# Data Mining
# Laboratory

Ewa Figielska

Warsaw School of Computer Science

# Outliers

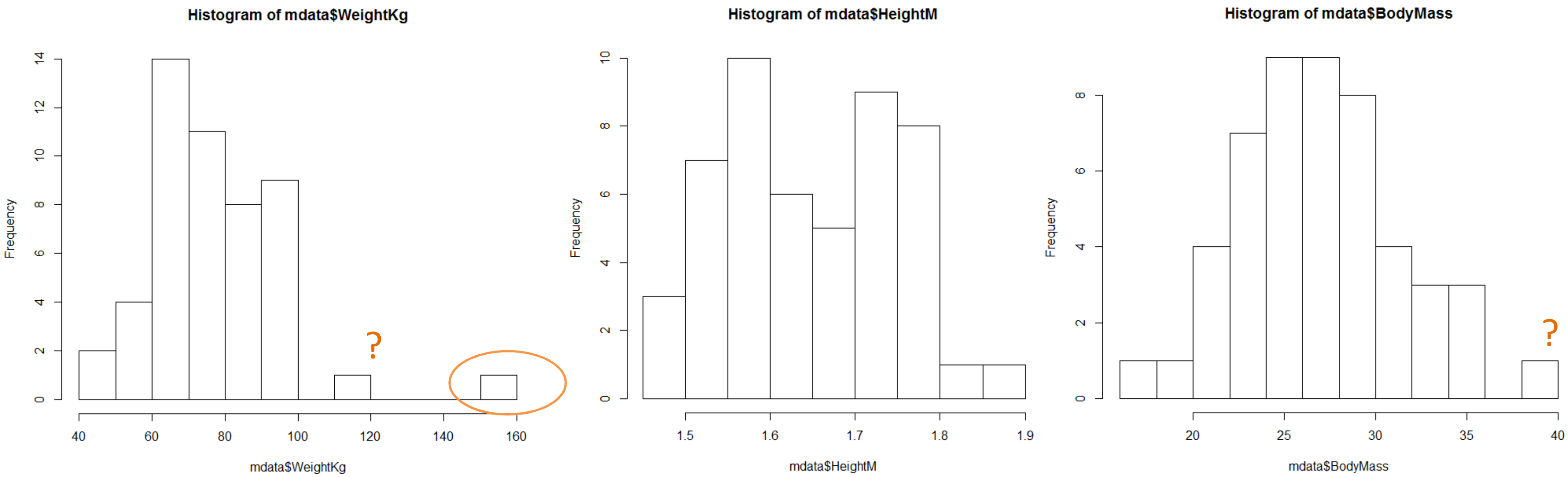
- *Outliers* are extreme values that lie near the limits of the data range or go against the trend of the remaining data.

- Identifying outliers is important because they may represent errors in data entry.
- If an outlier is a valid data point and not an error, certain statistical methods are sensitive to the presence of outliers and may deliver unstable results.

- Graphical methods for identifying outliers (for numeric variables):
  - histogram
  - two-dimensional scatter plot

- Numerical methods for identifying outliers:
  - Z-score method
  - interquartile range (IQR)

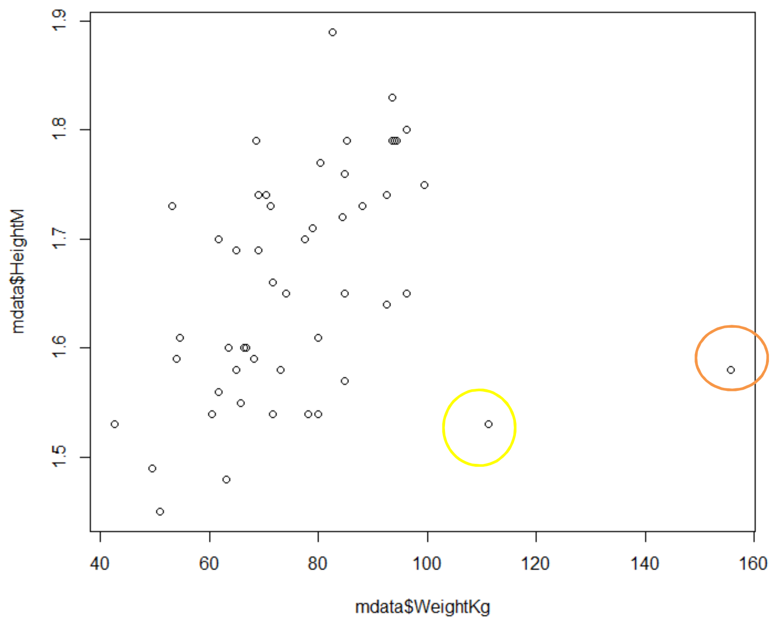
# Histogram

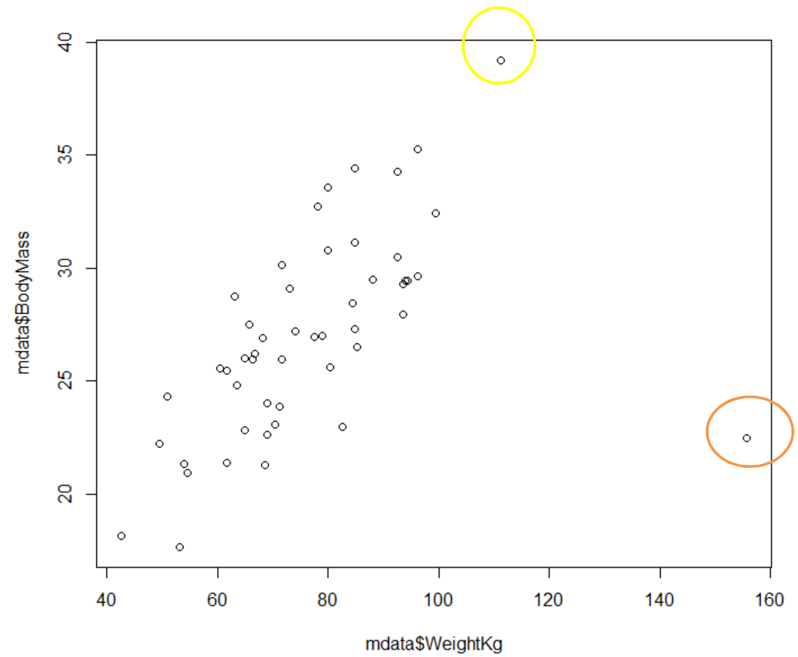



- Data set : “body_mass_index.csv”
- There appears to be one lonely value of Weight in the right tail of the distribution (record 19: 155.66, 1.58, 22.49). This point can be clearly identified as an outlier.
- Points with Weight = 111.13 and BodyMass = 39.22  (record 22) may represent  errors or unusual values (but it is not clear if they should be considered as outliers).

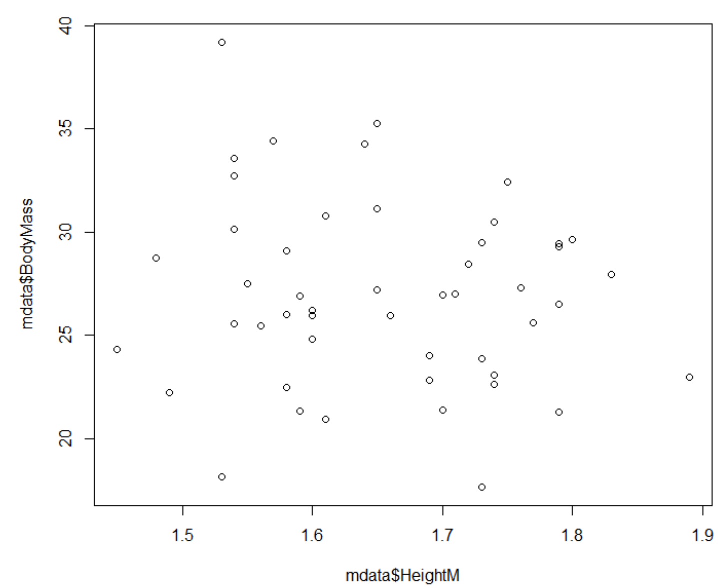
```
> mdata <- read.csv(file="body_mass_index.csv", header=TRUE, sep=",")
> hist(mdata$WeightKg,breaks=10)
> hist(mdata$HeightM,breaks=15)
> hist(mdata$BodyMass,breaks=15)
```

# Two-dimensional scatter plot




- The scatter plots of Height against Weight and of BodyMass against Weight show one outlier: with value of WeightKg equal to 155.66 (record 19).
- The point with the values 111.3 and 39.22 of the WeightKg and BodyMass variables (record 22), respectively, probably needs a closer look, too.
- The scatter plot of BodyMass against Height does not reveal any outliers.




```
> plot(mdata$WeightKg,mdata$HeightM)
```

# Z-score method

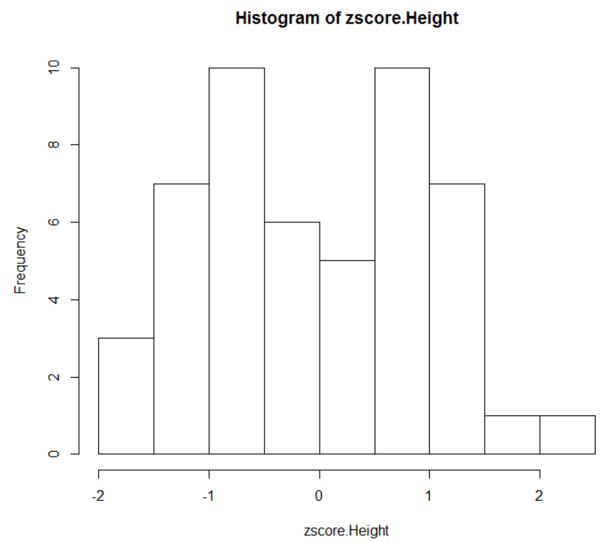
Z-score method for identifying outliers states that a data value is an outlier if it has a Z-score that is either less than -3 or greater than 3.

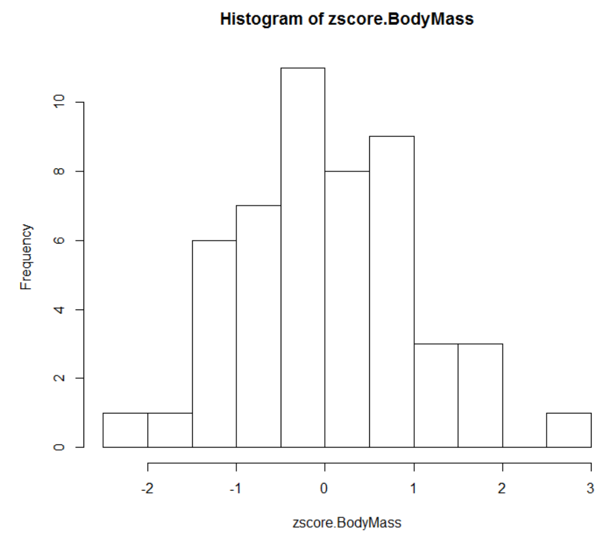
- **Standard normal distribution Z** - the normal distribution with $\mu = 0$ and standard deviation $\sigma = 1$

- **Z-score standardization** – takes the difference between the field value and the variable mean, and scales this difference by the standard deviation of the variable values.

$$\text{Z}-\text{score} = \frac{X-\text{mean}(X)}{\text{SD}(X)}$$

# After Z-score standardization


- Z-score standardization shows one outlier: 4.22 is the standardized value of Weight = 155.66.

  (See below the code for determining the standardized values of Weight which are less than -3 or greater than 3)

**1. zscore.Weight<-(mdata$WeightKg - mean(mdata$WeightKg))/sd(mdata$WeightKg)**
**2. hist(zscore.Weight,breaks=10)**

**3. zscore.Weight.outliers<-mdata$WeightKg [(zscore.Weight <(-3)) | (zscore.Weight >3)]**
**4. zscore.Weight.outliers**
**[1] 155.66**

#In line 1, the Z-score standardization of mdata$WeightKg is performed, and the values of this variable after standardization are stored in object zscore.Weight
# In line 3, the values of mdata$WeightKg are chosen for which zscore.Weight is less than (-3) or zscore.Weight is greater than 3 (I is the or operator)

# Z-score method drawbacks

- *Z*-score standardization depends on the mean and standard deviation.
- Both the mean and standard deviation are sensitive to the presence of outliers.
- If an outlier is added to a data set, the values of mean and standard deviation will be affected by this new data value.
- When choosing a method for evaluating outliers, it may not seem appropriate to use measures which are themselves sensitive to their presence.

# Interquartile range (IQR)

- The **quartiles** of a data set divide the data set into four parts, each containing 25% of the data.
  - The **first quartile** (Q1) is the 25th percentile.
  - The **second quartile** (Q2) is the 50th percentile, that is, the median.
  - The **third quartile** (Q3) is the 75th percentile.

(The $n^{th}$ percentile of an observation variable is the value that cuts off the first n percent of the data values when they are sorted in ascending order).

- The **interquartile range (IQR)** is a measure of variability.

  IQR = Q3 − Q1

  IQR represents the spread of the middle 50% of the data.

- A data value is an outlier if:
  a) It is located 1.5*IQR or more below Q1, or
  b) It is located 1.5*IQR or more above Q3.

Quartile calculations – example:

```
> data<-c(1,5,78,18,9,101,82,13,15,4,94,112)
> quantile(data,c(0.25,0.5,0.75))
    25% 50% 75%
      5   15   82
> sort(data)
[1] 1 4 5 9 13 15 18 78 82 94 101 112
```

#function c() yields the vector of numbers

# Detecting outliers using the IQR method

| WeightKg | HeightM | BodyMass |
|---|---|---|
| **Q1** = 65.0875<br>**Q2** = 73.4850<br>**Q3** = 85.1650<br>**IQR** = 20.0775<br><br>Q1 – 1.5*IQR =34.97125<br>Q3 +1.5*IQR = 115.2812 | **Q1** = 1.58<br>**Q2** = 1.65<br>**Q3** = 1.74<br>**IQR** = 0.16<br><br>Q1 – 1.5*IQR =1.34<br>Q3 +1.5*IQR = 1.98 | **Q1** = 23.9075<br>**Q2** = 26.9250<br>**Q3** = 29.4875<br>**IQR** = 5.58<br><br>Q1 – 1.5*IQR =15.5375<br>Q3 +1.5*IQR = 37.8575 |
| Outlier:<br>155.66<br>(Record 19: 155.66, 1.58, 22.49) | No outliers | Outlier:<br>39.22<br>(Record 22: 111.13, 1.53, 39.22) |

- Two outliers have been detected: Weight = 155.66 (record 19) and BodyMass = 39.22 (record 22)

```
1. o<-mdata$WeightKg[mdata$WeightKg>quantile(mdata$WeightKg,0.75)+1.5*IQR(mdata$WeightKg)]
2. o
[1] 155.66
```

#In line 1, outliers that lie above Q3 +1.5*IQR are detected

# English - polish dictionary

- outlier – punkt odstający/oddalony, obserwacja odstająca
- quartile – kwartyl
- percentile – percentyl
- interquartile range – rozstęp międzykwartylowy
- Z-score standardization - standaryzacja