ANOMALY DETECTION IN R

# What is an anomaly?

Alastair Rushworth
Data Scientist

# Defining the term anomaly

***Anomaly:*** *a data point or collection of data points that do not follow the same pattern or have the same structure as the rest of the data*

# Point anomaly

- A single data point
- Unusual when compared to the rest of the data

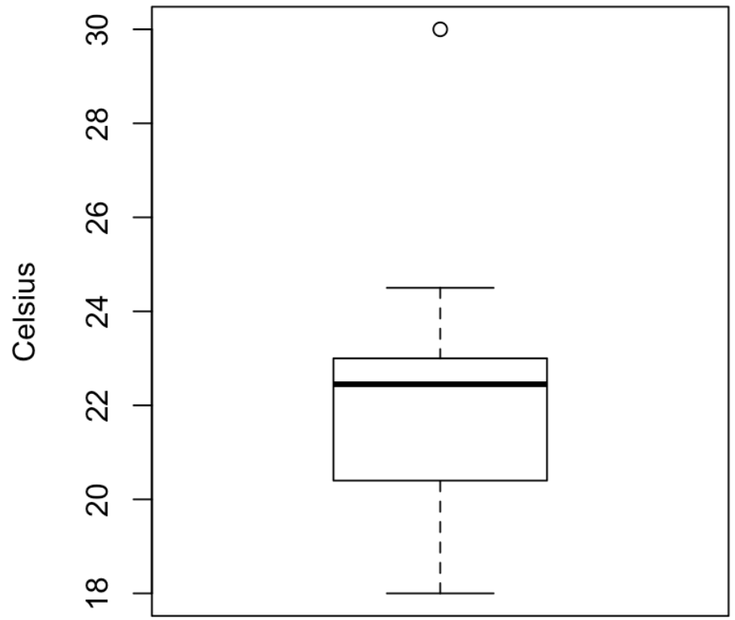
**Example:** A single 30C daily high temperature among a set of ordinary spring days

```
summary(temperature)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   20.45   22.45   22.30   22.98   30.00
```

# Visualizing point anomalies with a boxplot

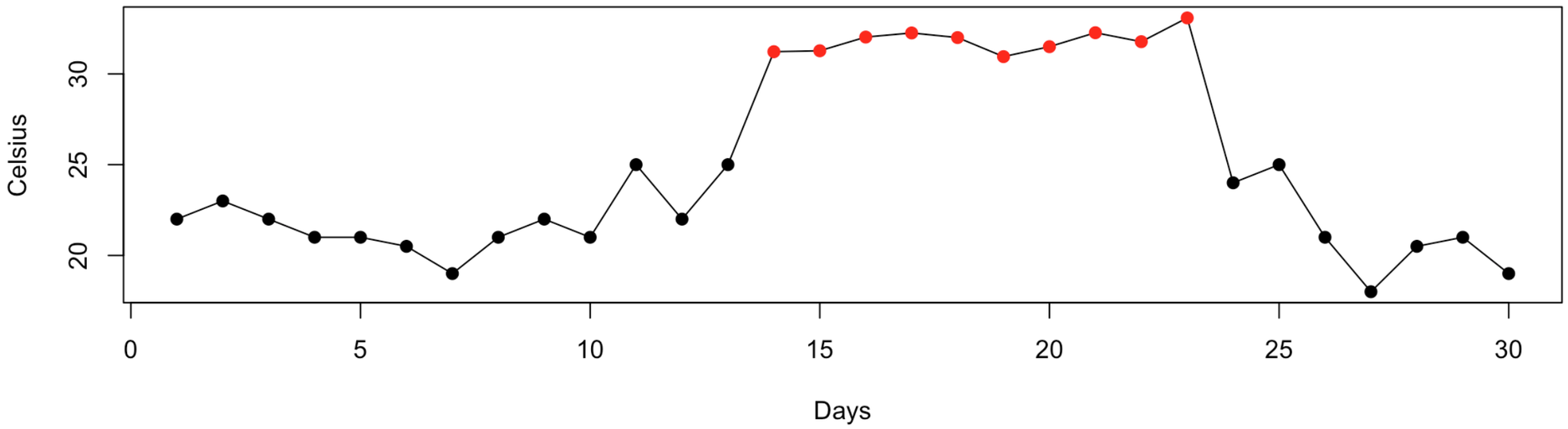
```
boxplot(temperature, ylab = "Celsius")
```



# Collective anomaly

- An anomalous collection of data instances
- Unusual when considered together

**Example:** 10 consecutive high daily temperatures



ANOMALY DETECTION IN R

# Let's practice!

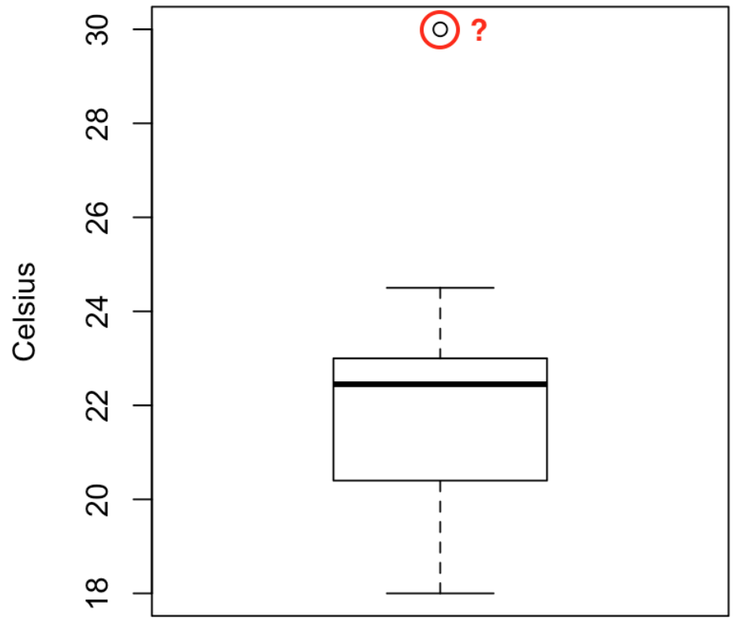
ANOMALY DETECTION IN R

# Testing the extremes with Grubbs' test

Alastair Rushworth
Data Scientist

# Visual assessment is not always reliable!

```
boxplot(temperature, ylab = "Celsius")
```

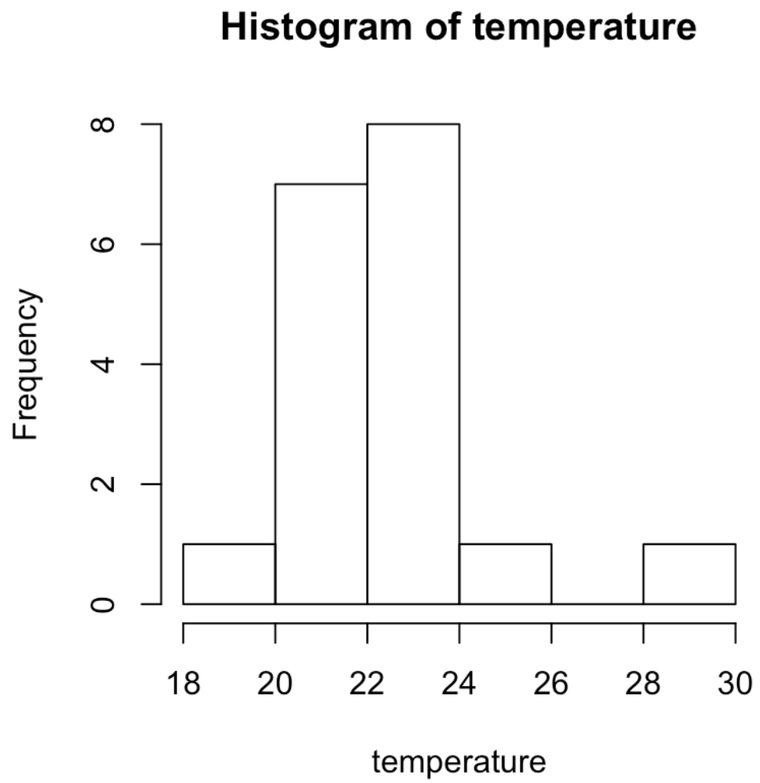


# Grubbs' test

- Statistical test to decide if a point is outlying
- Assumes the data are normally distributed
- Requires checking the normality assumption first

# Checking normality with a histogram

```
hist(temperature, breaks = 6)
```




Symmetrical & bell shaped?

# Running Grubbs' test

Use the `grubbs.test()` function:

```
grubbs.test(temperature)
```

```
    Grubbs test for one outlier
data:  temp
G = 3.07610, U = 0.41065, p-value = 0.001796
alternative hypothesis: highest value 30 is an outlier
```

# Interpreting the p-value

```
grubbs.test(temperature)
```

```
    Grubbs test for one outlier

data:  temperature
G = 3.07610, U = 0.41065, p-value = 0.001796
alternative hypothesis: highest value 30 is an outlier
```

`p-value`

- Near 0 - *stronger* evidence of an outlier
- Near 1 - *weaker* evidence of an outlier

# Get the row index of an outlier

## Location of the **maximum**

```
which.max(weights)
```

```
[1] 5
```

## Location of the **minimum**

```
which.min(temperature)
```

```
[1] 12
```

ANOMALY DETECTION IN R

# Let's practice!

ANOMALY DETECTION IN R

# Detecting multiple anomalies in seasonal time series

Alastair Rushworth
Data Scientist

# Monthly revenue data

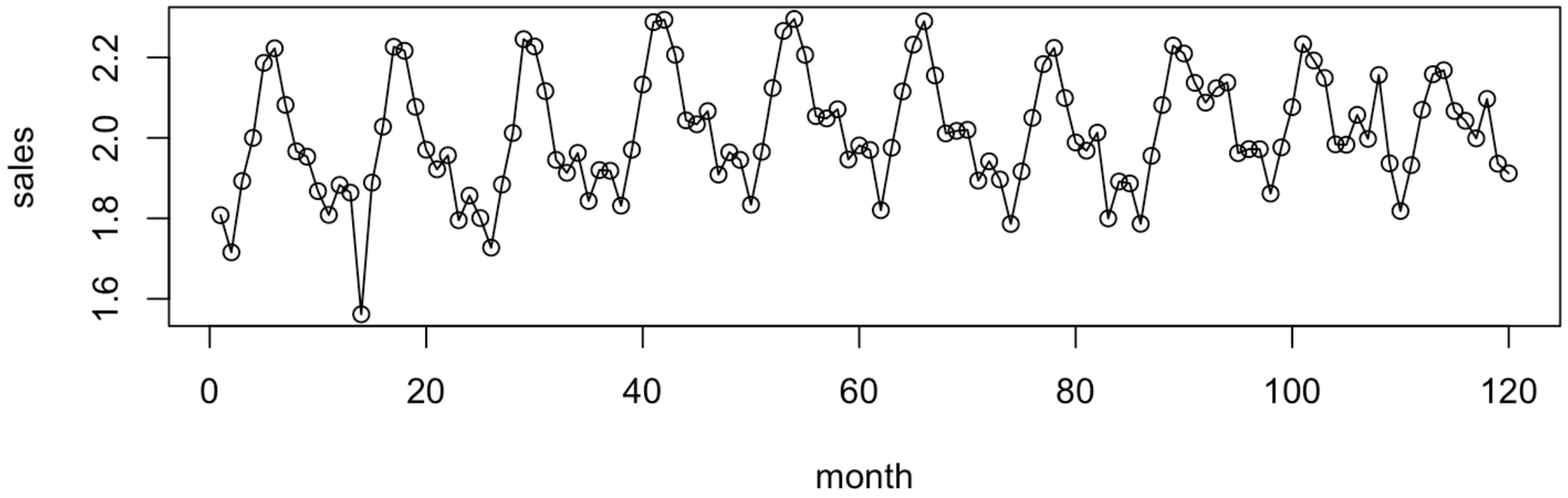
```
head(msales)

  sales month
1 6.068     1
2 5.966     2
3 6.133     3
4 6.230     4
5 6.407     5
6 6.433     6
```

Grubbs' test not appropriate here

- Seasonality may be present
- May be multiple anomalies

# Visualizing monthly revenue

```
plot(sales ~ month, data = msales, type = 'o')
```



# Seasonal-Hybrid ESD algorithm usage

```
library(AnomalyDetection)

sales_ad <- AnomalyDetectionVec(x = msales$sales, period = 12,
                                direction = 'both')
```

**Arguments**

- `x`: vector of values
- `period`: period of repeating pattern
- `direction`: find anomalies that are small (`'neg'`), large (`'pos'`), or both (`'both'`)

--

☐ Download from https://github.com/twitter/AnomalyDetection

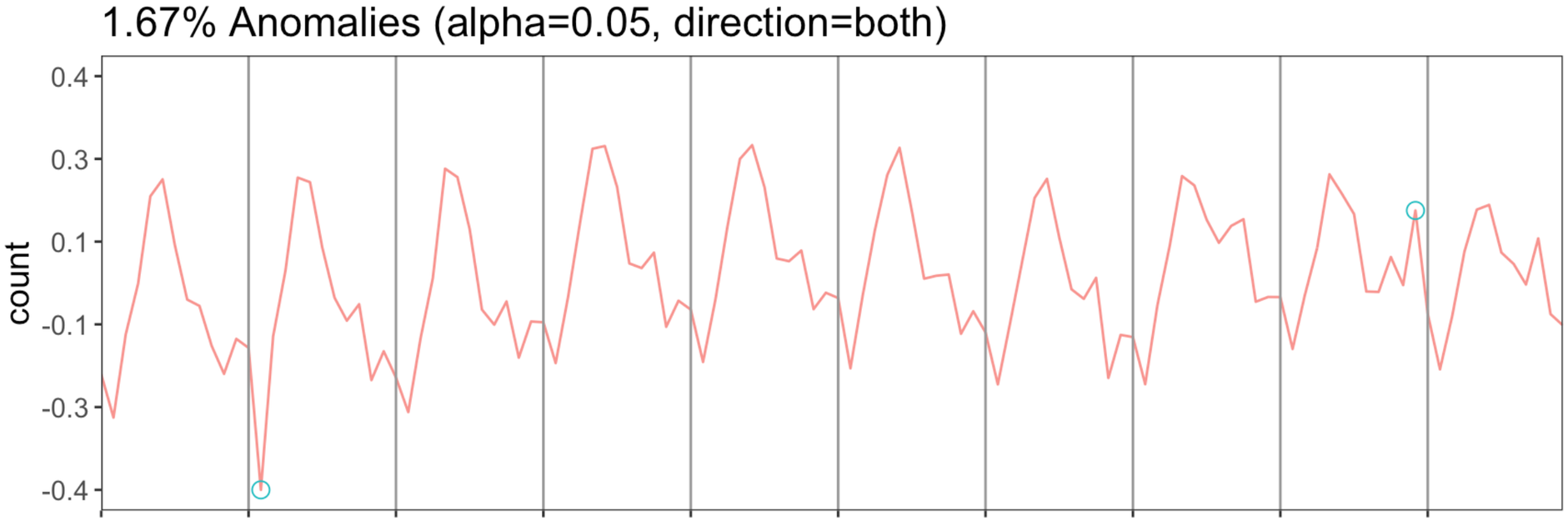
# Seasonal-Hybrid ESD algorithm output

```
sales_ad <- AnomalyDetectionVec(x = msales$sales, period = 12,
                                direction = 'both')
```

```
sales_ad$anoms

  index anoms
1    14 1.561
2   108 2.156
```

# Seasonal-Hybrid ESD algorithm plot

```
AnomalyDetectionVec(x = msales$sales, period = 12,
                    direction = 'both', plot = T)
```



ANOMALY DETECTION IN R

# Let's practice!