



ANOMALY DETECTION IN R

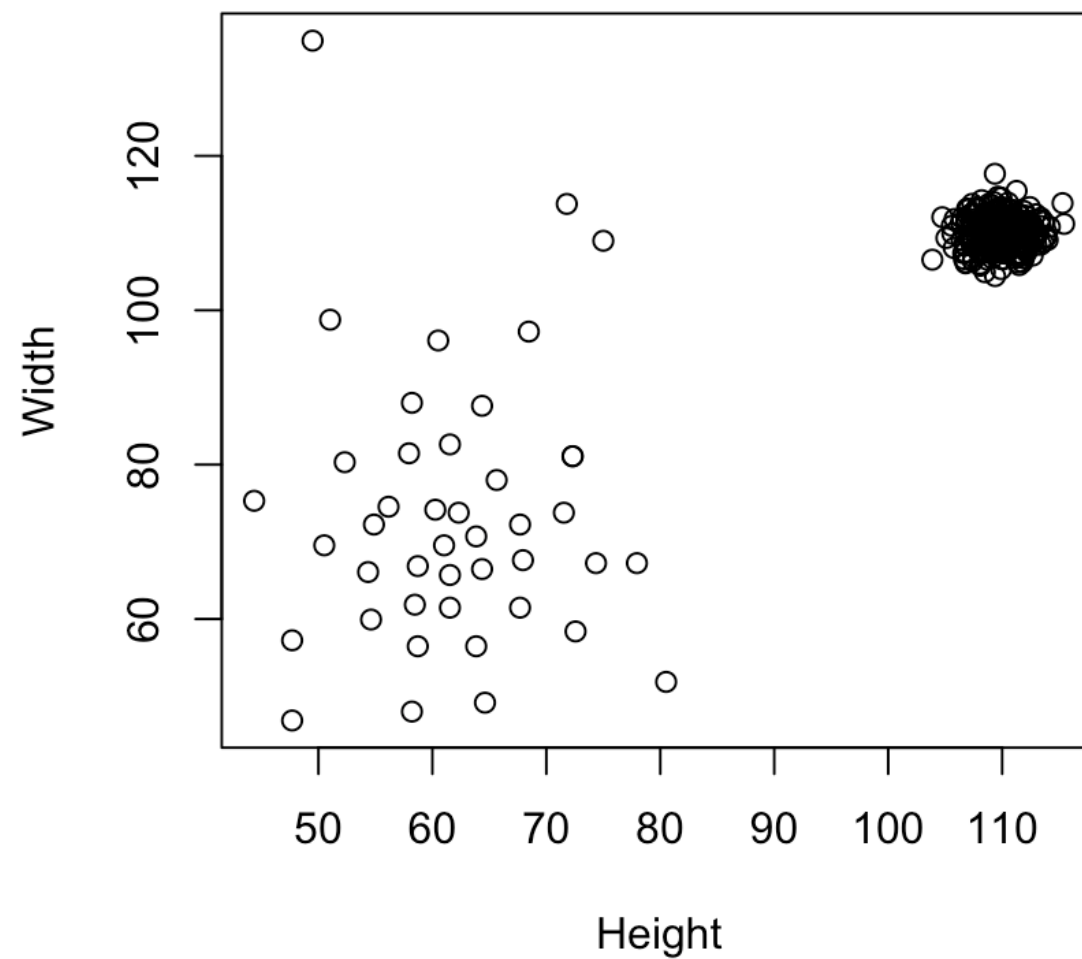
k-nearest neighbors distance score

Alastair Rushworth
Data Scientist



Furniture dimensions

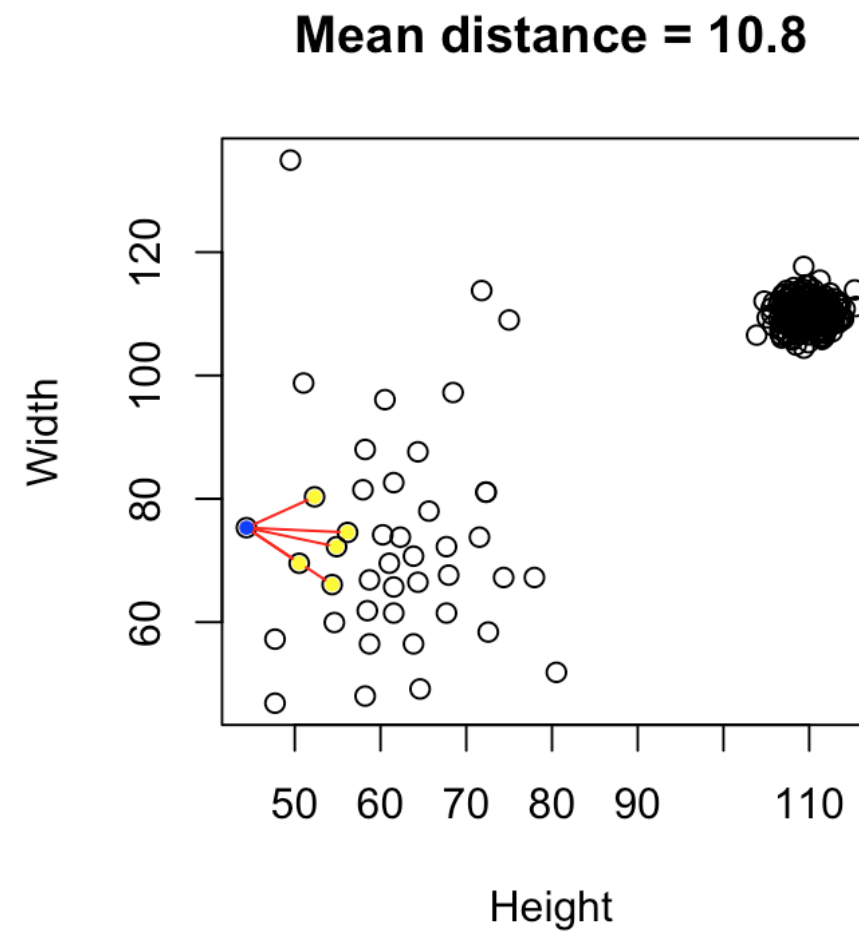
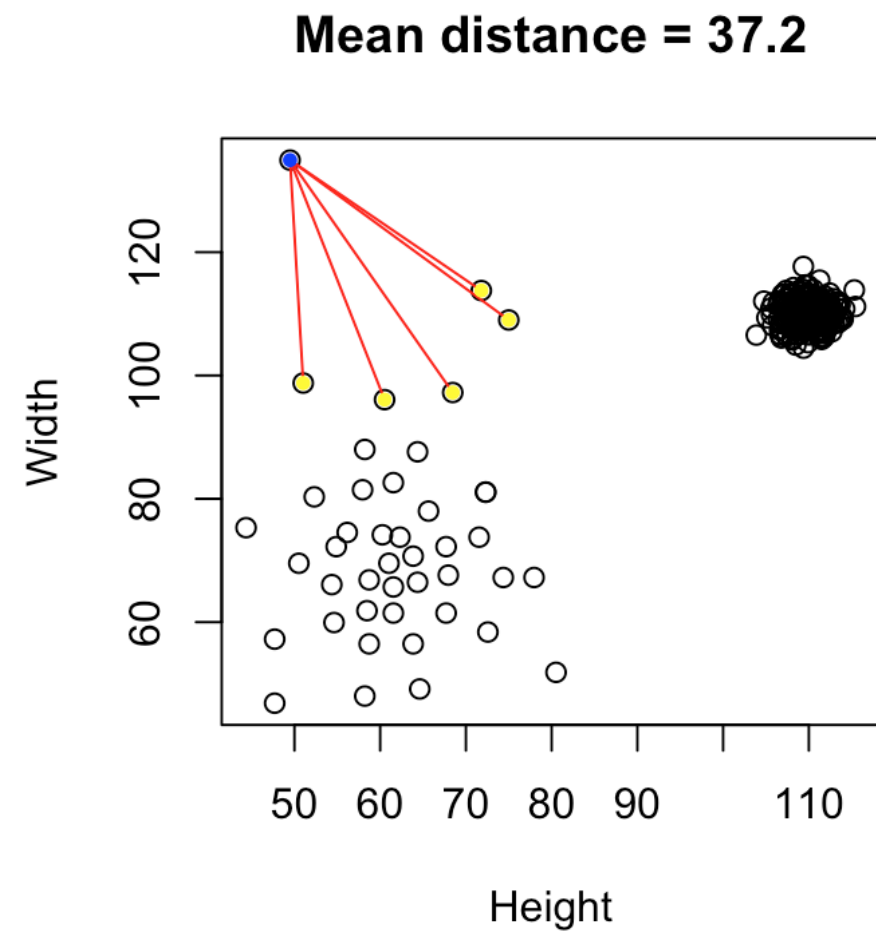
```
plot(Width ~ Height, data = furniture)
```





k-nearest neighbors (kNN) distance

Anomalies usually lie far from their neighbors





Inputs for distance matrix calculation

```
library(FNN)
furniture_knn <- get.knn(data = furniture, k = 5)
```

Arguments

- `data`: matrix of data
- `k`: the number of neighbors



Distance matrix output

`get.knn()` **returns two matrices**

```
names(furniture_knn)
```

```
[1] "nn.index" "nn.dist"
```

Distance matrix

```
head(furniture_knn$nn.dist, 3)
```

```
      [,1]      [,2]      [,3]      [,4]      [,5]  
[1,] 5.128300 5.367791 5.390801 5.740713 8.477025  
[2,] 4.300093 5.367791 6.159139 7.091966 7.428176  
[3,] 3.047502 3.545978 4.426266 5.006570 5.654202
```



kNN distance score

Average distance to nearest neighbors

```
furniture_score <- rowMeans(furniture_knn$nn.dist)
```

Largest score?

```
which.max(furniture_score)
```

```
[1] 29
```



ANOMALY DETECTION IN R

Let's practice!



ANOMALY DETECTION IN R

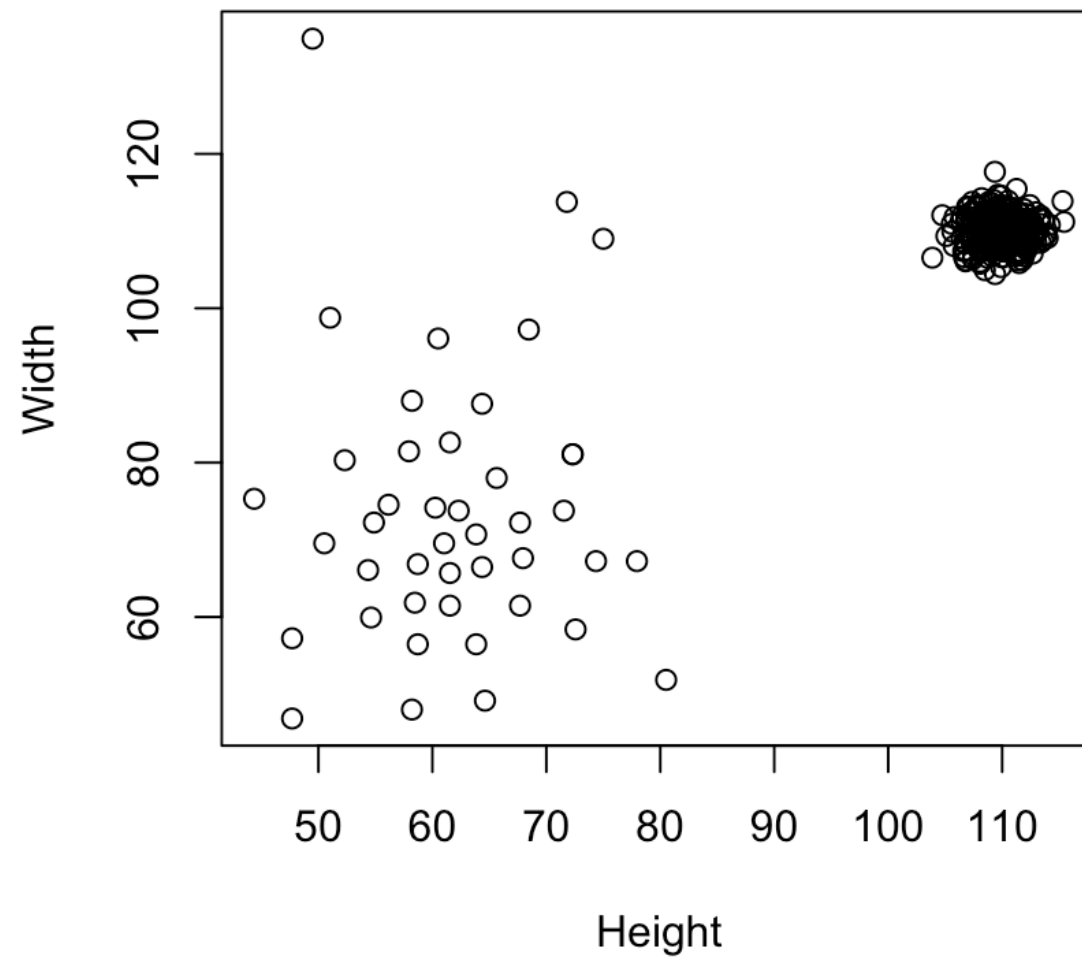
Visualizing kNN distance score

Alastair Rushworth
Data Scientist



Standardizing feature scales

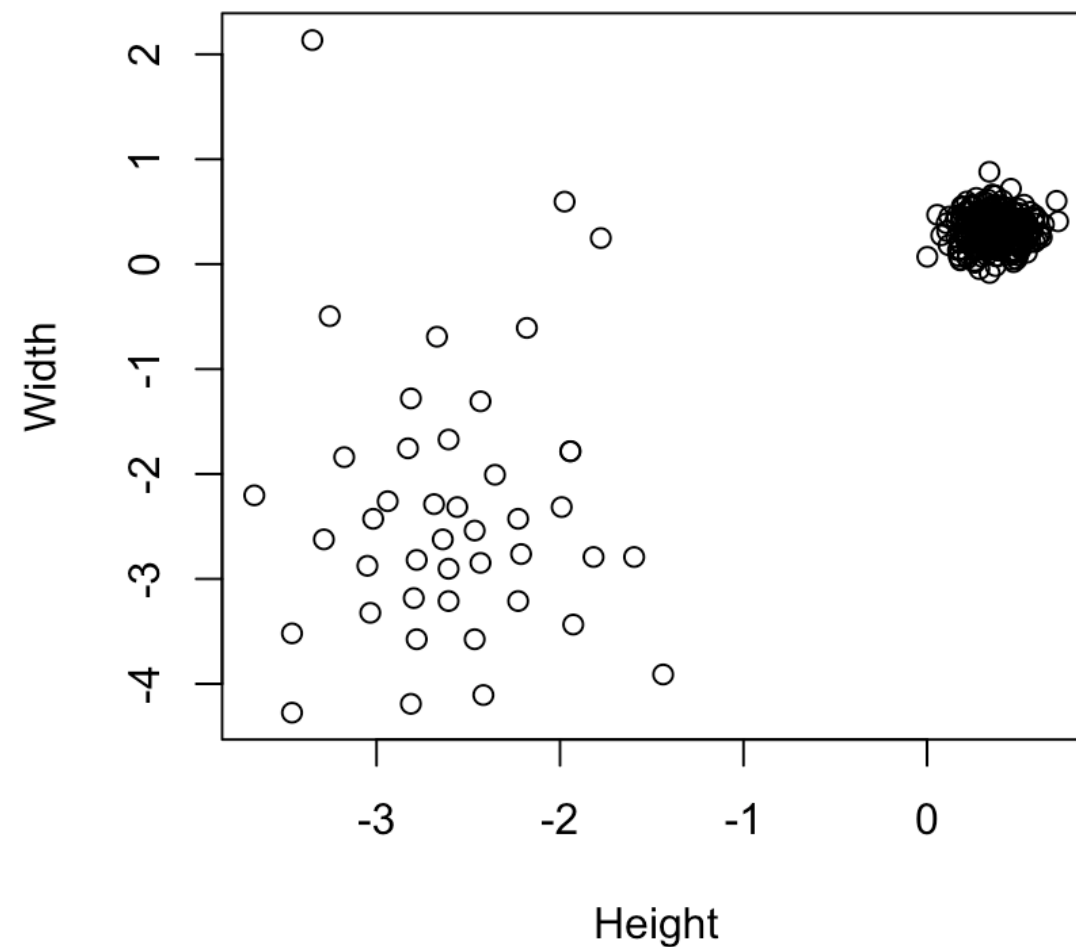
```
plot(Width ~ Height, data = furniture)
```



Standardizing features

```
furniture_scaled <- scale(furniture)
```

```
plot(Width ~ Height, data = furniture_scaled)
```





Create and append distance score

Distance matrix

```
furniture_scaled <- scale(furniture)
furniture_knn    <- get.knn(furniture_scaled, 5)
```

Calculate and append score

```
furniture$score <- rowMeans(furniture_knn$nn.dist)
```

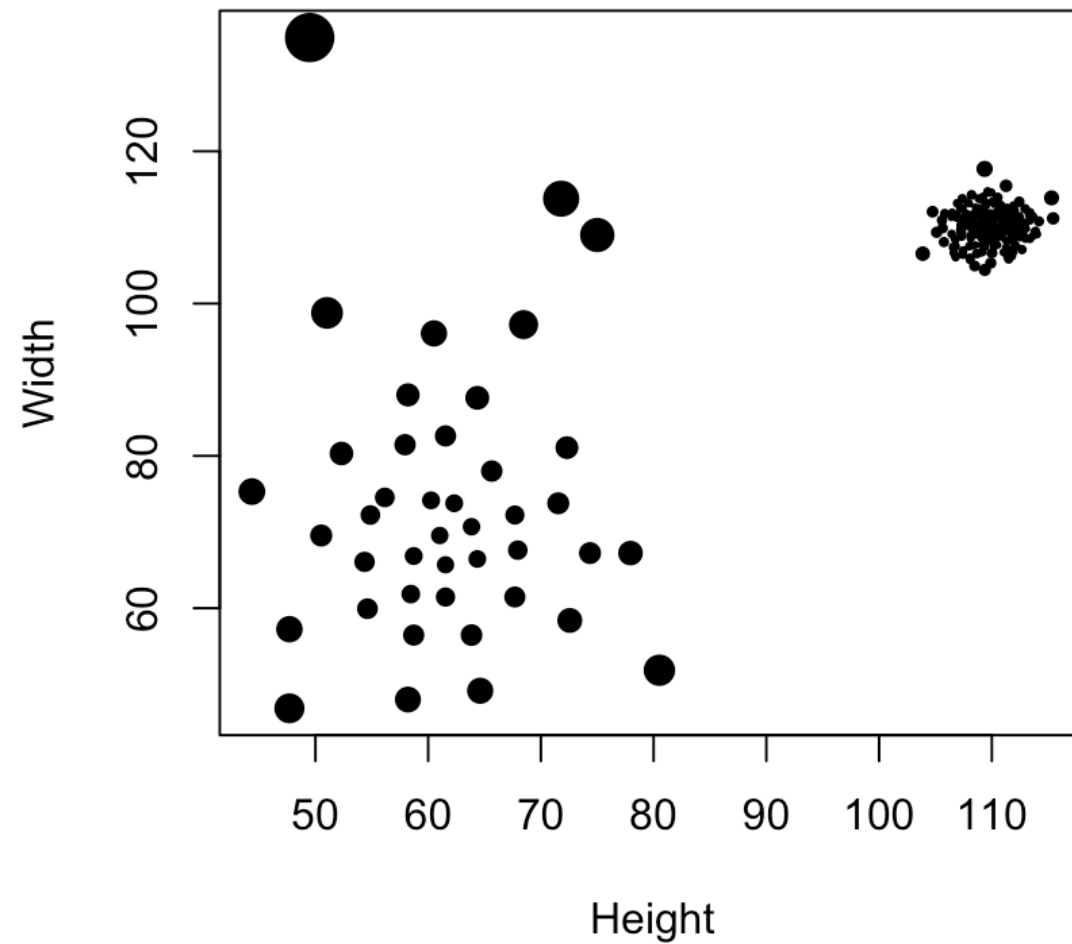
```
head(furniture, 4)
```

	Height	Width	score
1	58.7179	56.4663	0.4170000
2	54.6154	59.9279	0.3981695
3	58.7179	66.8510	0.2845042
4	63.8462	56.4663	0.4376807



Visualizing distance score

```
plot(Width ~ Height, cex = sqrt(score), data = furniture, pch = 20)
```





ANOMALY DETECTION IN R

Let's practice!



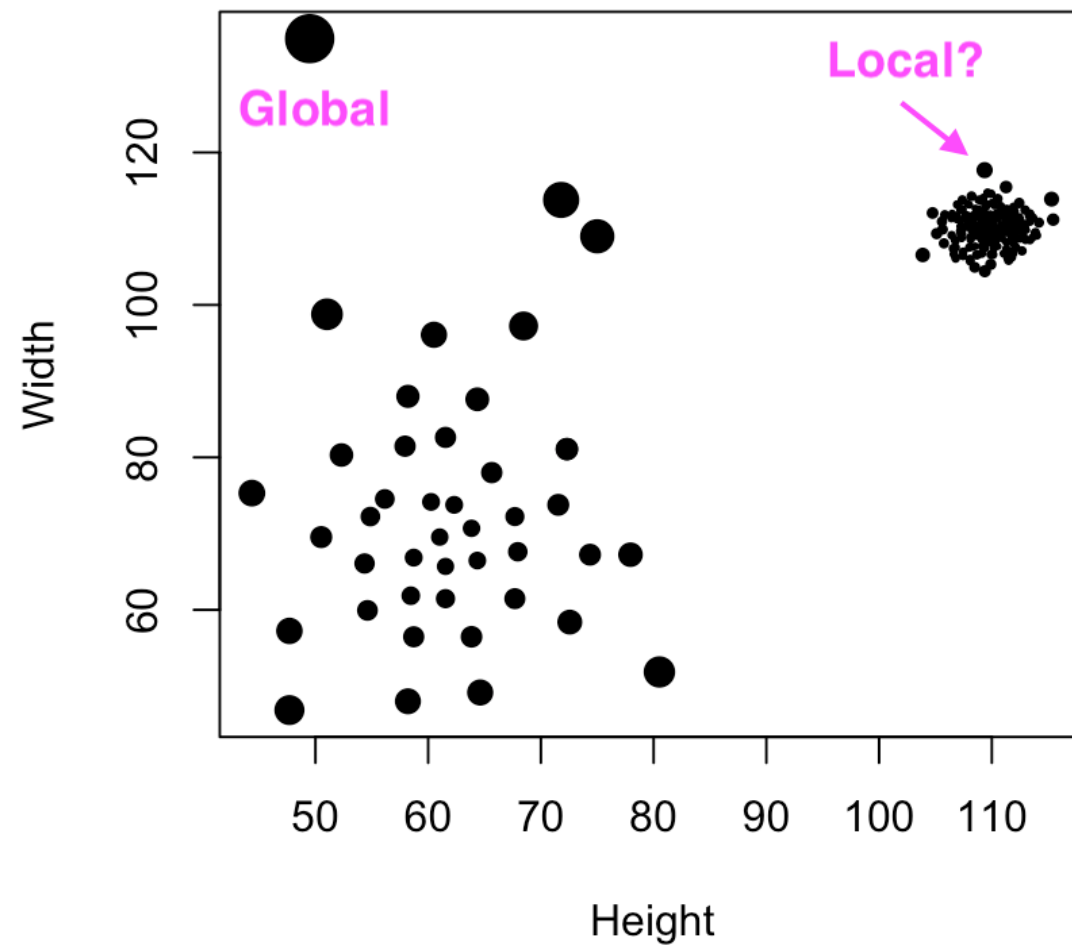
ANOMALY DETECTION IN R

The local outlier factor (LOF)

Alastair Rushworth
Data Scientist

Postmortem of kNN distance

Global versus local anomalies





Calculating LOF

Obtain LOF for furniture data

```
library(dbscan)
furniture_lof <- lof(scale(furniture), k = 5)
```

View the scores

```
furniture_lof[1:10]
```

```
[1] 1.0649669 1.1071205 0.9980290 1.0392385 0.9725305
[6] 1.1933199 1.3210459 1.1409659 1.0613144 1.0805445
```




Interpreting LOF

LOF is a ratio of densities

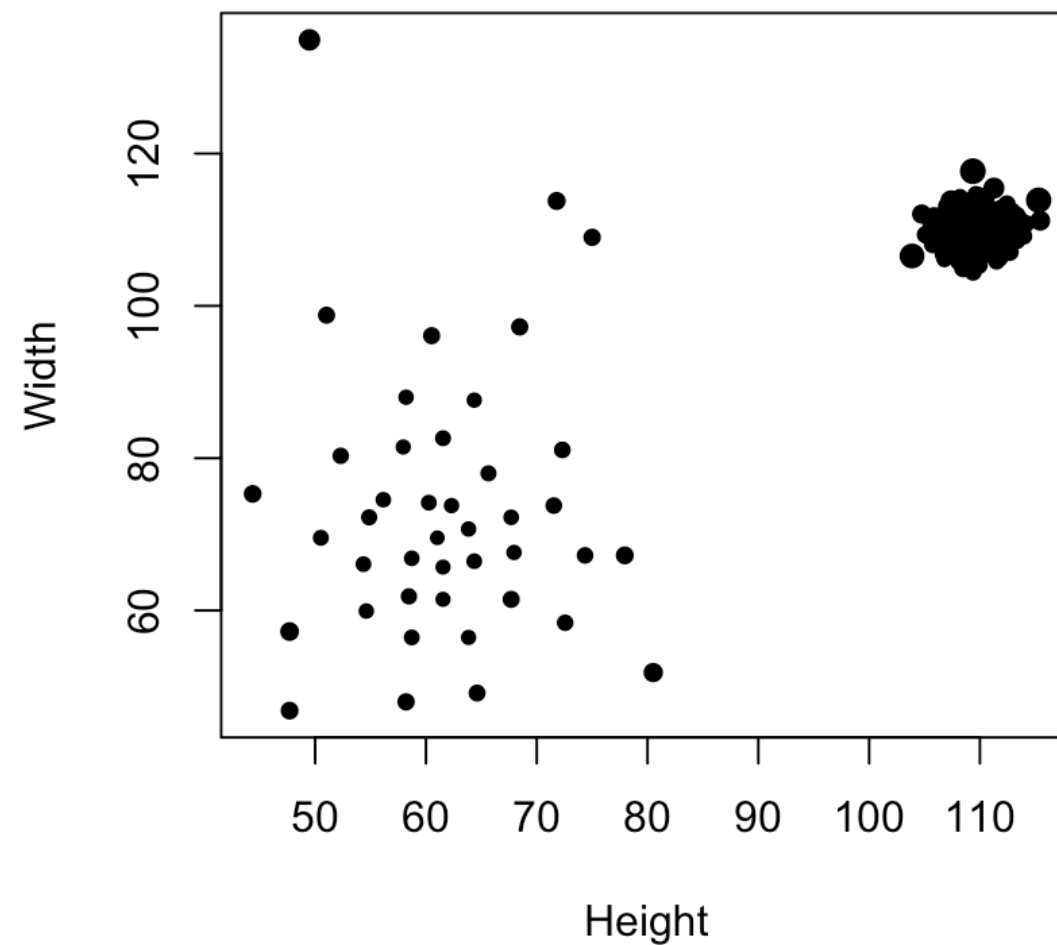
- $\text{LOF} > 1$ more likely to be anomalous
- $\text{LOF} \leq 1$ less likely to be anomalous

Large LOF values indicate more isolated points

Visualizing LOF

```
furniture$score_lof <- furniture_lof
```

```
plot(Width ~ Height, data = furniture, cex = score_lof, pch = 20)
```





ANOMALY DETECTION IN R

Let's practice!