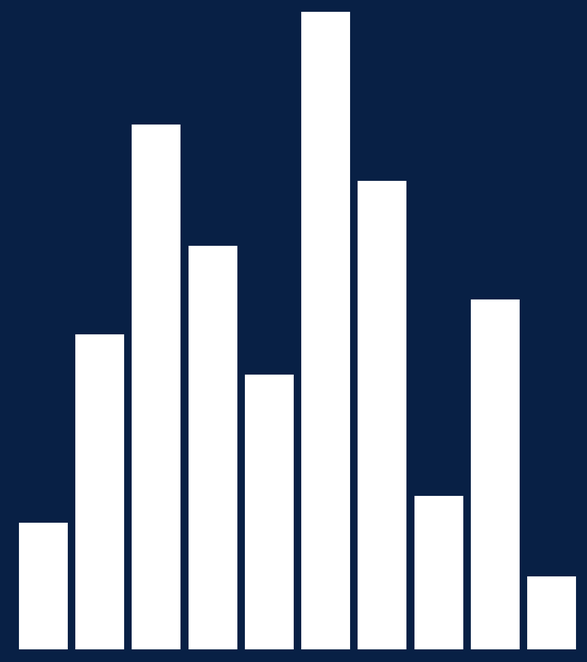


Uncover the Hidden Secrets of Your Data with Histograms

Are you interpreting histograms correctly?
Don’t let misreading them cost you crucial insights.

“A picture is worth a thousand words, but only if you know how to read it.”

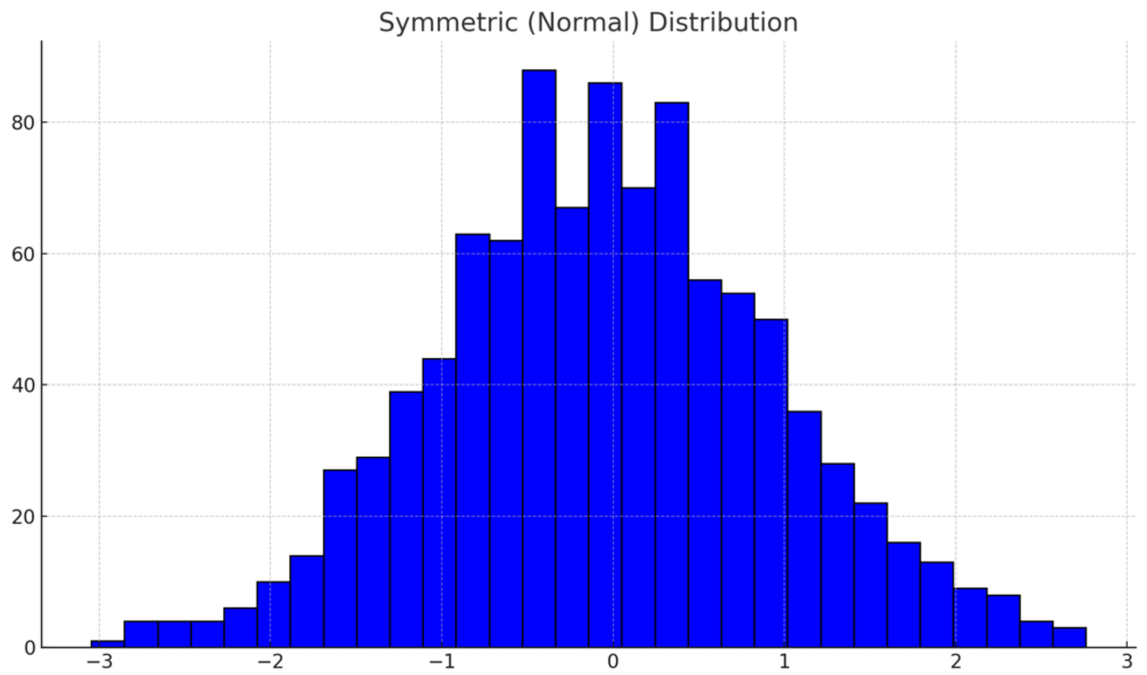
Histograms can reveal deep insights about your data.

But only if you know how to interpret them correctly.

Let’s dive into how you can read histograms like a pro. Your data decisions depend on it.

Addressing these characteristics in a histogram involves careful analysis and often requires a combination of data transformations, segmentation, and robust statistical methods.

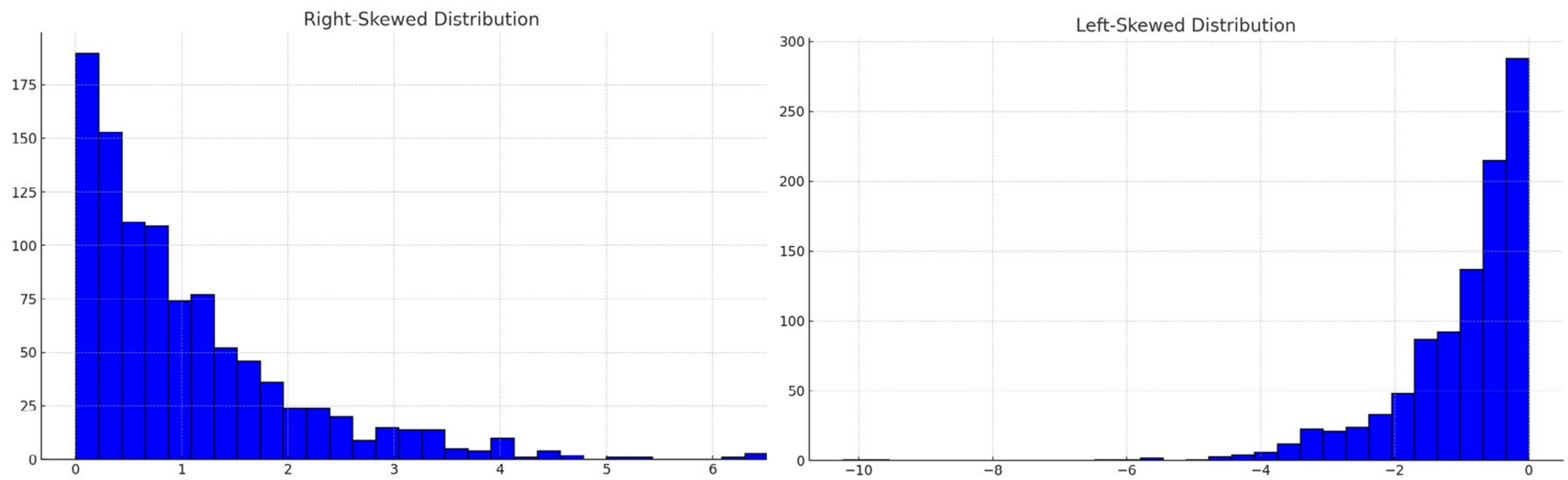
The key is understanding what the histogram reveals and then tailoring the approach to correct, enhance, or make sense of the distribution for more accurate interpretation and decision-making.



Symmetry vs. Skewness

Symmetrical Distribution:

The histogram has a roughly bell-shaped curve, with data evenly distributed on both sides of the central peak. This often indicates a normal distribution.



Skewed Distribution:

Right-Skewed (Positively Skewed):
The tail on the right side of the histogram is longer, indicating that a few values are much larger than the rest. The mean is usually more significant than the median in this case.

Left-Skewed (Negatively Skewed):
The tail on the left side is longer, suggesting a few much smaller values. The mean is typically less than the median.

Managing Skewed Distributions:

Right-Skewed Distribution:

Transformation: Apply a logarithmic, square root, or reciprocal transformation to reduce skewness and make the distribution more symmetrical.

Alternative Measures: Consider using the median instead of the mean to summarise the central tendency, as the median is less affected by skewness.

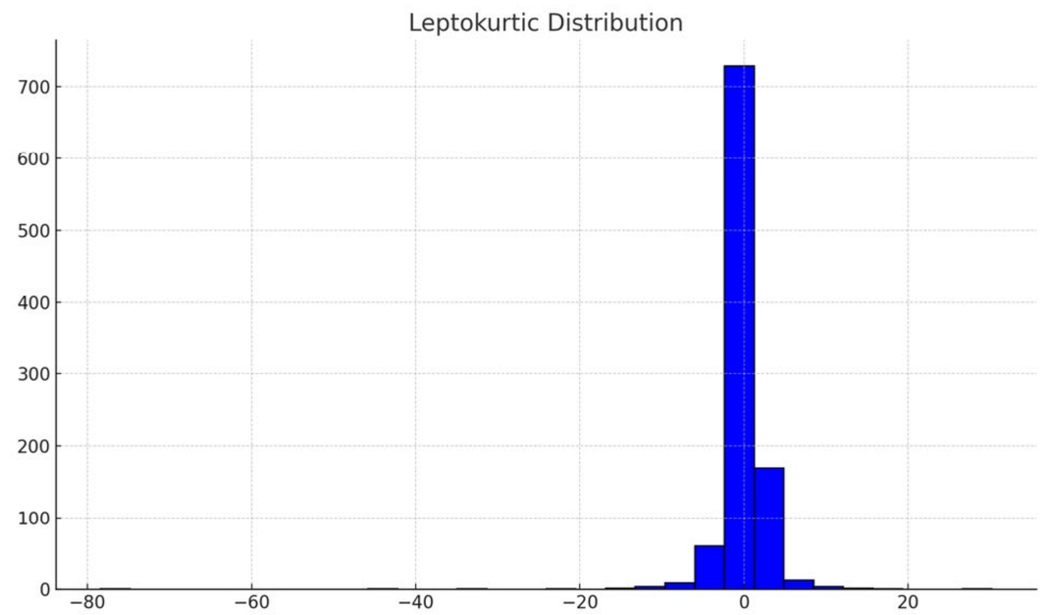
Segmentation: Analyse the skewed portion separately if it represents a distinct group or phenomenon.

Left-Skewed Distribution:

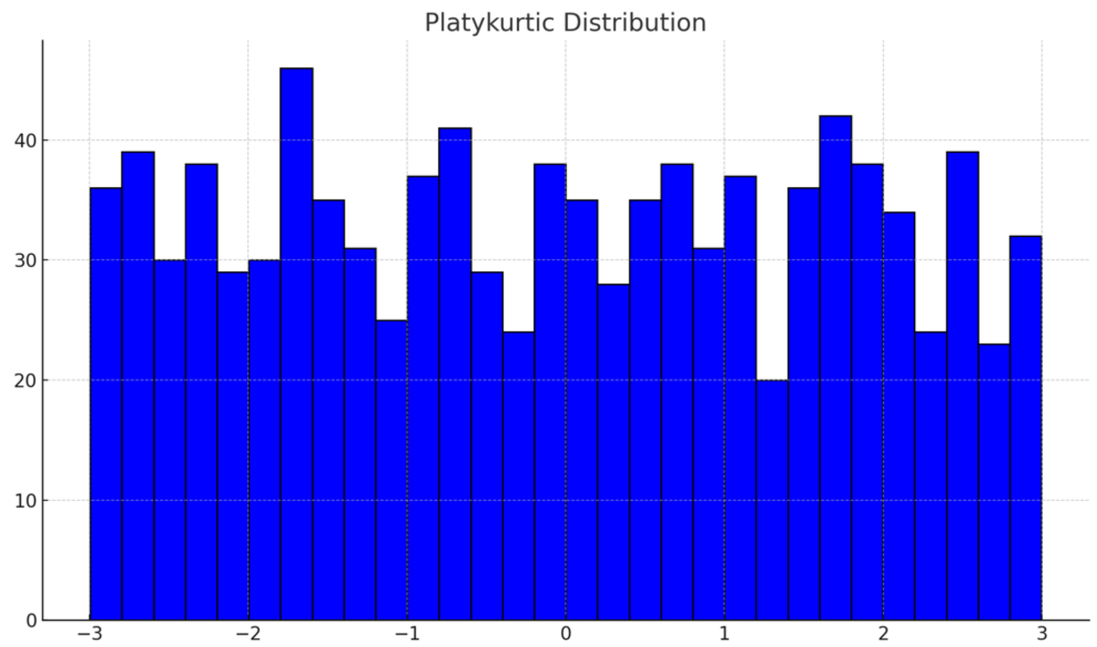
Transformation: Similar to right skew, but apply transformations in the opposite direction (e.g., squaring the data).

Alternative Summaries: Use the median and interquartile range (IQR) instead of mean and standard deviation.

Kurtosis



Leptokurtic:
A histogram with sharp peaks and fat tails indicates that data has heavy tails or outliers. This means more values are near or far from the mean, with fewer in the intermediate range.



Platykurtic:
A flatter distribution with thinner tails, indicating data is more evenly spread out with fewer extreme values.

Managing Kurtosis

Leptokurtic (High Kurtosis):

Outlier Management: Identify and potentially remove or cap outliers contributing to heavy tails. Use robust statistical methods that are less sensitive to outliers.

Alternative Distributions: Consider modelling the data with distributions that naturally accommodate heavy tails, like the t-distribution.

Platykurtic (Low Kurtosis):

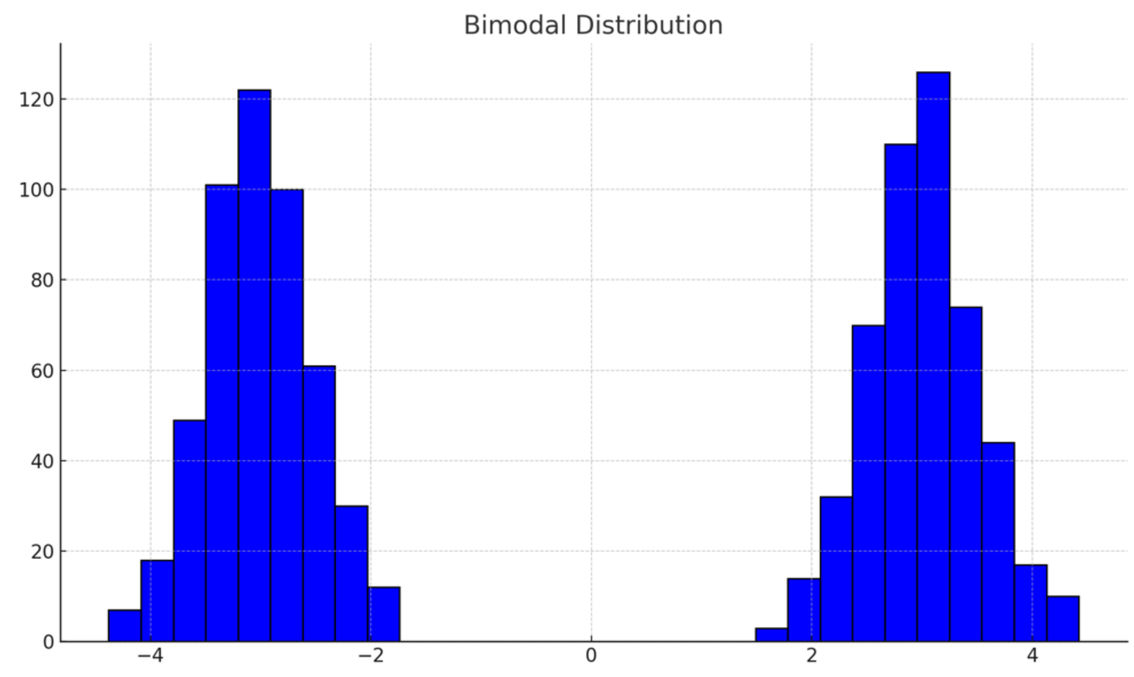
Summary Statistics: If the data is too flat, consider using other measures like range or variance to capture the spread more effectively.

Data Collection: Ensure the flatness isn't due to errors or insufficient data collection.

Peaks (Modes)

Unimodal:

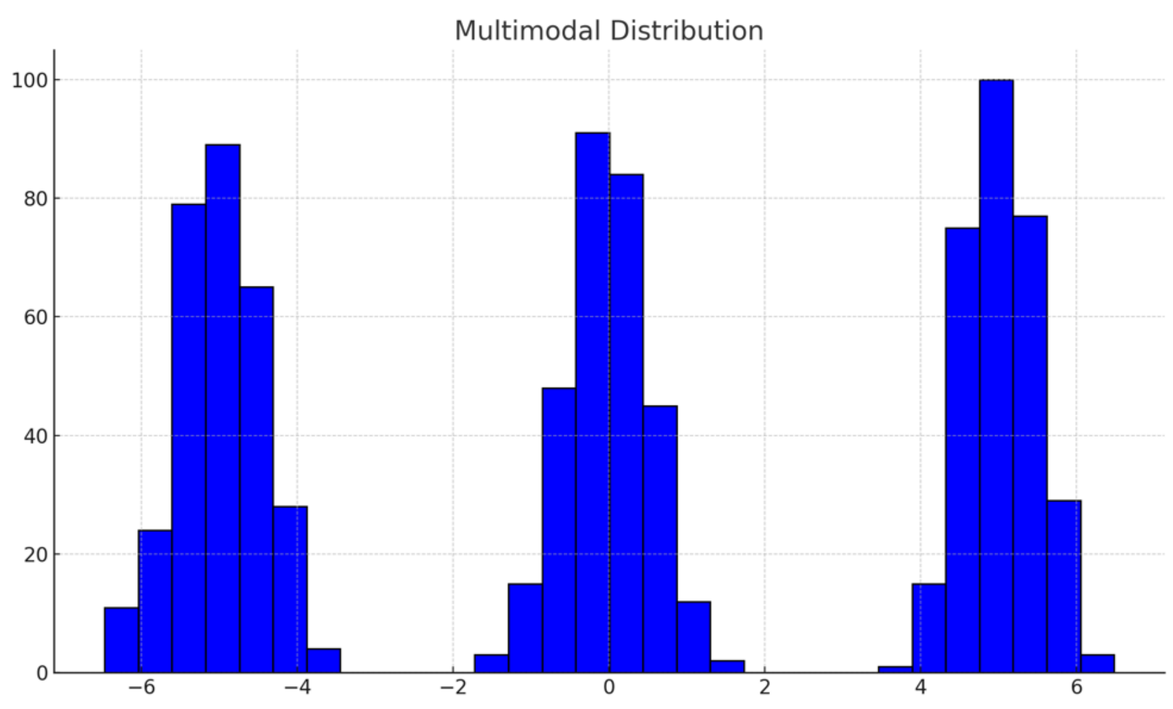
The histogram has one clear peak. This is common in many natural datasets and often indicates a single central tendency.



Bimodal:

Two distinct peaks suggest the presence of two different groups or processes in the data. For example, a histogram of heights might show two peaks if it includes both male and female data.

Peaks (Modes)



Multimodal:

More than two peaks, suggesting multiple underlying processes or groups within the data.

Managing Peaks

Unimodal:

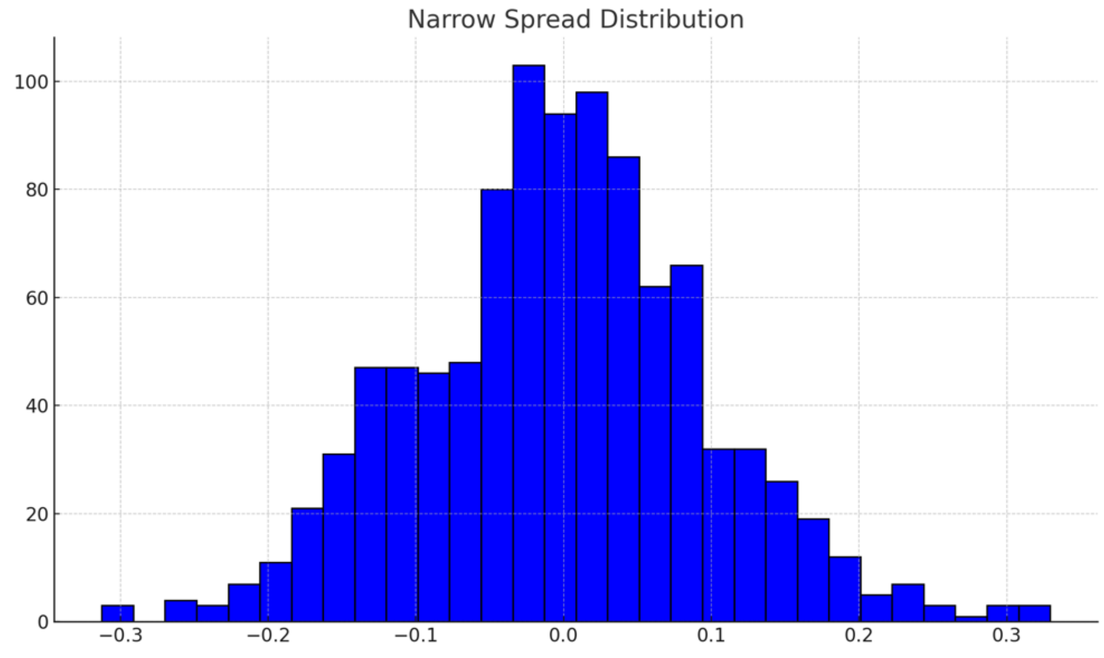
No Action Needed: No specific action is often required if the distribution is unimodal and the shape is as expected.

Bimodal/Multimodal:

Segmentation: Investigate if the different modes correspond to different subgroups or categories in the data. Analyse these groups separately.

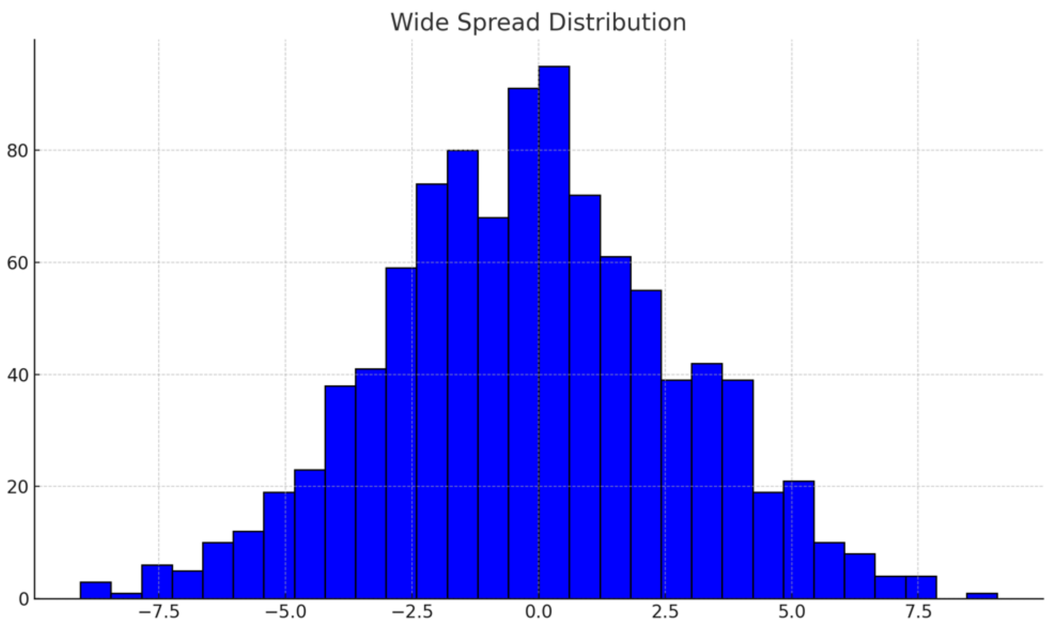
Further Investigation: Determine if the multiple peaks indicate different processes, sources, or populations and address them separately.

Spread /Variability



Narrow Spread

A histogram with a narrow spread indicates that the data values are concentrated around the centre. This suggests low variability.



Wide Spread

A wider spread suggests high variability, with data values distributed over a broader range.

Managing Spread

Narrow Spread

Precision Analysis: If variability is too low, ensure that measurement tools or methods are precise and sensitive enough to capture true variability in the data.

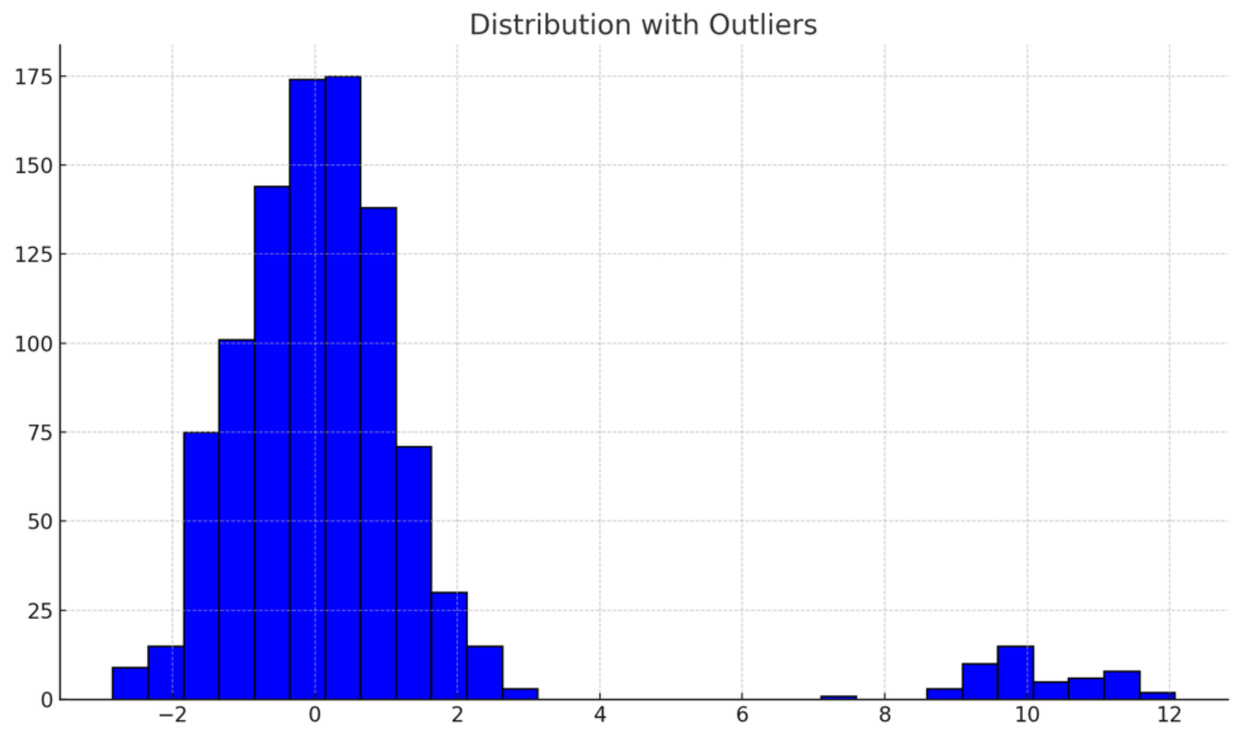
Wide Spread:

Subgroup Analysis: Identify the wide spread results from combining distinct groups. Separate them and analyse each group independently.

Data Transformation: Consider normalising or standardising the data to make it more manageable.

Outliers

Outliers can be identified by isolated bars far from the bulk of the data. These values may be genuine observations or indicate errors or anomalies in data collection.



Managing Outliers

Outlier Identification:
To identify outliers, use statistical methods like Z-scores, IQR (Interquartile Range), or Grubbs' test.

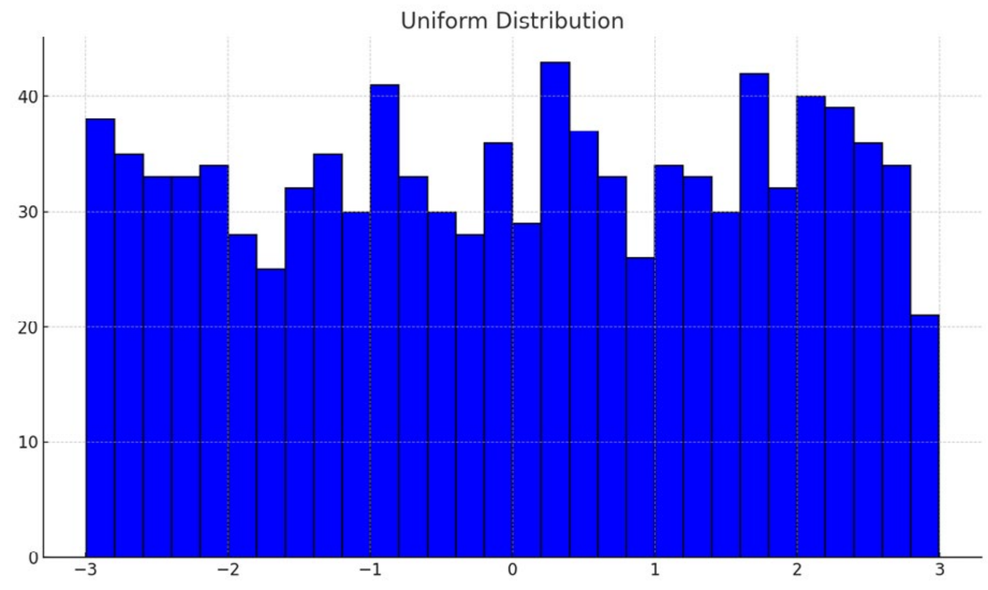
Outlier Management:

Investigate: Determine whether the outliers are due to data entry errors, measurement errors, or genuinely extreme cases.

Removal/Retention: Decide whether to remove, cap, or retain outliers based on their cause and impact on the analysis.

Robust Methods: Employ robust statistical methods that minimise the influence of outliers (e.g., using median instead of mean).

Uniform



If the histogram bars are roughly the same height across the range, it indicates a uniform distribution where all outcomes are equally likely.

Managing Uniformity

Data Segmentation: Check if the uniformity is appropriate for the context. If not, investigate whether it's masking underlying trends or subgroups.

Refined Analysis: Apply clustering or other segmentation techniques to uncover potential patterns hidden within the uniform distribution.

Gaps

Gaps between bars indicate ranges of data values where no observations occur. This could suggest distinct subgroups within the data or potential issues with data collection,

Managing Gaps

Data Quality Check: Ensure gaps aren't due to missing or faulty data. Review data collection methods.

Subgroup Analysis: Gaps might indicate different subgroups or processes. Analyse each segment separately to understand why there is a lack of data in specific ranges.

Modelling Gaps: Consider using methods like interpolation or imputation if gaps are due to missing data that needs to be addressed.

Granularity (Bin Width)

The granularity or the bin width in a histogram can affect interpretation. Too-wide bins may hide important details, while too narrow bins may exaggerate noise. The choice of bin size should be made carefully to ensure the histogram accurately represents the underlying distribution.

Managing Granularity

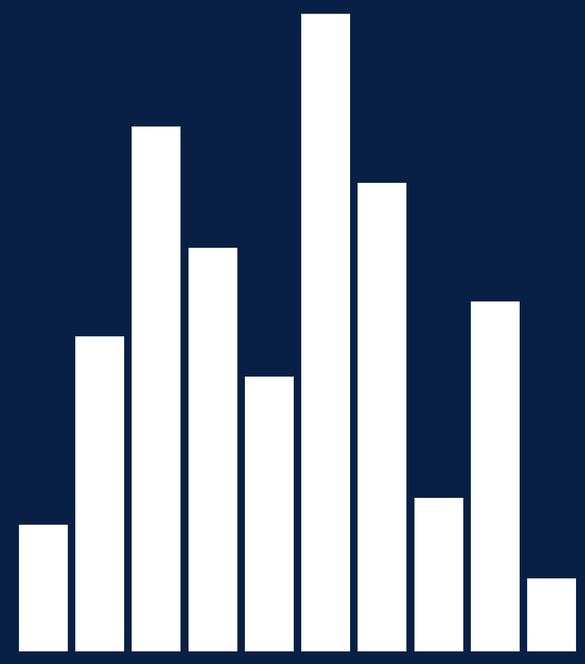
Adjusting Bin Width: Experiment with different bin widths to find a balance that reveals the underlying data structure without obscuring details or exaggerating noise.

Adaptive Binning: Consider using data-driven approaches like Freedman-Diaconis or Sturges' rule to determine optimal bin widths.

Cumulative Distribution: Complement the histogram with a cumulative distribution function (CDF) to provide a smoother overview of the data's distribution.

Key Takeaways

- Pareto Drilldown is a systematic approach to identify key factors or root causes.
- It helps focus on the “vital few” that truly matter.
- Each level of drilldown adds precision and clarity to the analysis.
- Use this method to convert overwhelming data into clear, actionable steps.



Reminder

Always look at the shape of the histogram first.

Don’t ignore skewness—it impacts mean and median.

Identify and understand outliers before acting on them.

Adjust bin width to reveal the most accurate data story.

Use peaks to identify different data segments.

Data tells a story—make sure you're reading it correctly!

Understanding histograms unlocks powerful insights.

Need assistance with analysis your data?
We stand ready to help.

What's the most surprising pattern you've found in your data?
Share your thoughts?

Also share this carousel with others.

Follow on LinkedIn

@pietropazzi

profitable processes

@profitableprocesses

profitableprocesses.com

info@profitableprocesses.com