Data Mining
Laboratory

Ewa Figielska

WARSAW SCHOOL
OF COMPUTER SCIENCE

Data mining and data mining tasks

- Data mining is the process of discovering useful patterns and trends in large data sets.

- Data mining tasks:
  - **Description** of patterns and trends lying within the data.

    E.g. a pollster may uncover evidence that those who have been laid off (and consequently are now less well off financially than before) are less likely to support the present incumbent in the presidential election and so would tend to prefer an alternative.

  - **Estimation** – approximating the value of a numeric target variable using a set of numeric and/or categorical predictor variables.

    E.g. estimating the amount of money a randomly chosen family of four will spend for back-to-school shopping this fall.

  - **Prediction** – similar to estimation, except that the results lie in future.

    E.g. predicting the price of a stock three months into the future.

Data mining tasks, cont.

- **Classification** – similar to estimation, except that the target variable is cathegorical rather than numeric.

  E.g.

  - diagnosing whether a particular disease is present,
  - determining whether a particular credit card transaction is fraudulent.

- **Clustering** - grouping of records, observations, or cases into classes of similar objects. Clustering differs from classification in that there is no target variable for clustering.

  E.g.

  - for accounting auditing purposes, to segmentize financial behavior into benign and suspicious categories,
  - as a dimension-reduction tool when the data set has hundreds of attributes.

- **Association** - finding which attributes “go together”.

  E.g.

  - examining the proportion of children whose parents read to them who are themselves good readers,
  - finding out which items in a supermarket are purchased together and which items are never purchased together.

Some definitions - **Measures of center**

Measures of center indicate where on the number line the central part of the data is located.

- **Mean** – arithmetic average of a data set.
  - **Sample mean**: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, n – the number of items in a sample
  - **Population mean**: $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$, N – the number of items in a population

- **Median** -
  - the middle data value, when there is an odd number of data values and the data have been sorted into ascending order,
  - the mean of the two middle values, if there is an even number of values.

  For (1, 1, 2, 2, 4, 6, 9), median = 2

- **Mode** – the data value that occurs with the greatest frequency.

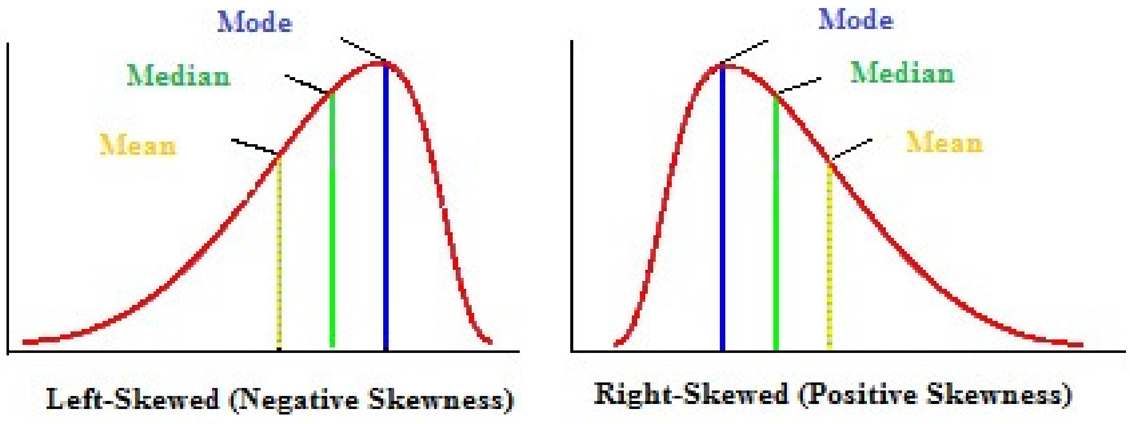
- **Midrange** – the average of the maximum and minimum values in a data set.

```
> x<-sample(1:10,10,replace = TRUE)
 > x
[1] 9 6 5 10 5 5 7 1 9 7
> mean(x)
[1] 6.4
> median(x)
[1] 6.5
> sort(x)
[1] 1 5 5 5 6 7 7 9 9 10
```

x<- ...   goes into x

Some definitions - **Skewness**

- for symmetric data, the mean and the median are approximately equal
- for right-skewed data, the mean is greater than the median
- for left-skewed data, the median is greater than the mean


Mode
Median
Mean
Left-Skewed (Negative Skewness)
Mode
Median
Mean
Right-Skewed (Positive Skewness)

Some definitions – **measures of variability**

Measures of variability quantify the amount of variation in the data.

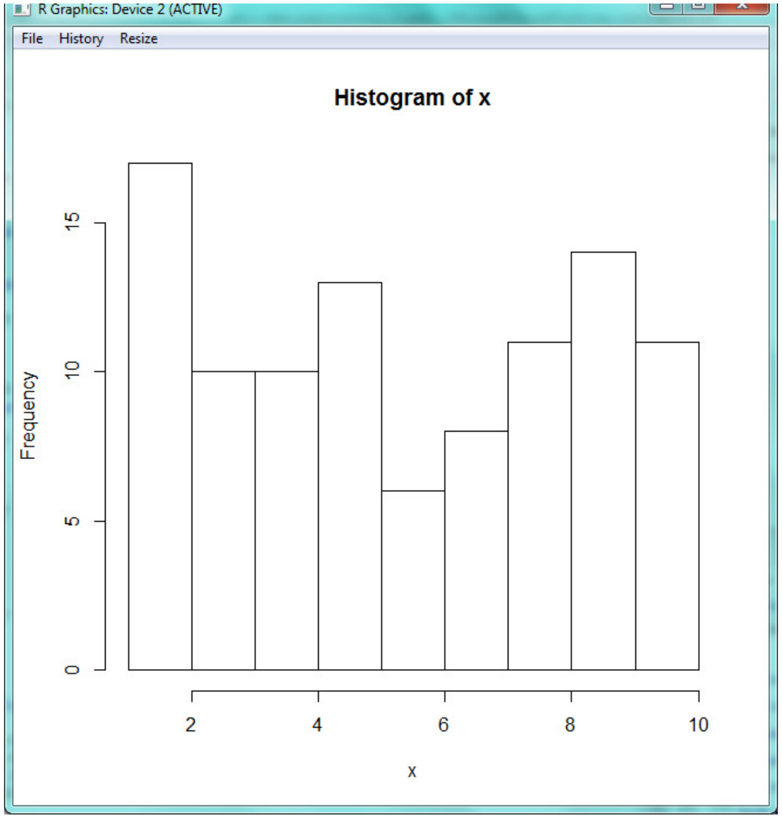
- **Range of variable** – the difference between the maximum and minimum values.
- **Deviation** – signed difference between a data value and the mean value.
- **Population variance** – mean of the squared deviations: $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$
- **Population standard deviation**: $\sigma = \sqrt{\sigma^2}$
- **Sample variance**: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
- **Sample standard deviation**: $s = \sqrt{s^2}$
- **Median absolute deviation**:
  $$MAD = median(|x - median(x)|)$$

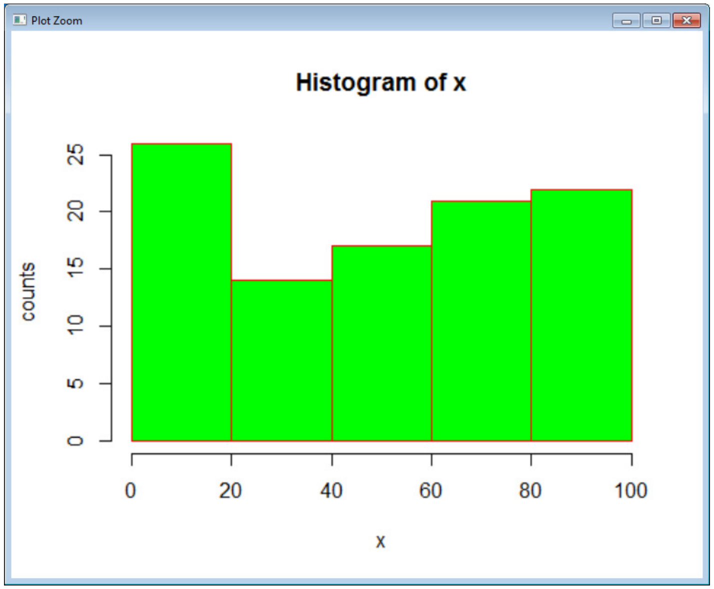
```
> x
[1] 9 6 5 10 5 5 7 1 9 7
> sd(x)
[1] 2.633122
> mad(x, constant=1)
[1] 1.5
```

Visualization of data - **histogram**

- Histogram is a graphical representation of a frequency distribution for a quantitative variable.

```
> x<-sample(1:10,100,replace = TRUE)
> x
[1] 7 4 8 1 4 10 4 7 5 1 8 7 5 5 7 6 4 9 8 3 9 8 6 9 5 5 10 7
[29] 7 7 6 2 10 9 4 10 8 10 10 9 1 10 5 4 10 3 1 1 1 9 8 5 7 3 5 8
[57] 9 2 9 8 2 6 2 1 3 1 10 1 2 8 3 9 3 3 5 9 2 5 4 5 4 9 2 3
[85] 8 6 5 10 6 4 10 5 9 9 3 1 3 8 9 4
> hist(x)
```





```
> hist(x,breaks=5,col="green",border="red",xlab="x",ylab="counts",main="Histogram of x")
```

Data preprocessing

- Much of the raw data contained in databases is incomplete and noisy. They may contain:
  - values that have expired,
  - values not consistent with common sense,
  - missing values,
  - outliers,
  - data in a form not suitable for data mining models

- To be useful for data mining purposes, the databases need to undergo preprocessing, in the form of data cleaning and data transformation.

Missing data

- Handling missing values:
  - omitting the records or fields with missing values – this may lead to a biased subset of data,
  - replacing the missing value with:
    - some constant, specified by the analyst,
    - the mean (for numerical variables) or the mode (for categorical variables),
    - a value generated at random from the observed variable distribution.

Example

- Reading file “_air1.csv”

```
> mdata <- read.csv(file="_air1.csv", header=TRUE, sep=",")
```

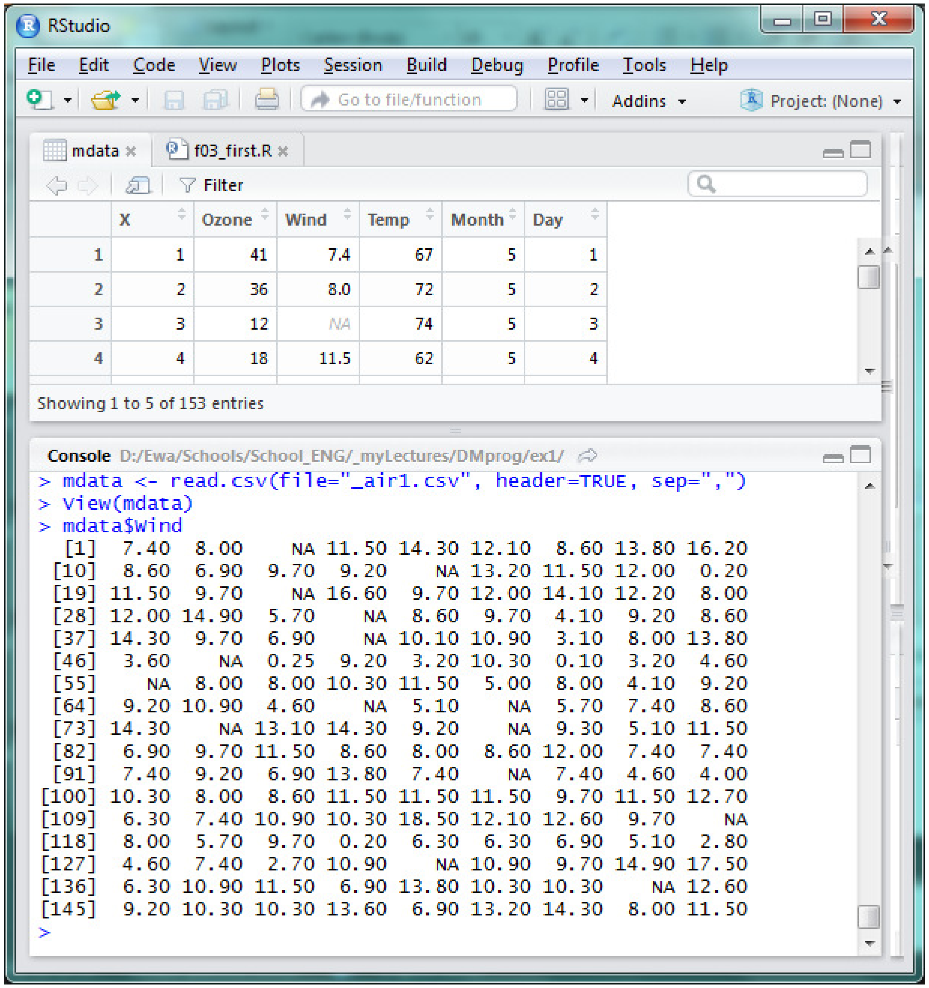
- Printing Wind data

```
> mdata$Wind
```

NA – not available

The mean and standard deviation after remowing NA values:

```
>mean(mdata$Wind, na.rm=TRUE)
[1] 9.180072
 >sd(mdata$Wind, na.rm=TRUE)
[1] 3.574063
```



	X	Ozone	Wind	Temp	Month	Day
1	1	41	7.4	67	5	1
2	2	36	8.0	72	5	2
3	3	12	NA	74	5	3
4	4	18	11.5	62	5	4

Showing 1 to 5 of 153 entries

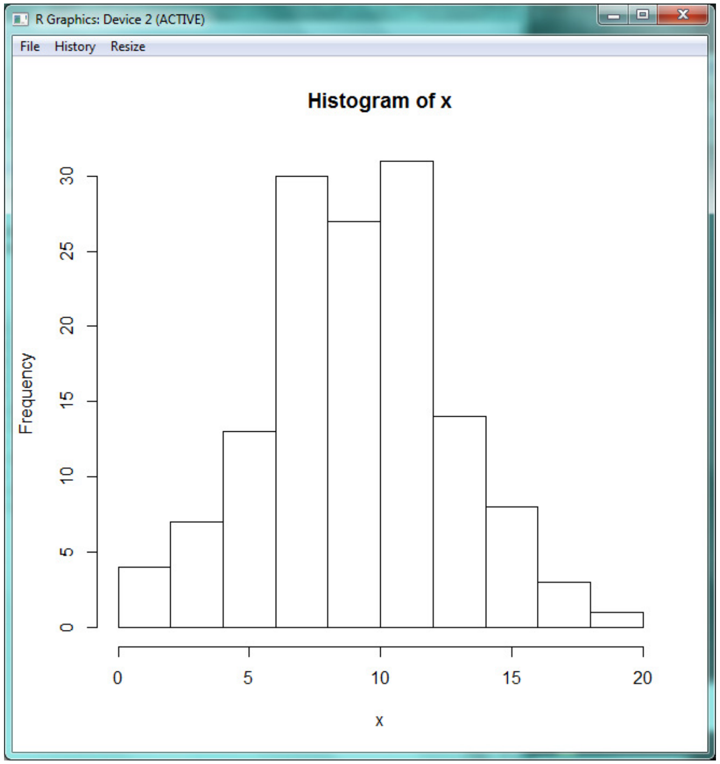
Console D:/Ewa/Schools/School_ENG/_myLectures/DMprog/ex1/

```
> mdata <- read.csv(file="_air1.csv", header=TRUE, sep=",")
> View(mdata)
> mdata$Wind
  [1]  7.40  8.00    NA 11.50 14.30 12.10  8.60 13.80 16.20
 [10]  8.60  6.90  9.70  9.20    NA 13.20 11.50 12.00  0.20
 [19] 11.50  9.70    NA 16.60  9.70 12.00 14.10 12.20  8.00
 [28] 12.00 14.90  5.70    NA  8.60  9.70  4.10  9.20  8.60
 [37] 14.30  9.70  6.90    NA 10.10 10.90  3.10  8.00 13.80
 [46]  3.60    NA  0.25  9.20  3.20 10.30  0.10  3.20  4.60
 [55]    NA  8.00  8.00 10.30 11.50  5.00  8.00  4.10  9.20
 [64]  9.20 10.90  4.60    NA  5.10    NA  5.70  7.40  8.60
 [73] 14.30    NA 13.10 14.30  9.20    NA  9.30  5.10 11.50
 [82]  6.90  9.70 11.50  8.60  8.00  8.60 12.00  7.40  7.40
 [91]  7.40  9.20  6.90 13.80  7.40    NA  7.40  4.60  4.00
[100] 10.30  8.00  8.60 11.50 11.50 11.50  9.70 11.50 12.70
[109]  6.30  7.40 10.90 10.30 18.50 12.10 12.60  9.70    NA
[118]  8.00  5.70  9.70  0.20  6.30  6.30  6.90  5.10  2.80
[127]  4.60  7.40  2.70 10.90    NA 10.90  9.70 14.90 17.50
[136]  6.30 10.90 11.50  6.90 13.80 10.30 10.30    NA 12.60
[145]  9.20 10.30 10.30 13.60  6.90 13.20 14.30  8.00 11.50
>
```

Setting the working Directory in R-Studio: Session/Set Working Directory

Drawing a histogram

- Historam of Wind after remowing NA values



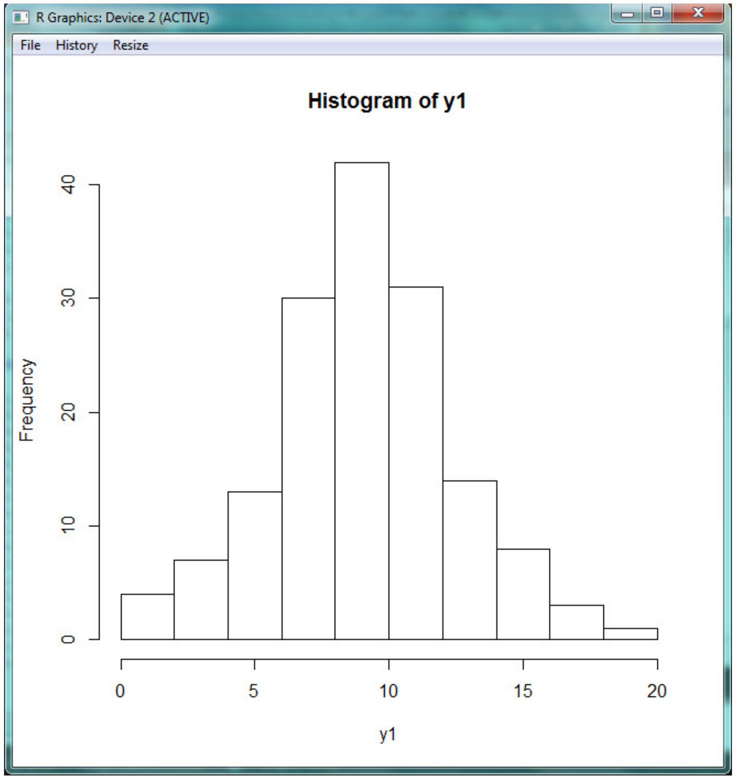
```
> mdata <- read.csv(file="_air1.csv", header=TRUE, sep=",")
> x<-na.omit(mdata$Wind)
> hist(x)
```

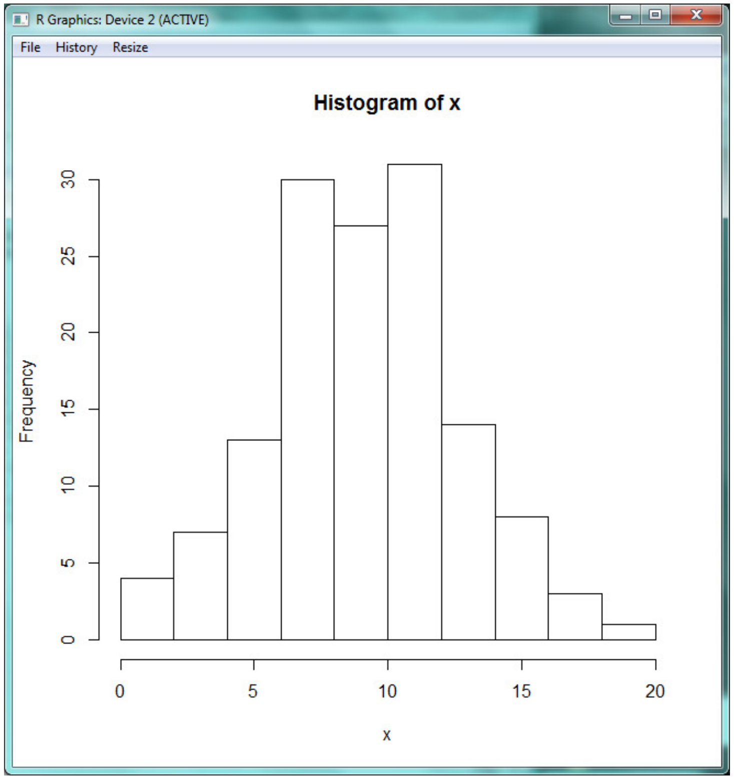
Replacing the missing values with the mean

```
> mdata <- read.csv(file="_air1.csv", header=TRUE, sep=",")
> y<-mdata$Wind
> y1 <- ifelse(is.na(y), mean(y, na.rm=TRUE), y)
[1] 7.400000 8.000000 9.180072 11.500000 14.300000 12.100 ...
```

After replacement:

```
> mean(y1)
[1] 9.180072
> sd(y1)
[1] 3.393131
```





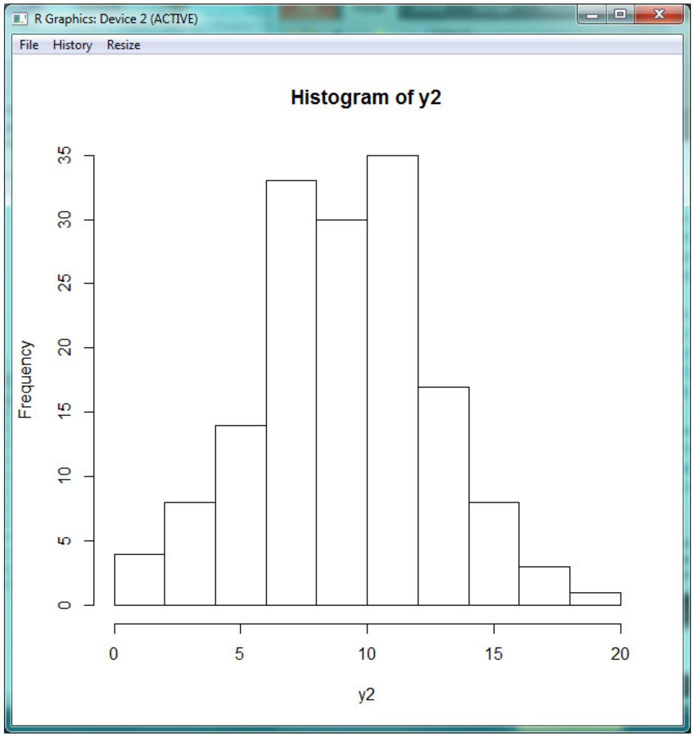
y1 contains Wind data with NA replaced by mean

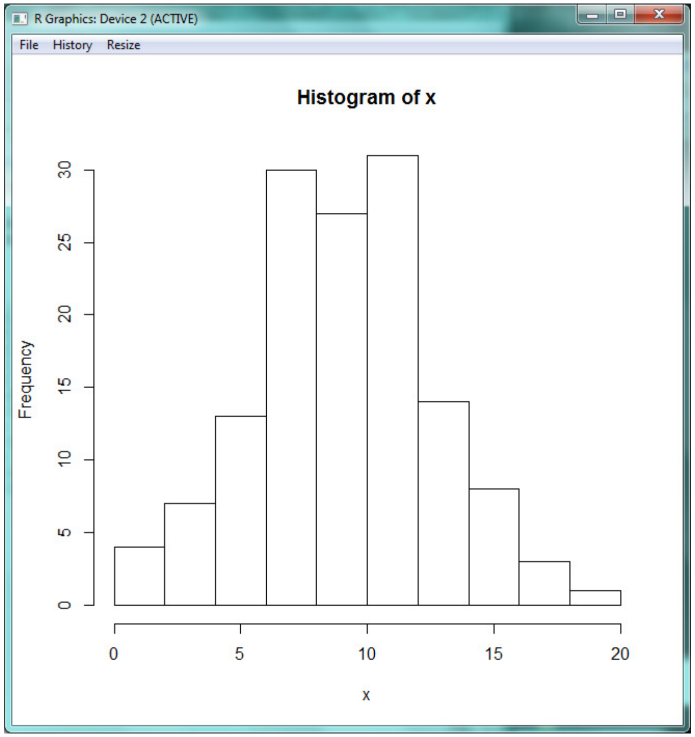
Replacing missing entries with values generated at random from the observed variable distribution

```
> mdata <- read.csv(file="_air1.csv", header=TRUE, sep=",")
> y<-mdata$Wind
> y2<-y                                    #alternatively y2 <- rep(NA, length(y))
> for(i in seq_along(y)){  y2[i]<- ifelse(is.na(y[i]), sample(na.omit(y),1), y[i]) }
> y2
[1] 7.40   8.00   12.60   11.50  14.30  12.10   8.60  13.80  16.20  8.60   6.90   9.70   9.20  13.60
```

After replacement:

```
> mean(y2)
[1] 9.225817
> sd(y2)
[1] 3.515277
```





y2 contains Wind data with NA replaced by values generated at random from the observed variable distribution

English - polish dictionary

- Measures of center – miary tendencji centralnej
- Quantitative and categorical variable – ilościowe i jakościowe (nominalne) zmienne
- Skewness - skośność