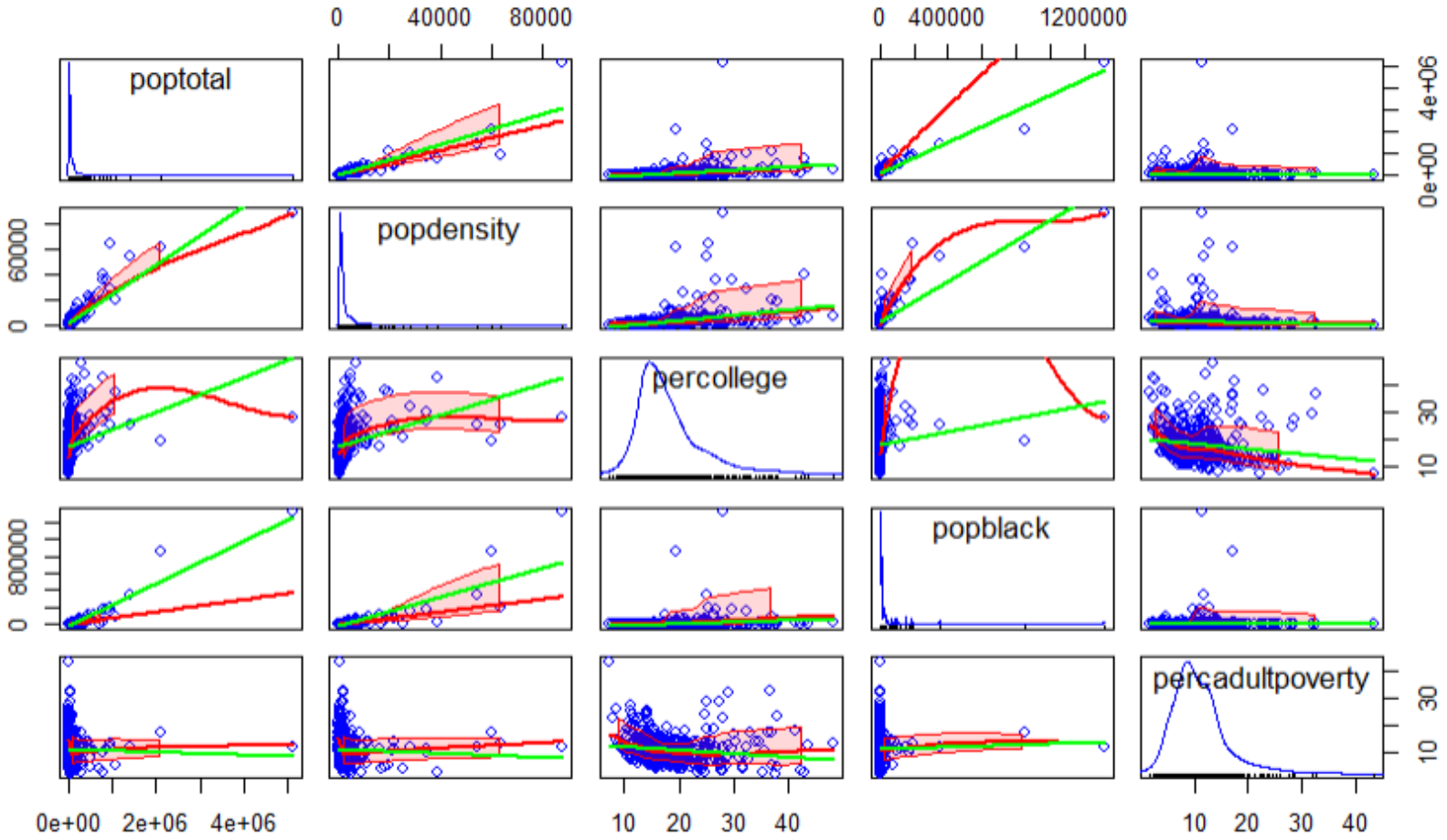
- An increased kurtosis (>3) can be visualized as a thin “bell” with a high peak whereas a decreased kurtosis corresponds to a broadening of the peak and “thickening” of the tails.
- Kurtosis > 3 is recognized as leptokurtic and < 3 as platykurtic.



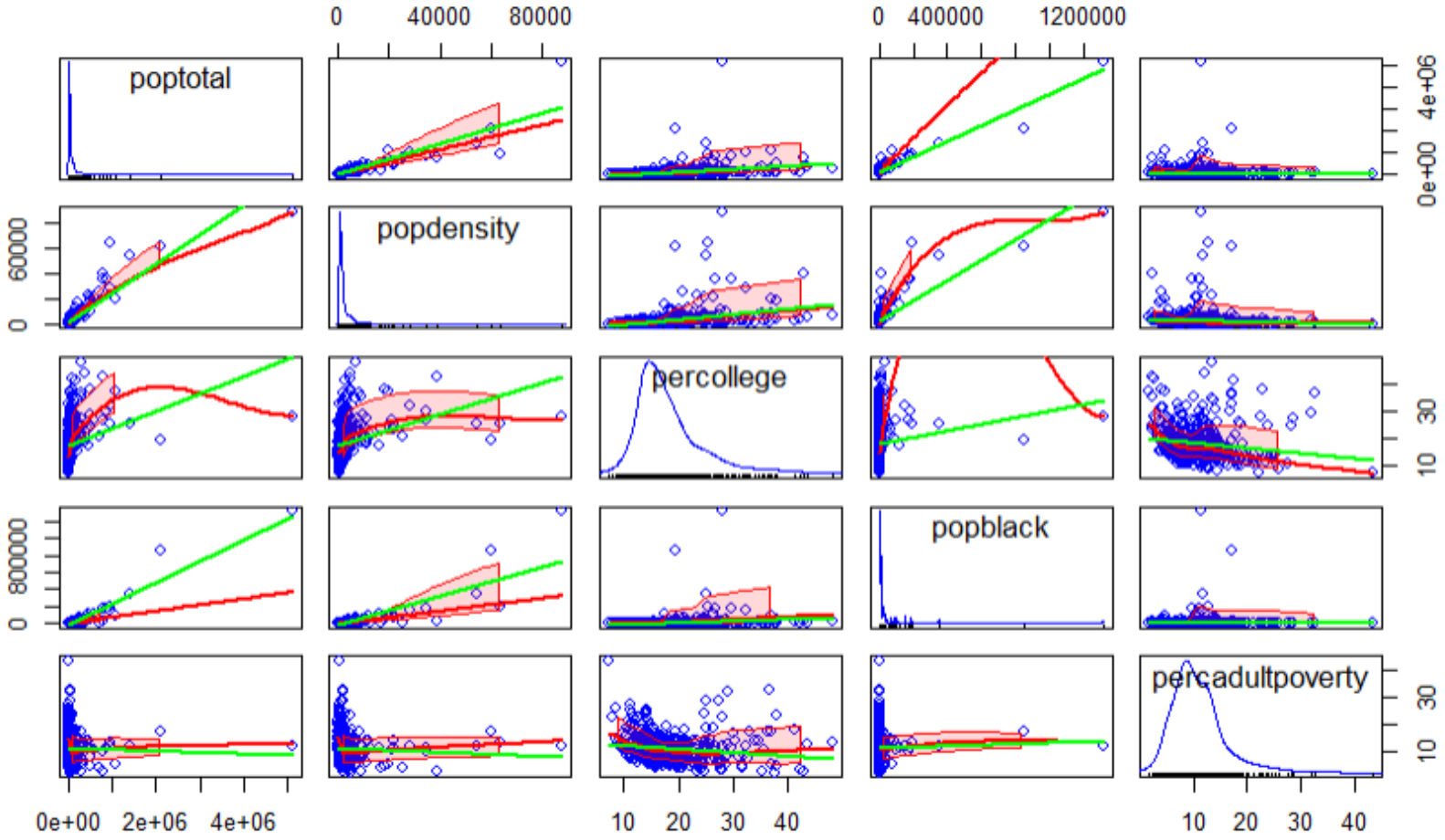
# SCATTERPLOT MATRIX

- A scatterplot matrix is a matrix of scatterplots that lets you understand the pairwise relationship between different variables in a dataset



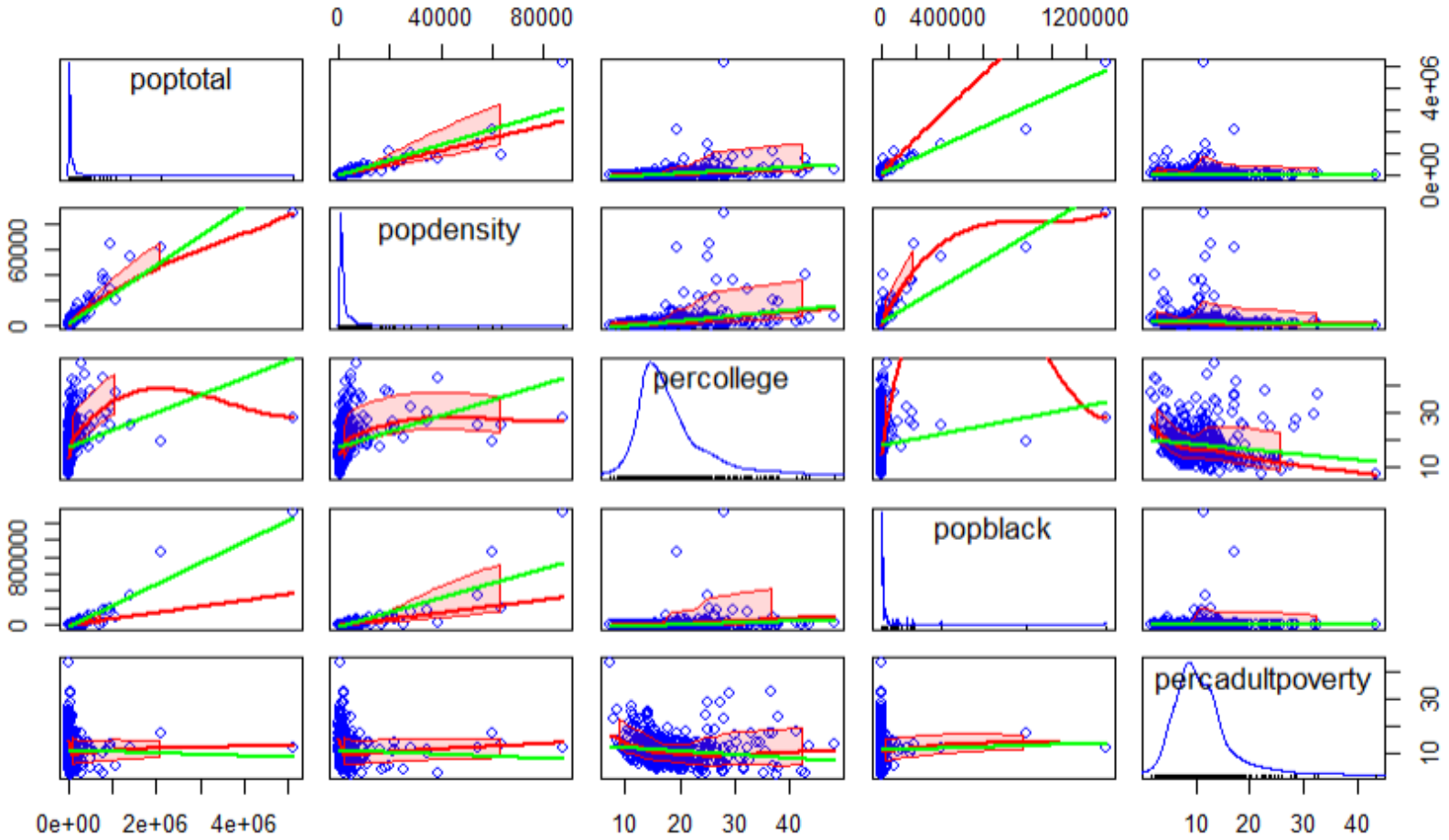
# A GRID OF SCATTERPLOT MATRIX

- Rows and column represent each variable.
- Each cell in the matrix shows a scatterplot of two variables (one on the x-axis, one on the y-axis).
- The diagonal cells typically contain the variable names or histograms of each variable.



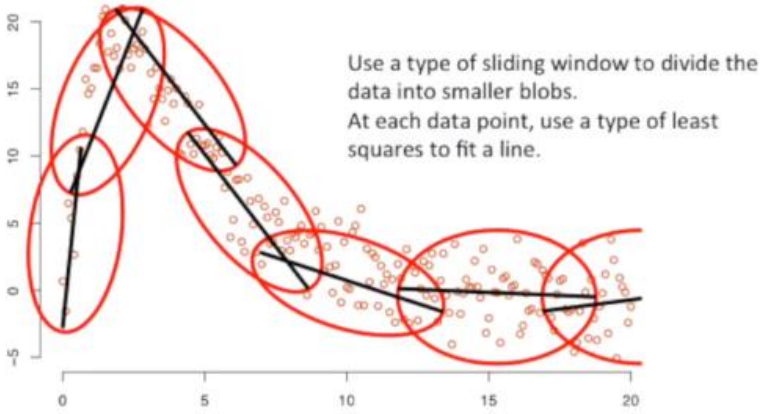
# REGRESSION LINE

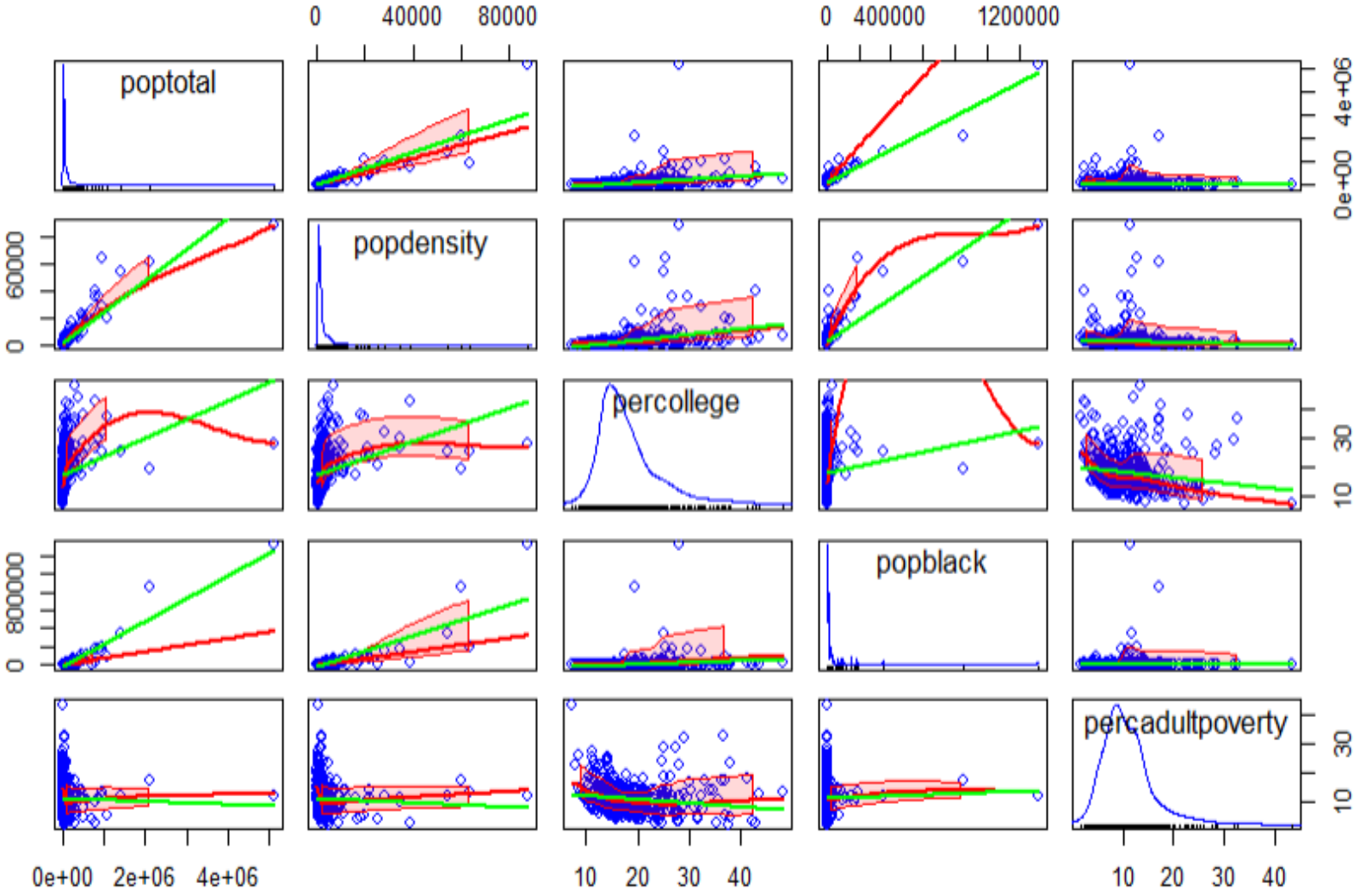
- The **green regression line** shows the **linear trend** (ordinary least squares).



# LOESS SMOOTH LINE

- The **red smooth line** shows a **non-linear trend** (LOESS curve) in the data.

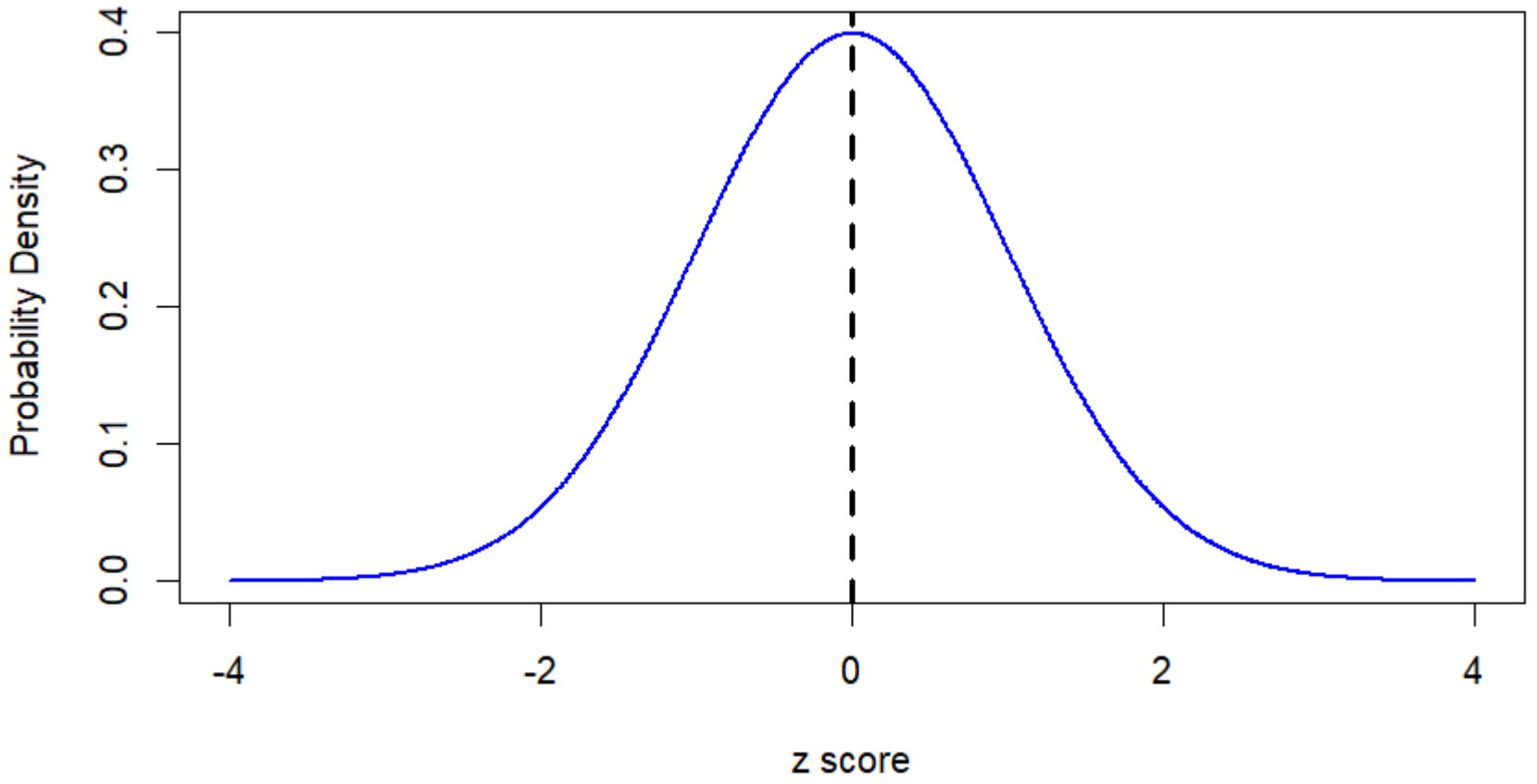


# 2.1.3 DATA TRANSFORMATION

# DATA TRANSFORMATION - STANDARDIZATION

- **Standardization** actually takes many forms, but the most commonly used is the z score or standard score.
- Methods that require data standardizing, for example, Support Vector Machine, Principal Component Analysis



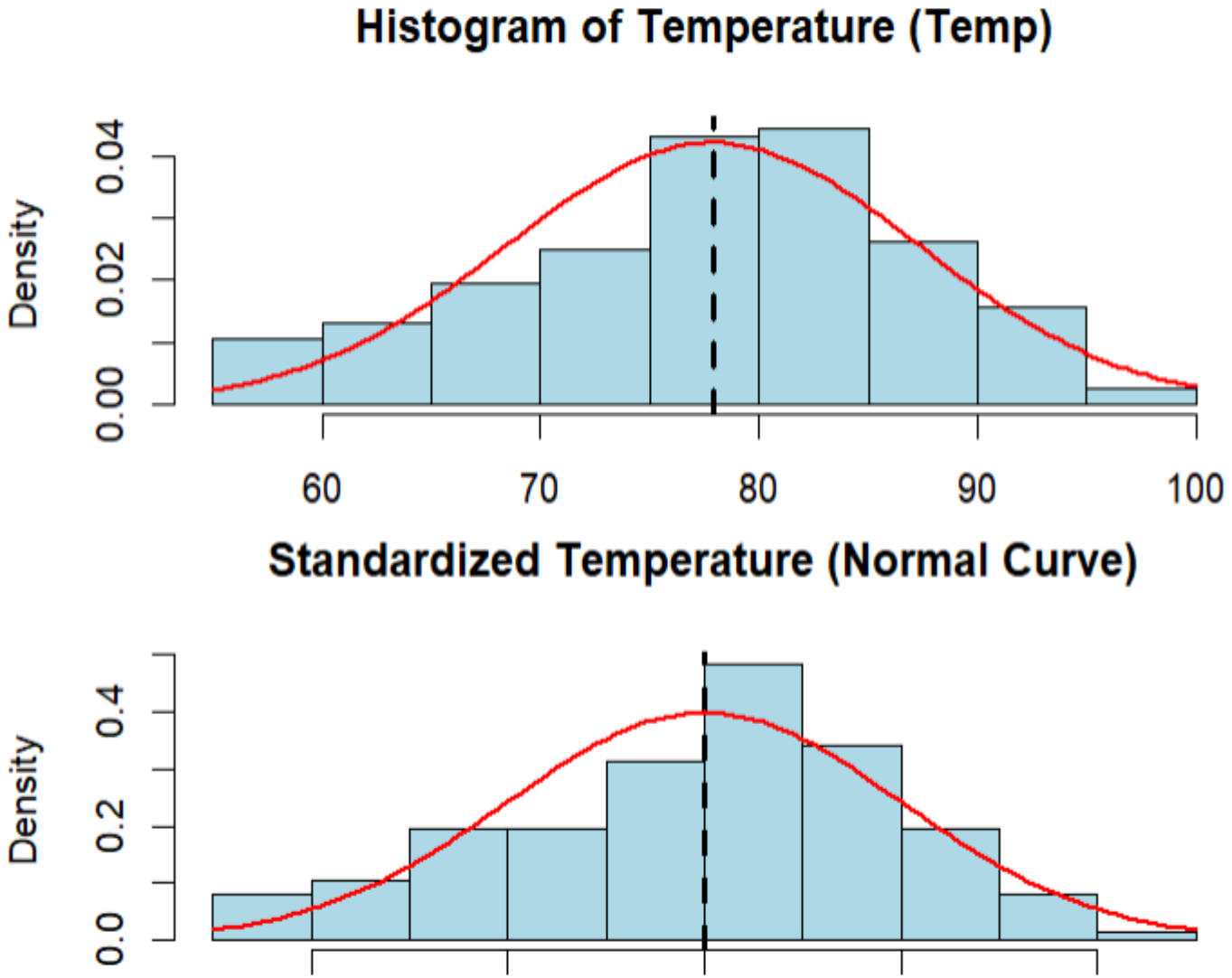

$z = N(0,1)$

# Z-SCORE

- Values above or below zero indicate the observation's difference from the variable mean in terms of standard deviations.

In a normal distribution, one observation has a z-score of 0.5 and another observation has a z-score of -1.2.

Which observation is farther from the mean? 🤔



# NORMALIZATION

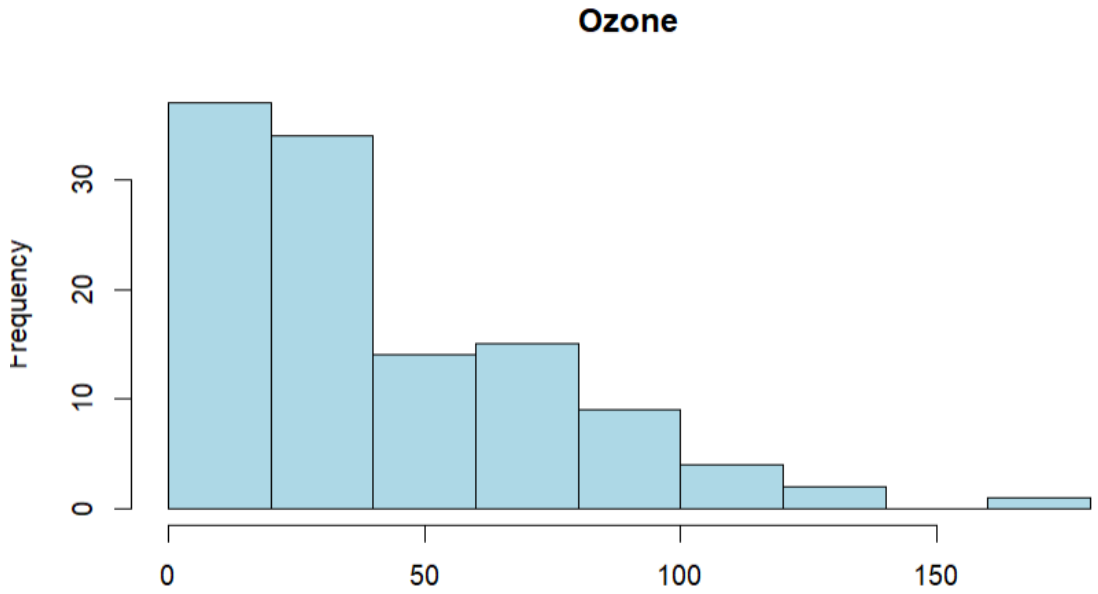
- **Normalization** is a technique used to **scale** the values of numerical features to ensure that all features contribute equally to a machine learning model. It helps improve **model performance, stability, and training speed**, especially for algorithms sensitive to the **magnitude and scale of input features**.

Why normalize data?

- **Different Scales Cause Bias**
  In a dataset, different features may have **different ranges**. For example:
  - **Age** might range from **0 to 100**.
  - **Income** might range from **$30,000 to $100,000**.
- Without normalization, machine learning models may **assign more importance to larger values**, leading to **biased results**.

# NORMALIZATION

- It is defined as:
  - $x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}}$

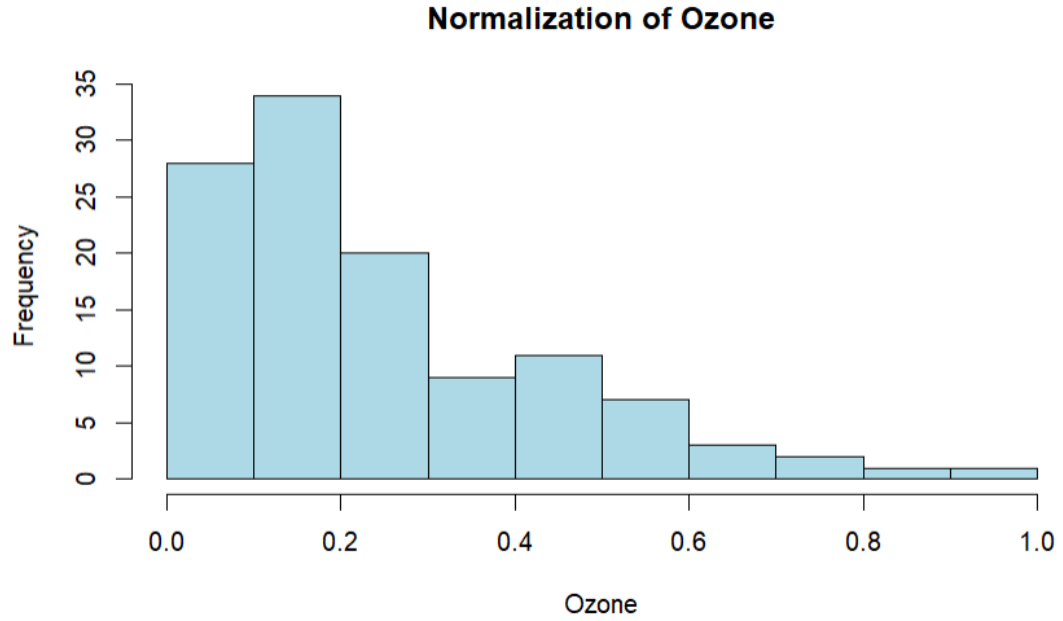



What properties of data change after normalization, and what properties remain unchanged? 🤔

a) **Skewness**

b) **Scale (range)**

c) **Kurtosis**



# WEEK 03

# LAB SESSION

Instructor: Yanan Wu
TA: Khadija Nisar

Spring 2025