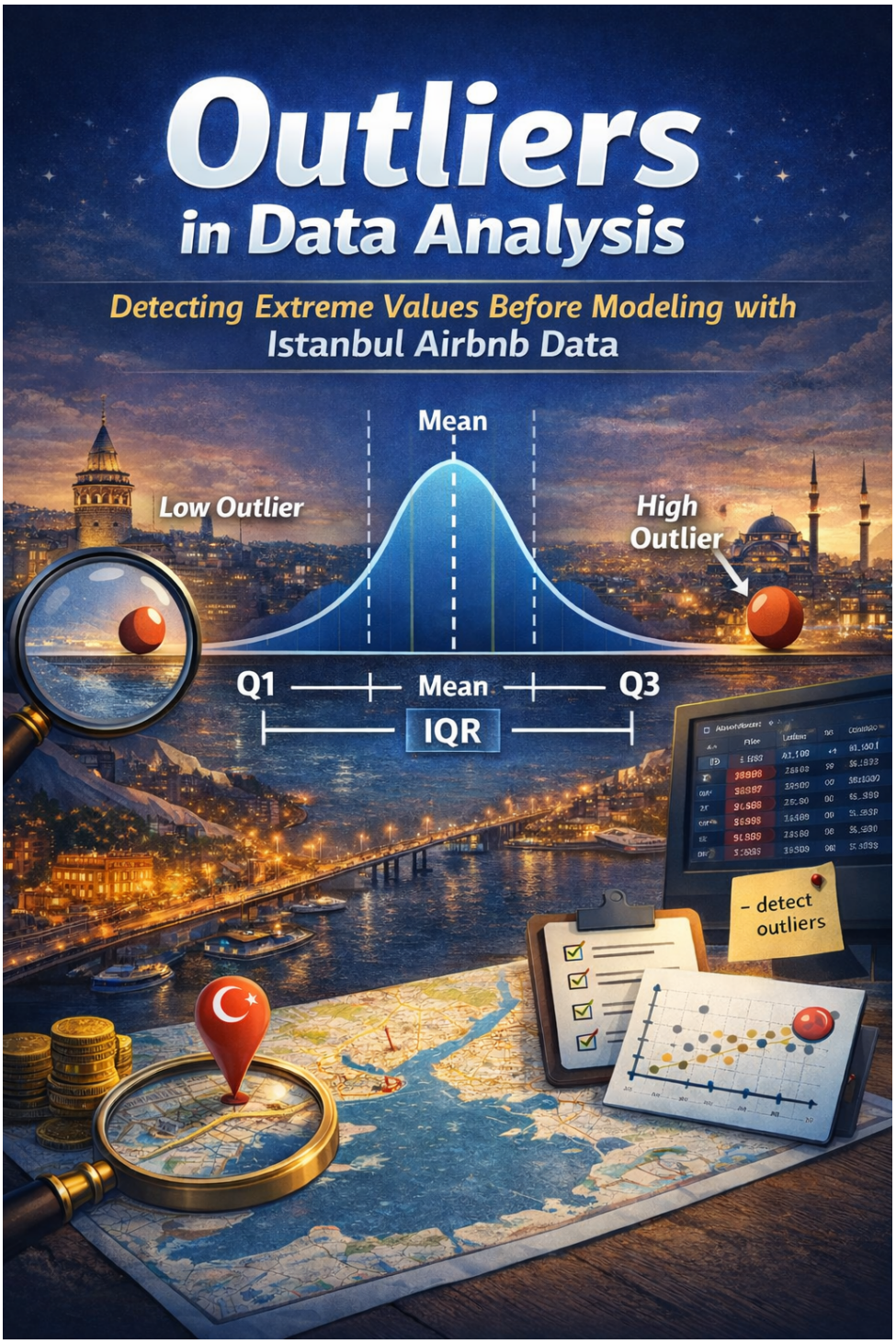
Outliers in Data Analysis: Detecting Extreme Values Before Modeling in R

M. Fatih Tüzen

2025-12-19

Table of contents


Outliers
in Data Analysis
Detecting Extreme Values Before Modeling with
Istanbul Airbnb Data
Mean
Low Outlier
High
Outlier
Q1
Mean
Q3
IQR
- detect
outliers

Introduction

Data preprocessing is often presented as a sequence of technical steps. However, each preprocessing decision implicitly embeds a statistical assumption.

In a previous article, I discussed how missing observations can bias analysis if they are ignored or handled improperly:

Handling Missing Data in R: A Comprehensive Guide:
https://medium.com/r-evolution/handling-missing-data-in-r-a-comprehensive-guide-eca195eaead3

This article continues that discussion by focusing on **outliers**. Unlike missing values, outliers are observed data points. The challenge is not their absence, but their *extremeness*.

Understanding whether an extreme value is informative or misleading is a crucial step before any modeling effort.

Why Outliers Matter

Outliers can affect statistical analysis in several fundamental ways:

- They distort summary statistics such as the mean and standard deviation
- They can dominate parameter estimates in regression models
- They influence distance-based methods such as clustering

More importantly, outliers force analysts to confront a key question:

> Are we observing rare but valid behavior, or a deviation from the assumed data-generating process?

What Is an Outlier?

Informally, an outlier is an observation that appears unusually large or small relative to the rest of the data.

Formally, an outlier is an observation that is inconsistent with the bulk of the data **under a given statistical model**.

Outliers are therefore not absolute objects. They depend on assumptions about distribution, scale, and structure.

The Dataset: Palmer Penguins

To demonstrate outlier detection methods, we use the `palmerpenguins` dataset.

```
library(palmerpenguins)
library(dplyr)

data(penguins)
```

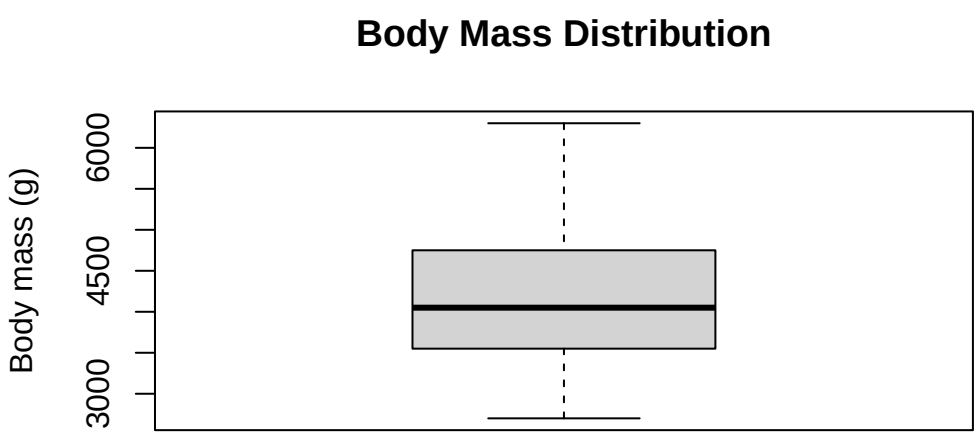
The dataset contains physical measurements of penguins observed in the Palmer Archipelago. In this article, we focus on the variable `body_mass_g`, measured in grams.

The dataset contains natural biological variability but does not exhibit extreme anomalies. This makes it suitable for illustrating *detection logic* rather than aggressive data cleaning.

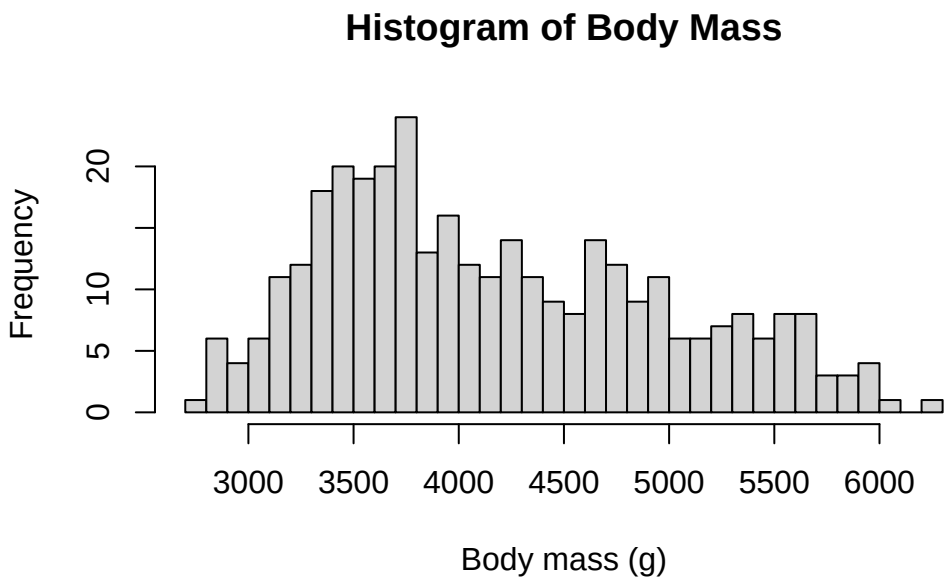
Visual Inspection: The Starting Point

Before applying formal rules, it is essential to examine the distribution visually.

```
boxplot(
  penguins$body_mass_g,
  main = "Body Mass Distribution",
  ylab = "Body mass (g)"
)
```




```
hist(
  penguins$body_mass_g,
  breaks = 30,
  main = "Histogram of Body Mass",
  xlab = "Body mass (g)"
)
```



Visual inspection highlights observations that lie far from the central mass of the data. At this stage, no decision is made. These plots serve as *signals*, not verdicts.

The IQR Method: A Robust Approach

Quartiles and the Interquartile Range

Quartiles divide an ordered dataset into four equal parts.

- The first quartile corresponds to the 25th percentile
- The third quartile corresponds to the 75th percentile

The interquartile range (IQR) is defined as:

$$\text{IQR} = Q_3 - Q_1$$

The IQR measures the spread of the middle 50 percent of the data and is robust to extreme values.

Defining Outliers Using the IQR Rule

A commonly used rule defines outliers as observations lying outside the interval:

$$[Q_1 - 1.5 \times \text{IQR},\ Q_3 + 1.5 \times \text{IQR}]$$

This rule does not rely on distributional assumptions and works well for skewed data.

IQR-Based Detection in R

```
Q1 <- quantile(penguins$body_mass_g, 0.25, na.rm = TRUE)
Q3 <- quantile(penguins$body_mass_g, 0.75, na.rm = TRUE)

IQR_value <- Q3 - Q1

penguins_iqr <- penguins %>%
  mutate(
    outlier_iqr =
      body_mass_g < Q1 - 1.5 * IQR_value |
      body_mass_g > Q3 + 1.5 * IQR_value
  )
```

This step **labels** potential outliers instead of removing them.

The Z-Score Method

Definition of the Z-Score

A Z-score measures how far an observation deviates from the mean in units of standard deviation.

The Z-score is defined as:

$$z = \frac{x - \mu}{\sigma}$$

where:

- μ is the sample mean
- σ is the sample standard deviation

A common heuristic flags observations satisfying:

$$|z| > 3$$

Assumptions and Limitations

The Z-score method assumes approximate symmetry and stable variance.

Extreme observations inflate the mean and standard deviation, reducing their own Z-scores. As a result, Z-score methods are sensitive to the very values they aim to detect.

Z-Score Detection in R

```
penguins_z <- penguins %>%
  mutate(
    z_score = as.numeric(scale(body_mass_g)),
    outlier_z = abs(z_score) > 3
  )
```

The method is simple and intuitive but should not be used in isolation.

Comparing IQR and Z-Score Methods

Different methods often flag different observations.

- IQR is rank-based and robust
- Z-scores rely on the mean and variance

Rather than asking which method is correct, the more relevant question is:

> Which assumptions are appropriate for this dataset and modeling goal?

Should Outliers Be Removed?

Outlier detection does not imply automatic removal.

Possible strategies include:

- verifying and correcting data errors
- applying transformations or robust estimators
- explicitly modeling rare events

Understanding *why* an observation is extreme is often more informative than deleting it.

Final Remarks

Outlier detection is a conceptual step that naturally follows missing data analysis and precedes scaling or transformation.

This article focused on **understanding and identifying** extreme values rather than eliminating them. Extreme values are not noise by default. They are questions posed by the data.

References and Further Reading

- Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.
- Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.
- James, G., Witten, D., Hastie, T., Tibshirani, R. (2021). *An Introduction to Statistical Learning*. Springer.
- NIST/SEMATECH e-Handbook of Statistical Methods – Outliers
  https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm
- Wickham, H., Grolemund, G. *R for Data Science*
  https://r4ds.hadley.nz/
- Palmer Penguins Dataset
  https://allisonhorst.github.io/palmerpenguins/