SAPIENZA
UNIVERSITÀ DI ROMA

UNIVERSITÀ
DEGLI STUDI DELLA
TUSCIA

Statistics for business and decision making Course

Dr. Ilaria Benedetti

Course Introduction

Introduction

In this session we will discuss approaches for summarising the distribution of the data.

Why is this important?

This is important because we would like to summarise a large number of data using only a couple of numbers.

These numerical measures provide information about a “typical” observation in the data and are referred to as measures of central tendency.

In this section, we present numerical measures — **the mean, median, and mode**—In response to questions concerning the location of the center of a data set.

One measure of central tendency that quickly comes to mind is the **arithmetic mean**, usually just called the mean, or average.

The arithmetic mean (or simply mean) of a set of data is the sum of the data values divided by the number of observations. If the data set is the entire population of data, then the population mean, m, is a parameter given by

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

where N = population size and Σ means “the sum of.”
If the data set is from a sample, then the sample mean, x, is a statistic given by

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where n = sample size.
The mean is appropriate for numerical data.

EXAMPLE 1 Mean

We have the following 15 observations

35,36,37,40,42,42,42,45,45,46,47,49,49,51,51

The mean is calculated is as follows

Mean= $\frac{35+36+37+\cdots+51}{15} = 43.8$

Mode

Mode

The mode, if one exists, is the most frequently occurring value. A distribution with one mode is called unimodal; with two modes, it is called bimodal; and with more than two modes, the distribution is said to be multimodal.

The mode is most commonly used with categorical data.

In a frequency distribution (with classes of different size) the modal class is the class with the highest frequency density.

Example: Suppose we have the following 10 observations

35	35	37	40	42	42	42	45	45	46

Xi	ni
35	2
37	1
40	1
42	3
45	2
46	1

Mode? = 42

EXAMPLE 2 mode for frequency distribution

n. employees (xj)	number of sales points (freq. nj)
3	2
4	1
6	3
7	1
10	2
	9

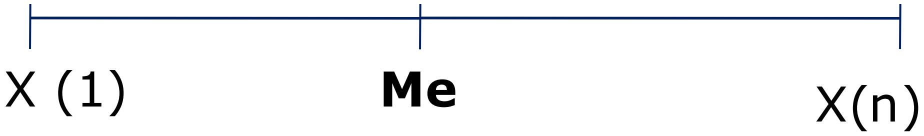
Mode? **6 employees**

Median

The mean can be misleading when a data set has an outlier because the mean will shift toward the value of that outlier. For this reason, there is a need for alternative measures of central tendency.

The **median** is the middle value in a distribution of data listed in non-decreasing (numeric) order.

Consequently, the median is the value at which the distribution of our data is divided into two equal parts.
The median is the second quartile (q2) of a distribution.



How to calculate the median?

Median

- For a **set of observations** placed in numeric order (ranked according to their magnitude):

- n ODD: the median position is of order: $\frac{(n+1)}{2}$

$$Me = x_{\left(\frac{n+1}{2}\right)}$$

- n EVEN: there are two median position of order: $\frac{n}{2}$ and $\frac{n}{2}+1$

$$x_{\left(\frac{n}{2}\right)} \le Me \le x_{\left(\frac{n}{2}+1\right)}$$

then: $$Me = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$$

Median

An example:

We have the following 15 observations

35,36,37,40,42,42,42,45,45,46,47,49,49,51,51

Median? **Median is 45**

n is ODD: the median position is of order: $\frac{(n+1)}{2}$

$$Me = x_{\left(\frac{n+1}{2}\right)}$$

Mean, Mode, Median: how to choose

Guidelines for Choosing a Measure of Central Tendency

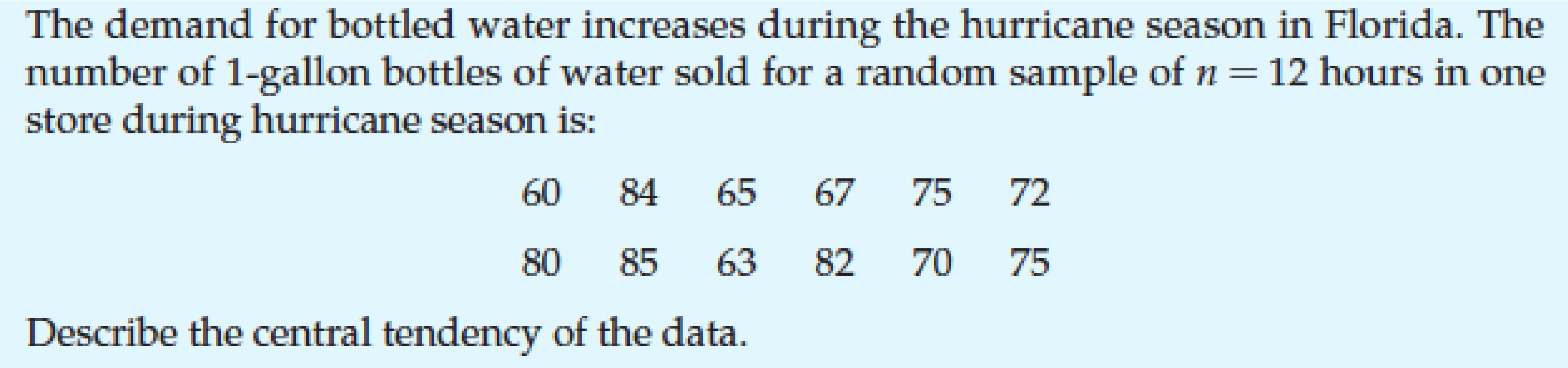
The mean is more commonly used. However, the mean is affected by extreme values (also known as outliers) …

Example:

Think of the variable income and assume that we have two multimillionaires in our dataset. What is the effect of these two individuals on the mean income?

- ✓ The mean is affected by outliers.
- ✓ On the other hand the median is not affected by extreme values.
- ✓ Median is a robust (to outliers) measure of central tendency
- ✓ The mean is easier to work with than the median
- ✓ The mode is used less often

EXAMPLE

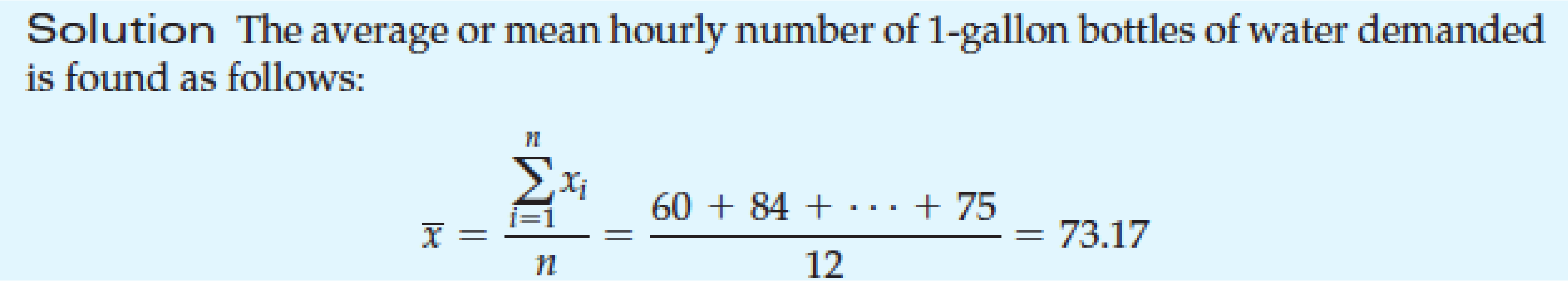


The demand for bottled water increases during the hurricane season in Florida. The number of 1-gallon bottles of water sold for a random sample of $n = 12$ hours in one store during hurricane season is:

60 84 65 67 75 72

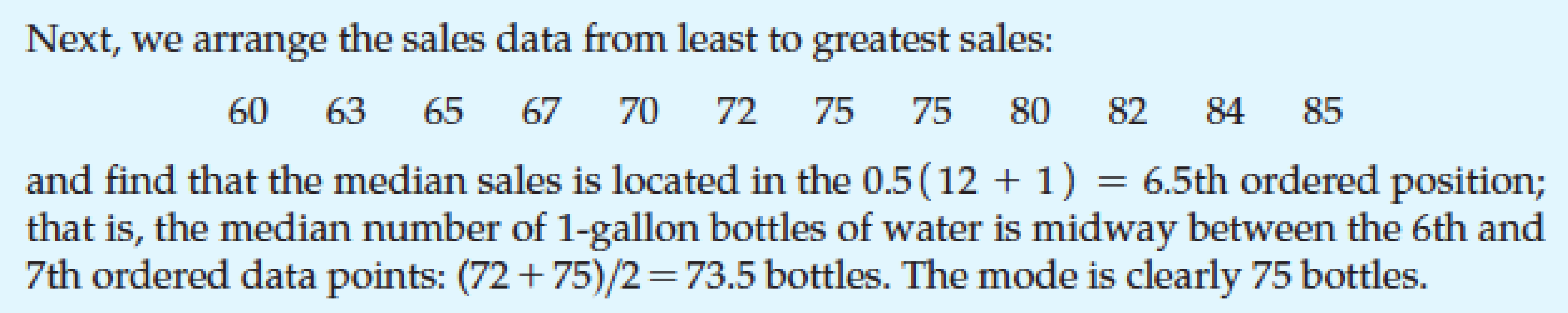
80 85 63 82 70 75

Describe the central tendency of the data.



Solution The average or mean hourly number of 1-gallon bottles of water demanded is found as follows:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{60 + 84 + \cdots + 75}{12} = 73.17$$



Next, we arrange the sales data from least to greatest sales:

60 63 65 67 70 72 75 75 80 82 84 85

and find that the median sales is located in the $0.5(12 + 1) = 6.5$th ordered position; that is, the median number of 1-gallon bottles of water is midway between the 6th and 7th ordered data points: $(72 + 75)/2 = 73.5$ bottles. The mode is clearly 75 bottles.

Percentiles and Quartiles

To find percentiles and quartiles, data must first be arranged in **order from** the smallest to the largest values.

The P_{th} percentile is a value such that approximately *P%* of the observations are at or below that number.

Percentiles separate large ordered data sets into 100ths.

The P_{th} percentile is found as follows:

P_{th} percentile = value located in the *(P/100) (n+1) th-ordered position*.

Quartiles are descriptive measures that separate large data sets into four quarters.

- The **first quartile**, Q1, (or 25th percentile) separates approximately the smallest 25% of the data from the remainder of the data;
- The **second quartile**, Q2, (or 50th percentile) is the median;
- The **third quartile**, Q3, (or 75th percentile), separates approximately the smallest 75% of the data from the remaining largest 25% of the data.

Q1 = the value in the 0.25(n+1)th ordered position
Q2 = the value in the 0.50(n+1)th ordered position
Q3 = the value in the 0.75(n+1)th ordered position

In describing numerical data, we often refer to the five-number summary.

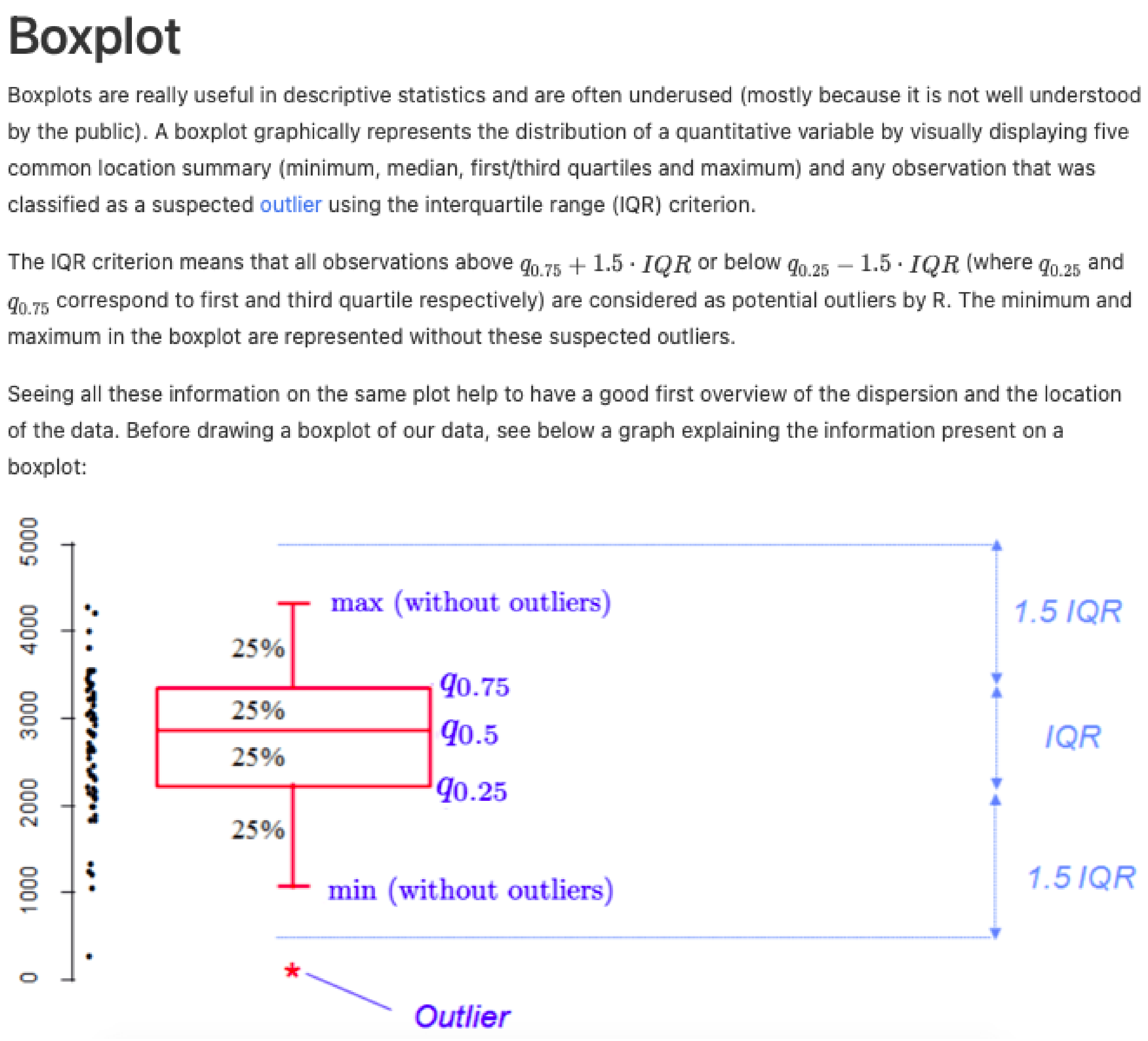
The **five-number summary** refers to the five descriptive measures: minimum, first quartile, median, third quartile, and maximum:

minimum < Q1 < median < Q3 < maximum

We present a graph of the five-number summary called a box-and-whisker plot.

A **box-and-whisker plot** is a graph that describes the shape of a distribution in terms of the five-number summary:

- the minimum value,
- first quartile (25th percentile),
- the median,
- the third quartile (75th percentile),
- the maximum value

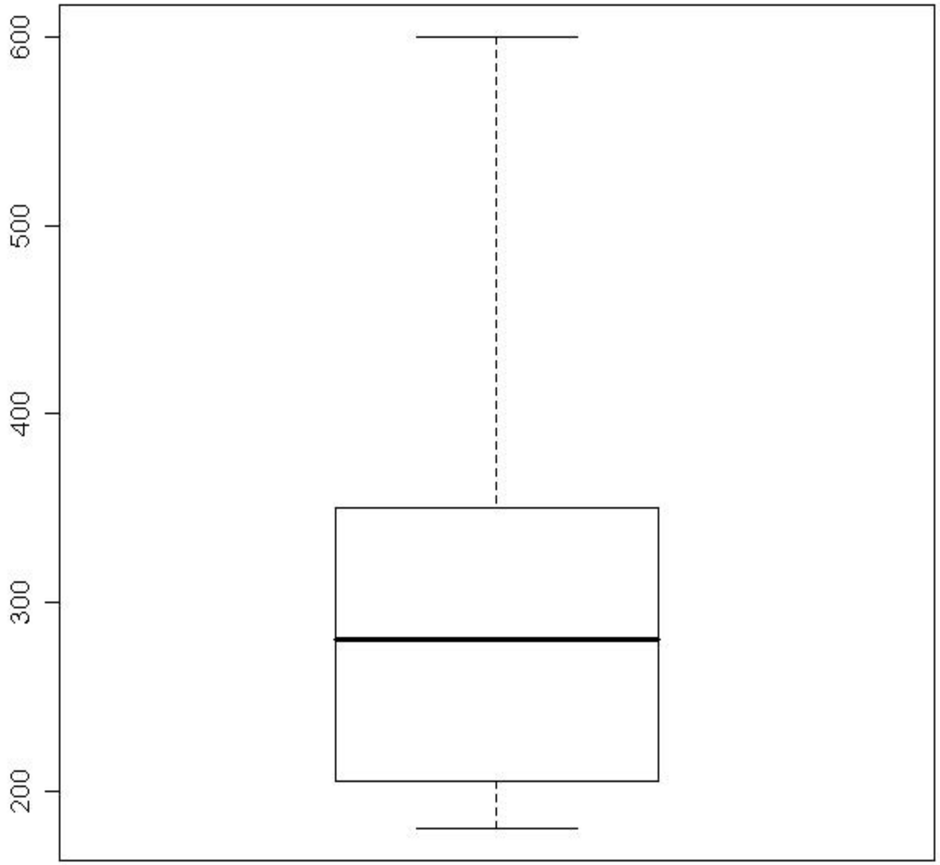


Boxplot

Boxplots are really useful in descriptive statistics and are often underused (mostly because it is not well understood by the public). A boxplot graphically represents the distribution of a quantitative variable by visually displaying five common location summary (minimum, median, first/third quartiles and maximum) and any observation that was classified as a suspected outlier using the interquartile range (IQR) criterion.

The IQR criterion means that all observations above $q_{0.75} + 1.5 \cdot IQR$ or below $q_{0.25} - 1.5 \cdot IQR$ (where $q_{0.25}$ and $q_{0.75}$ correspond to first and third quartile respectively) are considered as potential outliers by R. The minimum and maximum in the boxplot are represented without these suspected outliers.

Seeing all these information on the same plot help to have a good first overview of the dispersion and the location of the data. Before drawing a boxplot of our data, see below a graph explaining the information present on a boxplot:

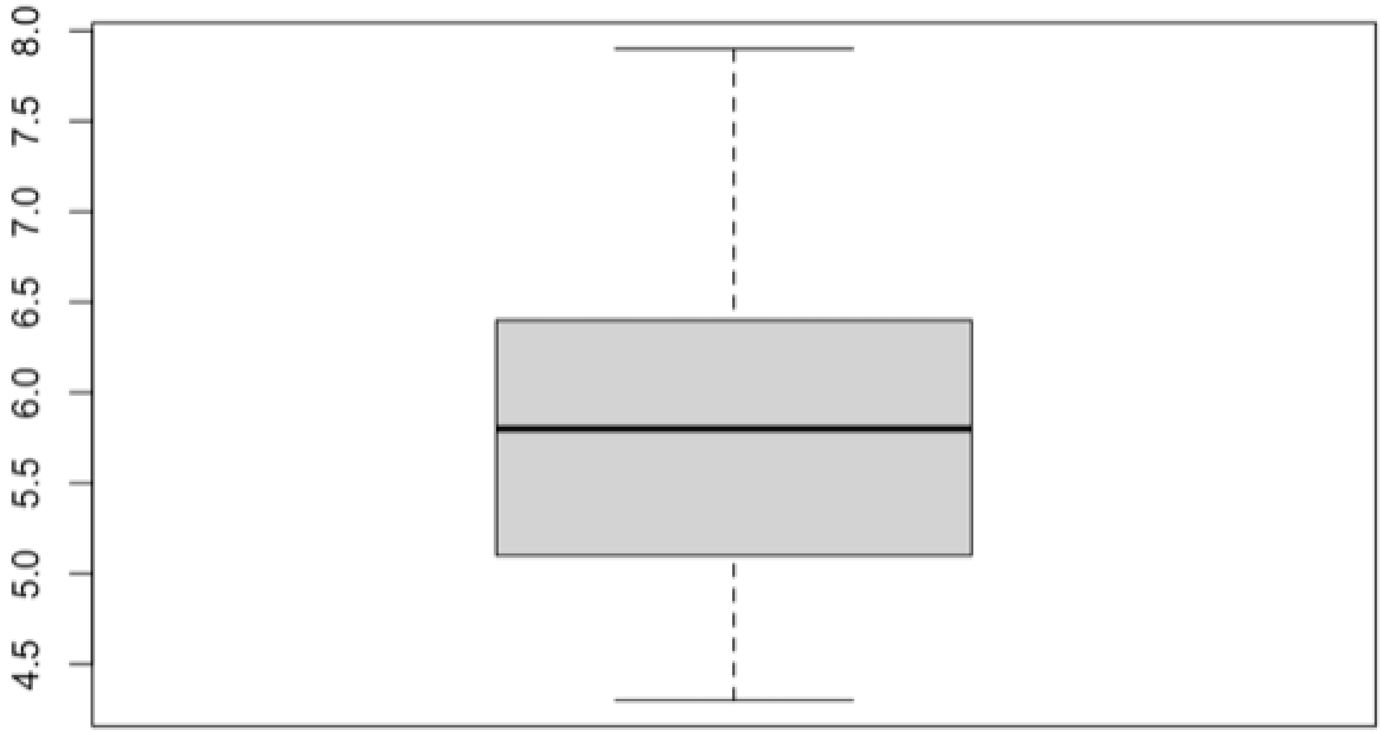


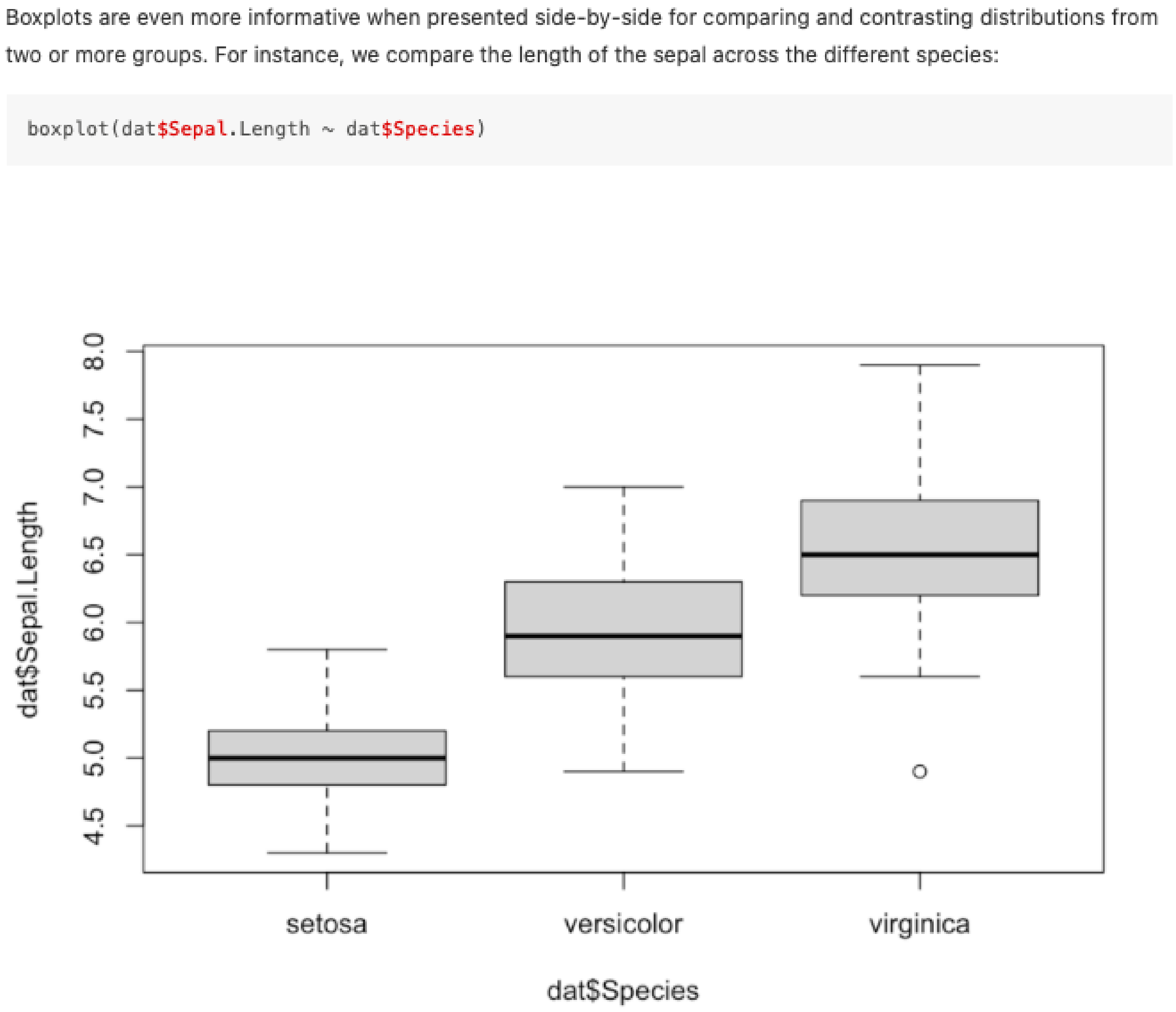
The inner box shows the numbers that span the range from the first to the third quartile.

A line is drawn through the box at the median.

There are two “whiskers.” One whisker is the line from the 25th percentile to the minimum value; the other whisker is the line from the 75th percentile to the maximum value

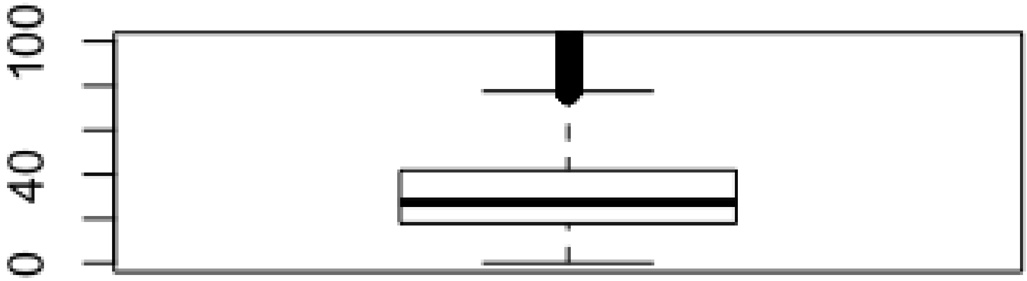
```
boxplot(dat$Sepal.Length)
```



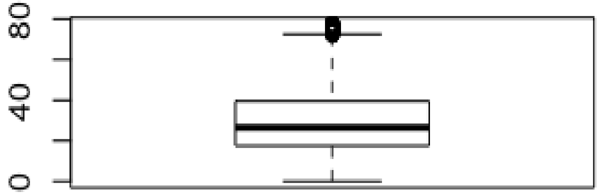


Boxplots are even more informative when presented side-by-side for comparing and contrasting distributions from two or more groups. For instance, we compare the length of the sepal across the different species:

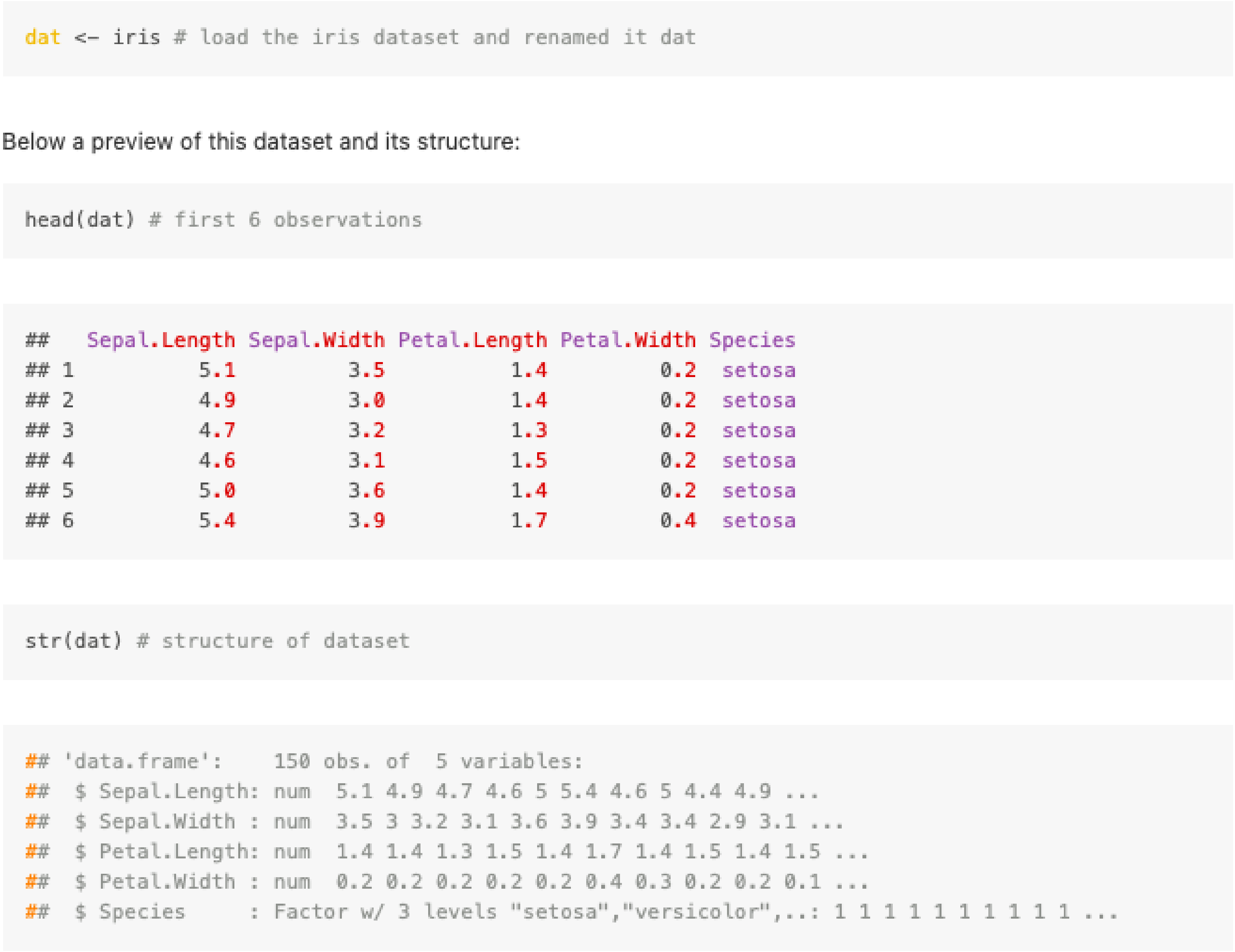
```
boxplot(dat$Sepal.Length ~ dat$Species)
```



```
# identification of outliers Q1=first quintile, Q3=third quintile
# x>Q3+lambda(Q3-Q1)
# x<Q1-lambda(Q3-Q1)

# lambda is a positive constant variable equal to 1.5

Q3=quantile(your_variable, 0.75 ) # define 3rd quintile
Q1=quantile(your_variable, 0.25 ) # define 1st quintile
x_upper<-Q3+1.5*(Q3-Q1)    # upper bound
x_lower<-Q1-1.5*(Q3-Q1)    # lower bound



tousand[tousand>77.58]=NA
summary(tousand) # now there are 348 NA's

boxplot(tousand)
```



To perform the analysis, please load into R the dataset called “iris”



```
dat <- iris # load the iris dataset and renamed it dat
```

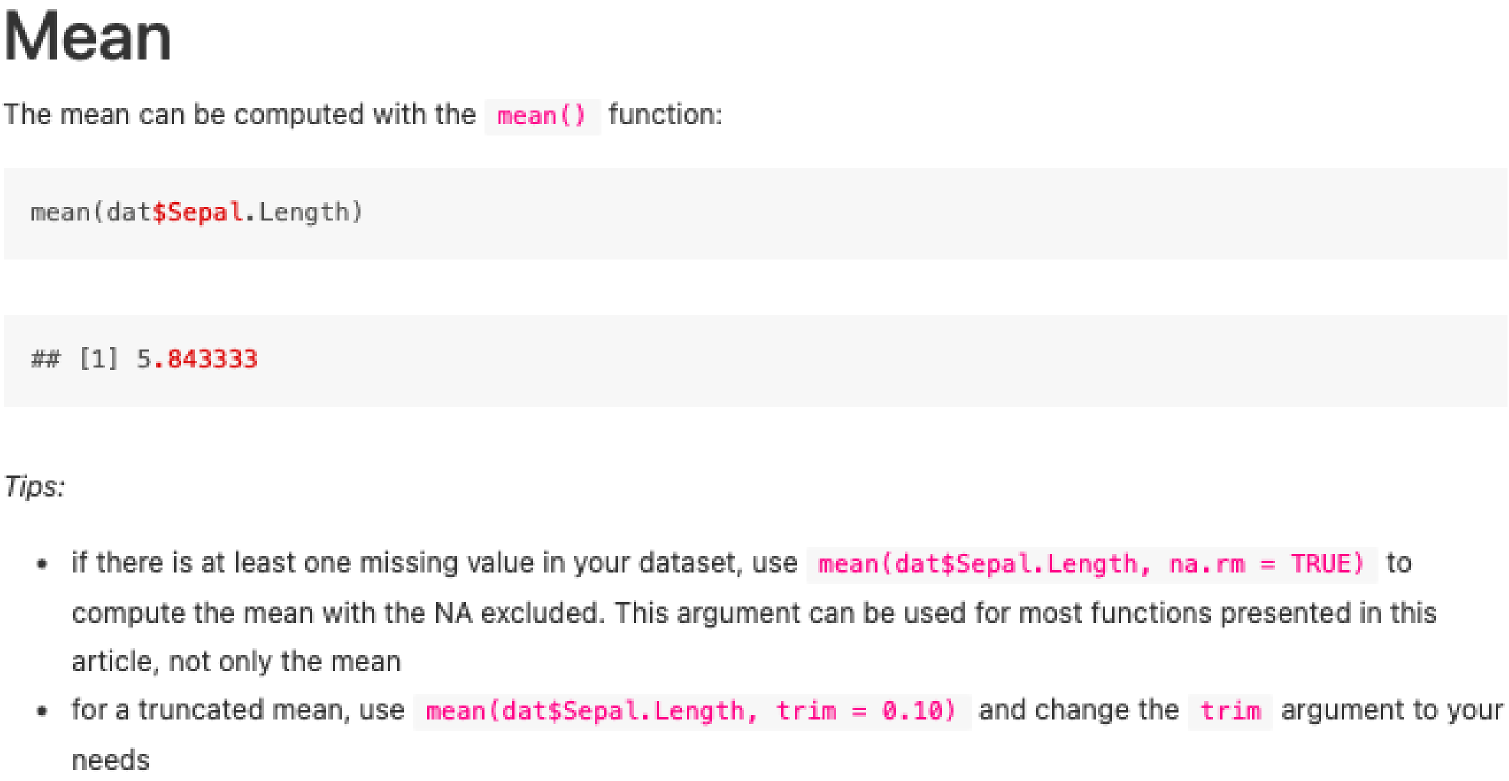
Below a preview of this dataset and its structure:

```
head(dat) # first 6 observations
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
str(dat) # structure of dataset
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```



Mean

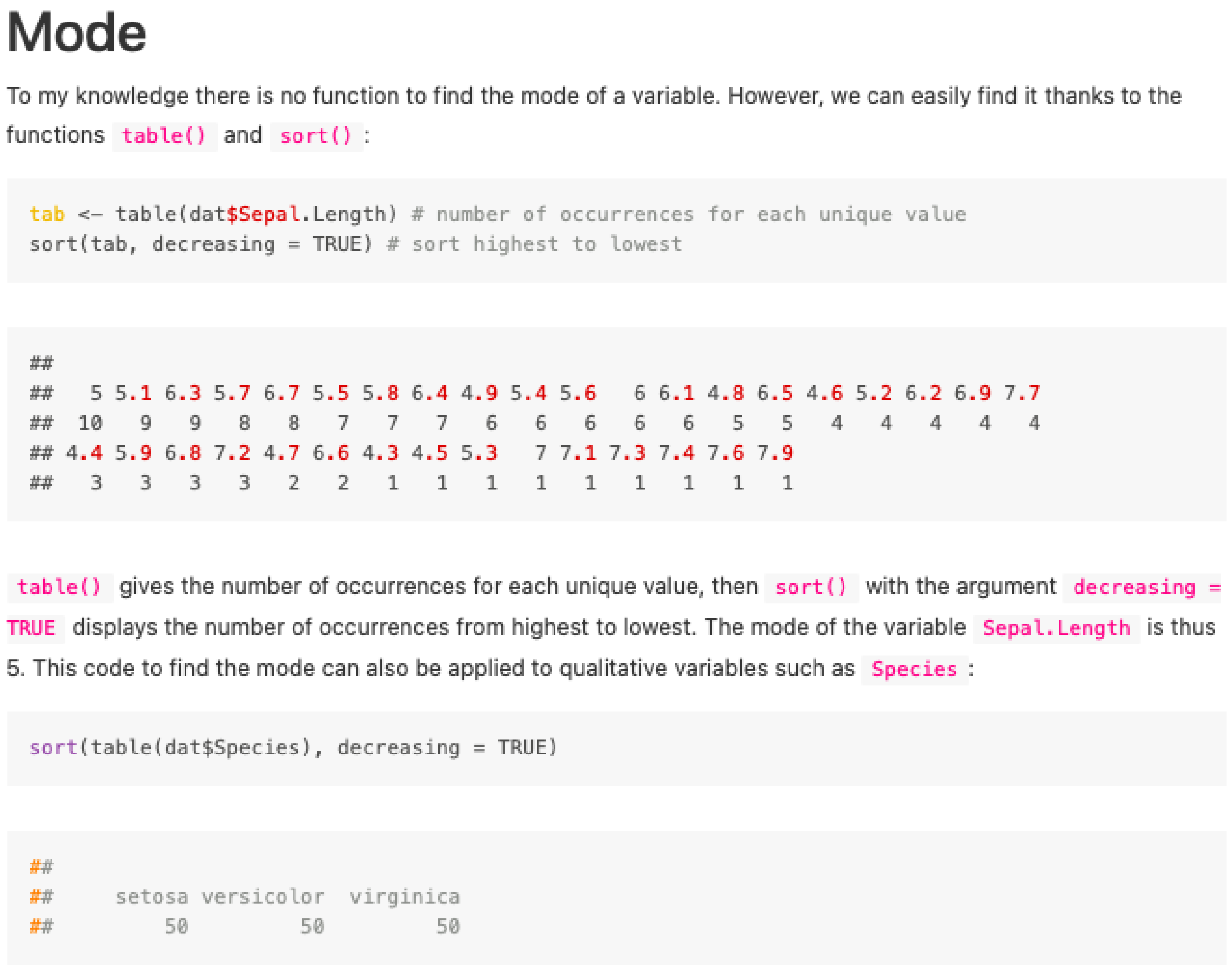
The mean can be computed with the `mean()` function:

```
mean(dat$Sepal.Length)
```

```
## [1] 5.843333
```

Tips:

- if there is at least one missing value in your dataset, use `mean(dat$Sepal.Length, na.rm = TRUE)` to compute the mean with the NA excluded. This argument can be used for most functions presented in this article, not only the mean
- for a truncated mean, use `mean(dat$Sepal.Length, trim = 0.10)` and change the `trim` argument to your needs



Mode

To my knowledge there is no function to find the mode of a variable. However, we can easily find it thanks to the functions `table()` and `sort()` :

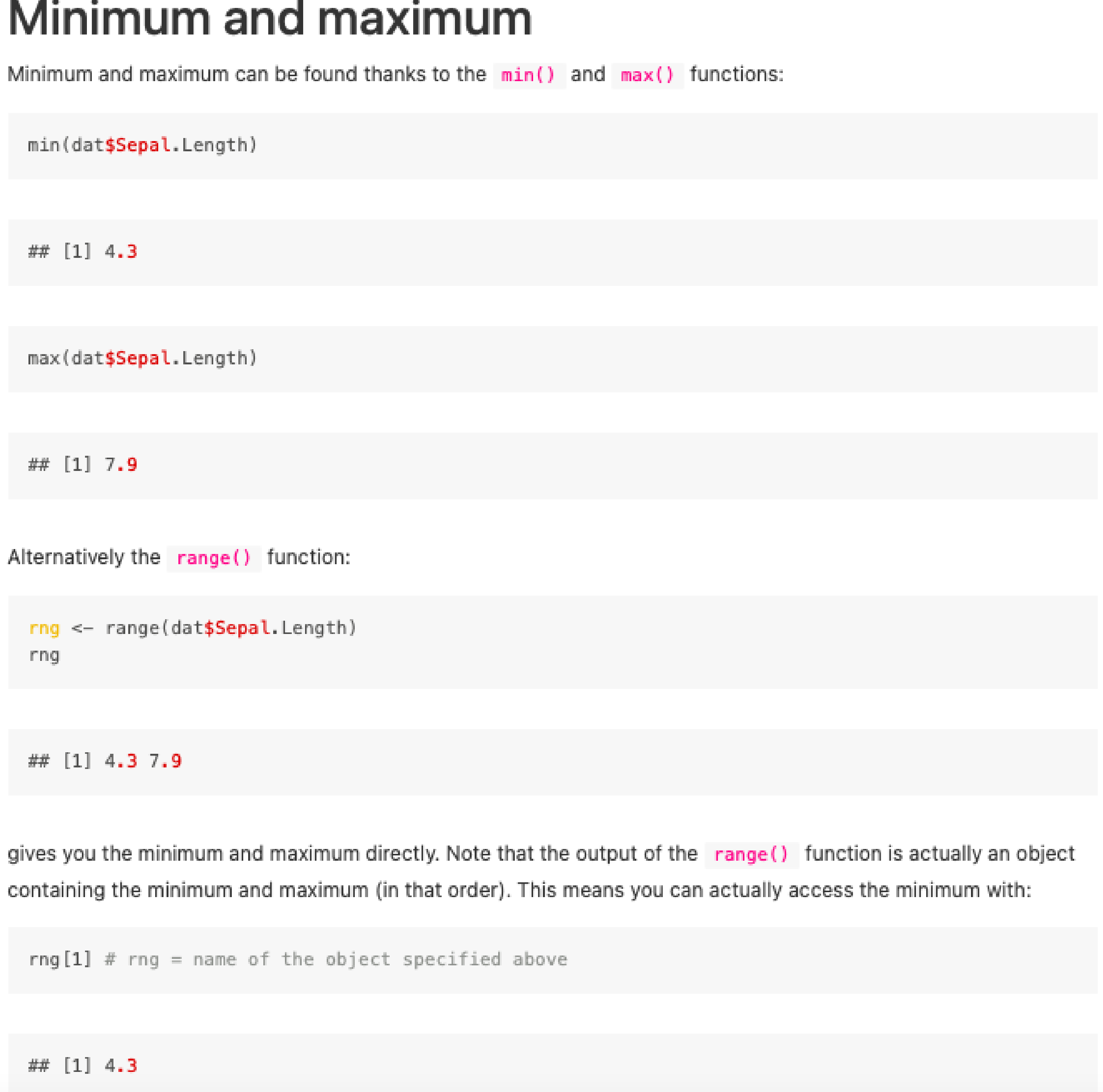
```
tab <- table(dat$Sepal.Length) # number of occurrences for each unique value
sort(tab, decreasing = TRUE) # sort highest to lowest
```

```
## 
##   5 5.1 6.3 5.7 6.7 5.5 5.8 6.4 4.9 5.4 5.6   6 6.1 4.8 6.5 4.6 5.2 6.2 6.9 7.7 
##  10   9   9   8   8   7   7   7   6   6   6   6   6   5   5   4   4   4   4   4 
## 4.4 5.9 6.8 7.2 4.7 6.6 4.3 4.5 5.3   7 7.1 7.3 7.4 7.6 7.9 
##   3   3   3   3   2   2   1   1   1   1   1   1   1   1   1
```

`table()` gives the number of occurrences for each unique value, then `sort()` with the argument `decreasing = TRUE` displays the number of occurrences from highest to lowest. The mode of the variable `Sepal.Length` is thus 5. This code to find the mode can also be applied to qualitative variables such as `Species` :

```
sort(table(dat$Species), decreasing = TRUE)
```

```
## 
##     setosa versicolor  virginica 
##         50         50         50
```



Minimum and maximum

Minimum and maximum can be found thanks to the `min()` and `max()` functions:

```
min(dat$Sepal.Length)
```

```
## [1] 4.3
```

```
max(dat$Sepal.Length)
```

```
## [1] 7.9
```

Alternatively the `range()` function:

```
rng <- range(dat$Sepal.Length)
rng
```

```
## [1] 4.3 7.9
```

gives you the minimum and maximum directly. Note that the output of the `range()` function is actually an object containing the minimum and maximum (in that order). This means you can actually access the minimum with:

```
rng[1] # rng = name of the object specified above
```

```
## [1] 4.3
```

Median

The median can be computed thanks to the `median()` function:

```
median(dat$Sepal.Length)
```

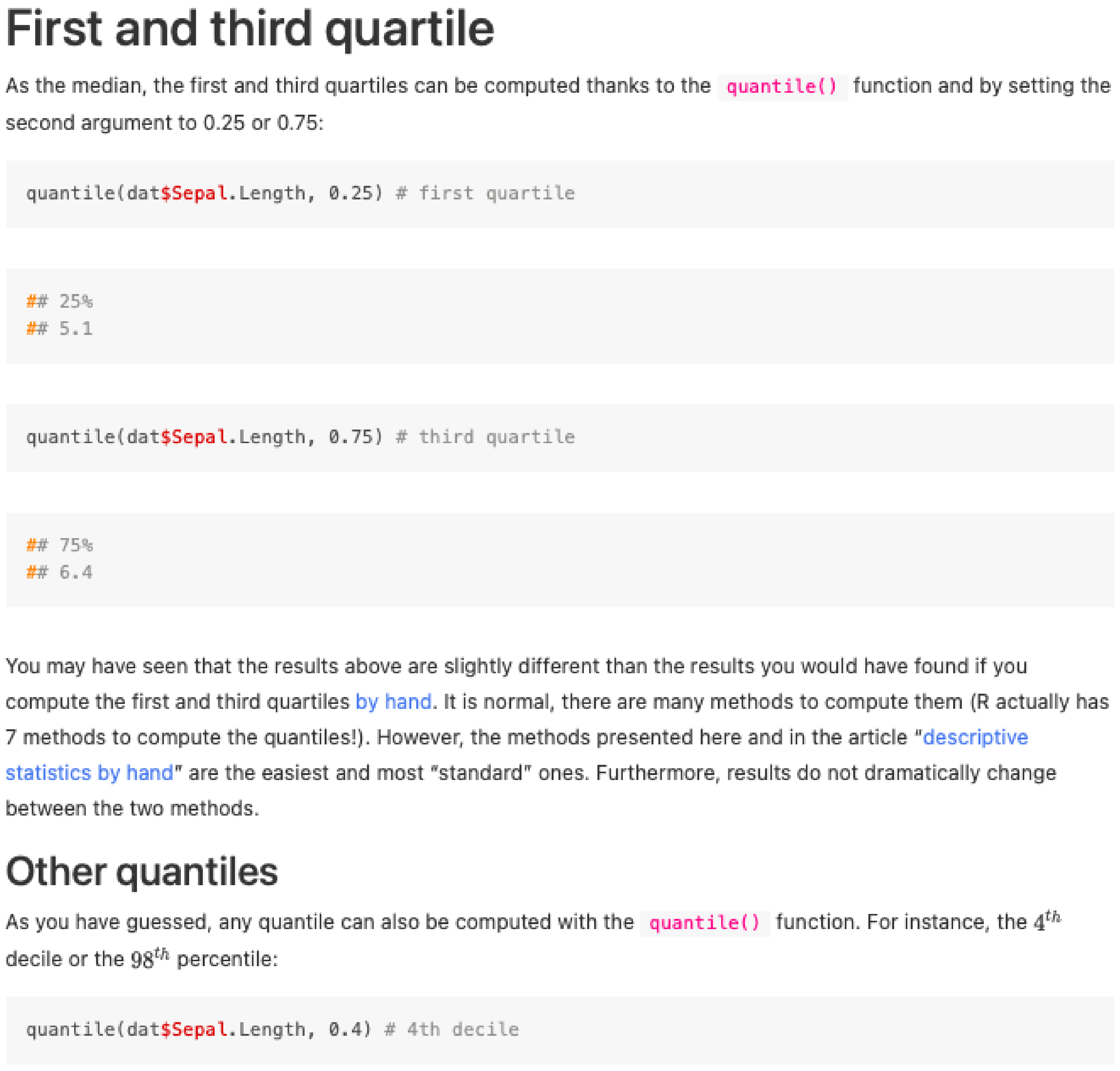
```
## [1] 5.8
```

or with the `quantile()` function:

```
quantile(dat$Sepal.Length, 0.5)
```

```
## 50%
## 5.8
```

since the quantile of order 0.5 ($q_{0.5}$) corresponds to the median.



First and third quartile

As the median, the first and third quartiles can be computed thanks to the `quantile()` function and by setting the second argument to 0.25 or 0.75:

```
quantile(dat$Sepal.Length, 0.25) # first quartile
```

```
## 25%
## 5.1
```

```
quantile(dat$Sepal.Length, 0.75) # third quartile
```

```
## 75%
## 6.4
```

You may have seen that the results above are slightly different than the results you would have found if you compute the first and third quartiles by hand. It is normal, there are many methods to compute them (R actually has 7 methods to compute the quantiles!). However, the methods presented here and in the article "descriptive statistics by hand" are the easiest and most "standard" ones. Furthermore, results do not dramatically change between the two methods.

Other quantiles

As you have guessed, any quantile can also be computed with the `quantile()` function. For instance, the 4^{th} decile or the 98^{th} percentile:

```
quantile(dat$Sepal.Length, 0.4) # 4th decile
```

From central tendency …
to variability measures

The mean alone does not provide a complete or sufficient description of data. Variation exists in all areas…

- ✓ In sports, the star basketball player might score five 3-pointers in one game and none in the next or play 40 minutes in one game and only 24 minutes in the next.
- ✓ The weather varies greatly from day to day and even from hour to hour;
- ✓ grades on a test differ for students taking the same course with the same instructor;
- ✓ a person’s blood pressure, pulse, cholesterol level, and caloric intake will vary daily.
- ✓ In business, variation is seen in sales, advertising costs, the percentage of product complaints, the number of new customers, and so forth. …

While two data sets could have the same mean, the individual observations in one set could vary more from the mean than do the observations in the second set.

Consider the following two sets of sample data:

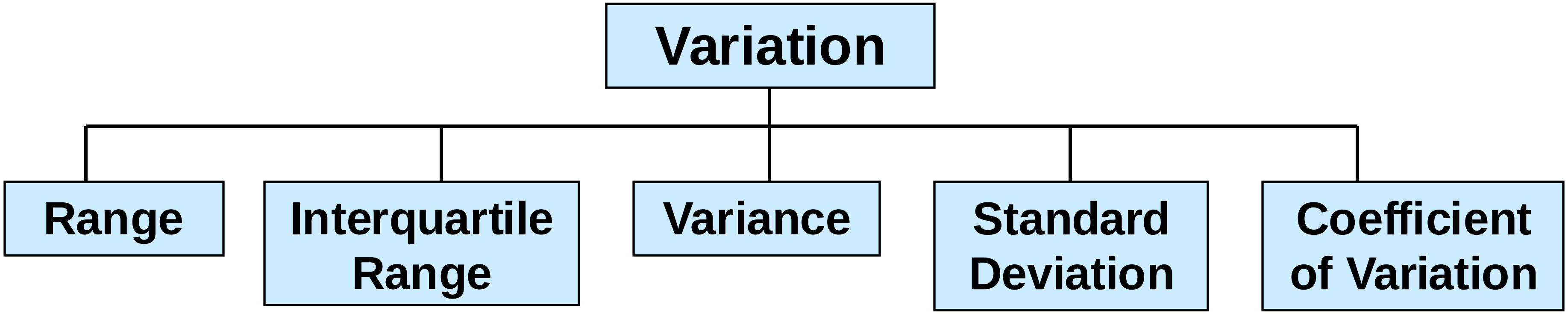
Sample A: 1 - 2 - 1 - 36
Sample B: 8 - 9 - 10 -13

Although the mean is 10 for both samples, clearly the data in sample A are farther from 10 than are the data in sample B.

We need descriptive numbers to measure this spread …

In this section we present descriptive numbers that measure the variability or spread of the observations from the mean. In particular, we include the **range, interquartile range, variance, standard deviation, and coefficient of variation**.

Measures of Variability



Variation

Range

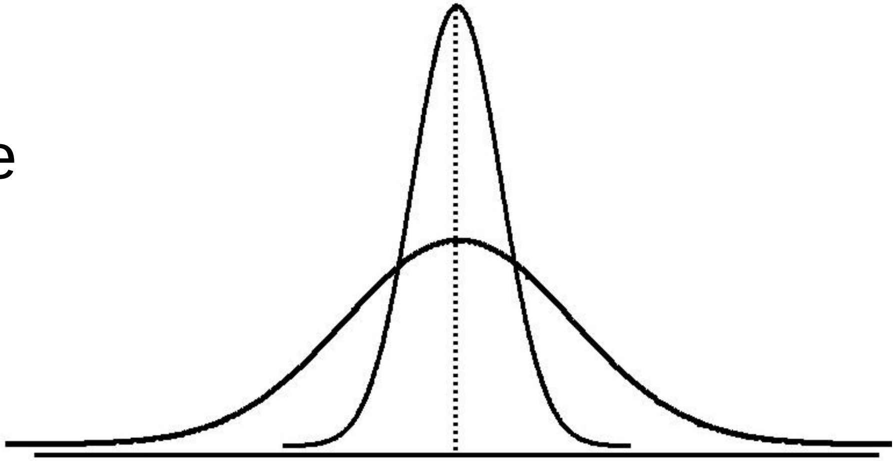
Interquartile Range

Variance

Standard Deviation

Coefficient of Variation

- Measures of variation give information on the spread or variability of the data values.



Same center, different variation

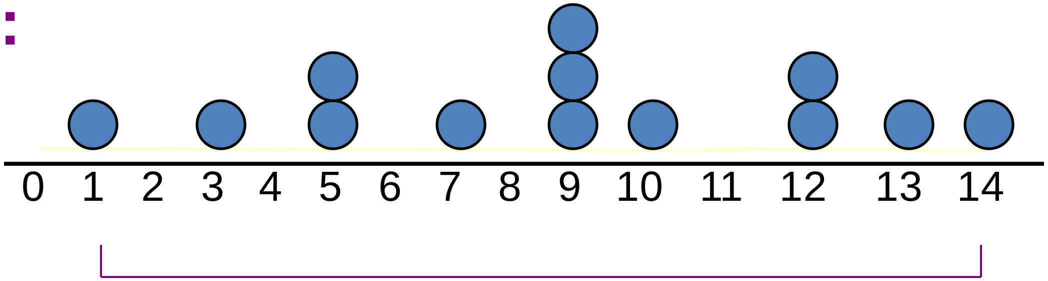
Range

The **range** is the difference between the largest and smallest observations in set of observations.

Simplest measure of variation

$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

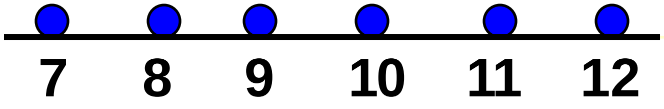
Example:



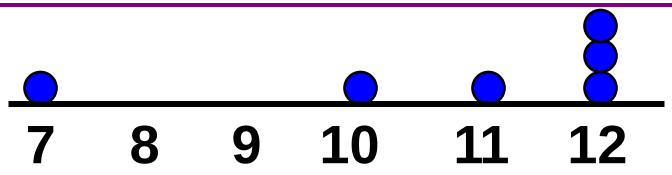
Range = 14 - 1 = 13

Disadvantages of the Range

Ignores the way in which data are distributed



Range = 12 - 7 = 5



Range = 12 - 7 = 5

Sensitive to outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5

Range = 5 - 1 = 4

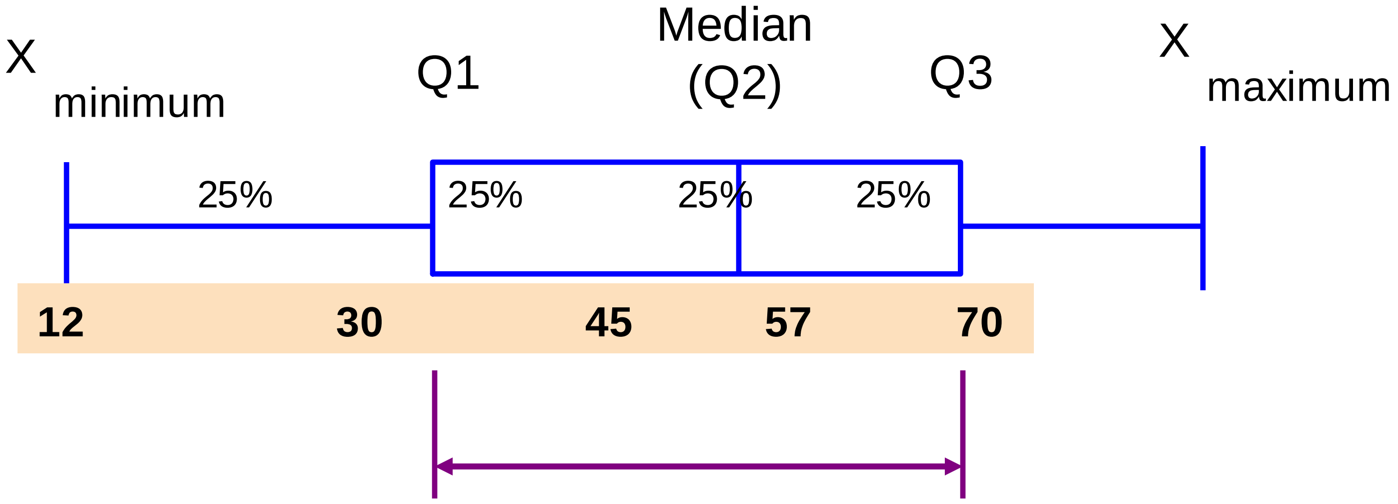
1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

Range = 120 - 1 = 119

Interquartile Range

- Can eliminate some outlier problems by using the **interquartile range**
- Eliminate high- and low-valued observations and calculate the range of the middle 50% of the data
- Interquartile range = 3rd quartile – 1st quartile

$$IQR = Q_3 - Q_1$$



$X_{minimum}$

Q1

Median (Q2)

Q3

$X_{maximum}$

25% 25% 25% 25%

12 30 45 57 70

Interquartile range
= 57 – 30 = 27

Variance

The variance measures the average amount by which the data differ from the mean. It is obtained as the mean of the squares of deviations of individual values from the average value.

$$\sigma^2 = \frac{1}{n}\sum_{1=1}^{n}(x_i - \overline{x})^2 \qquad \sigma^2 \geq 0$$

✓ The variance is affected by extreme observations

✓ The numerator of the variance is called Deviance:

$$\mathrm{Dev}(X) = \sum_{i=1}^{n}(x_i - \overline{x})^2$$

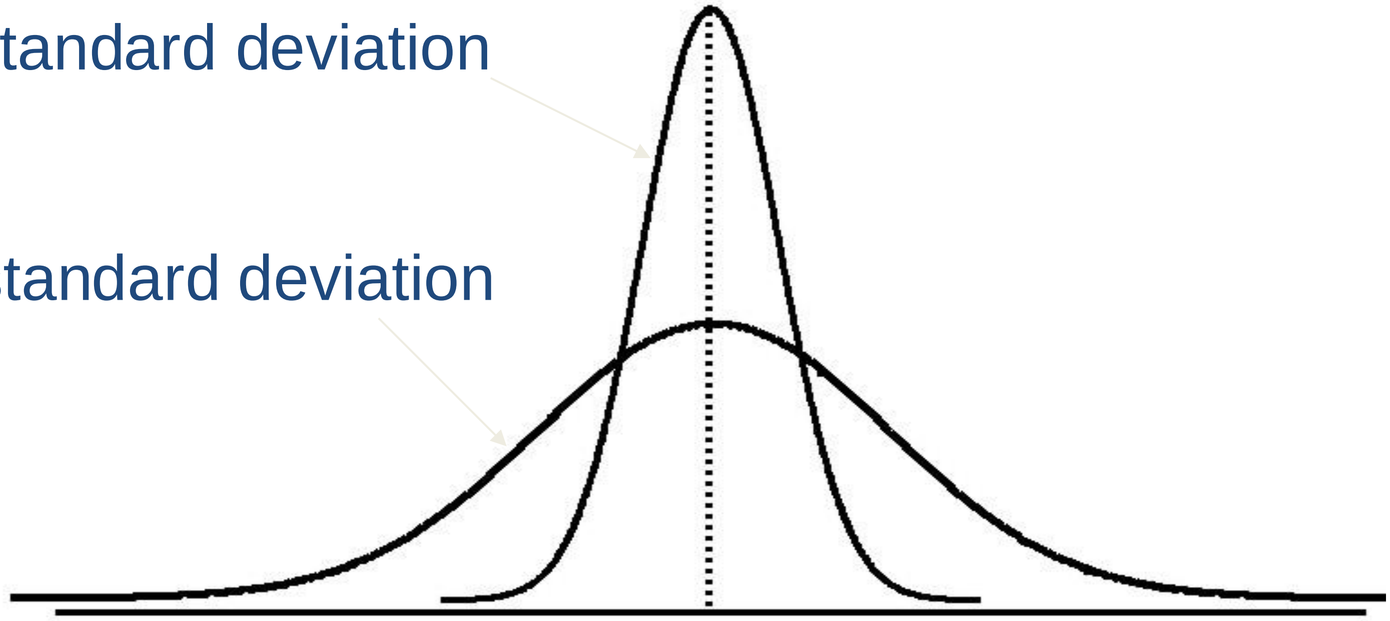
The Standard Deviation (SD)

The standard deviation is simply computed by taking the square root of the variance.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

- ✓ Most commonly used measure of variation
- ✓ Shows variation about the mean
- ✓ Has the same units as the original data

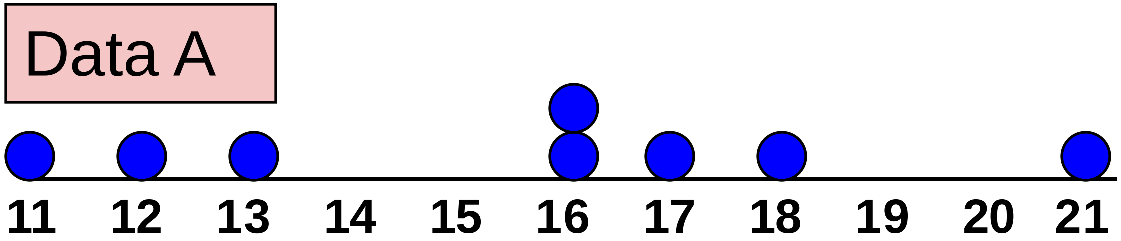
Measuring variation



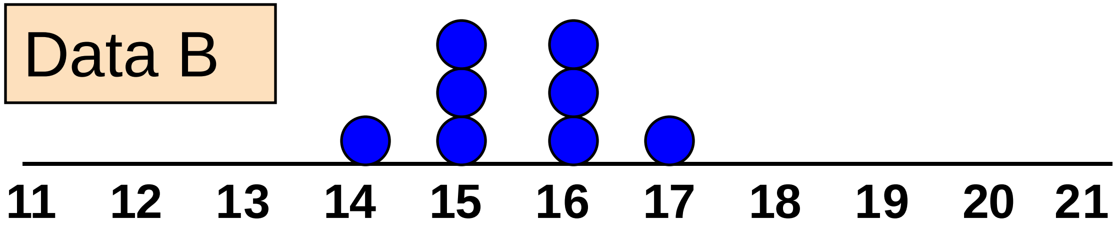
Small standard deviation

Large standard deviation

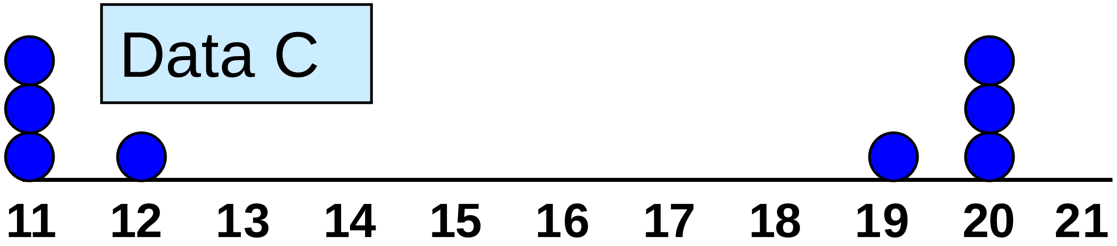
Comparing Standard Deviations



Mean = 15.5
s = 3.338



Mean = 15.5
s = 0.926



Mean = 15.5
s = 4.570

Coefficient of Variation

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare two or more sets of data measured in different units

$$CV = \frac{\sigma}{\overline{x}} 100 \qquad \overline{x} > 0$$

Comparing Coefficient of Variation

- Stock A:
  - Average price last year = $50
  - Standard deviation = $5

$$CV_A = \left(\frac{s}{\bar{x}}\right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = \boxed{10\%}$$

- Stock B:
  - Average price last year = $100
  - Standard deviation = $5

$$CV_B = \left(\frac{s}{\bar{x}}\right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = \boxed{5\%}$$

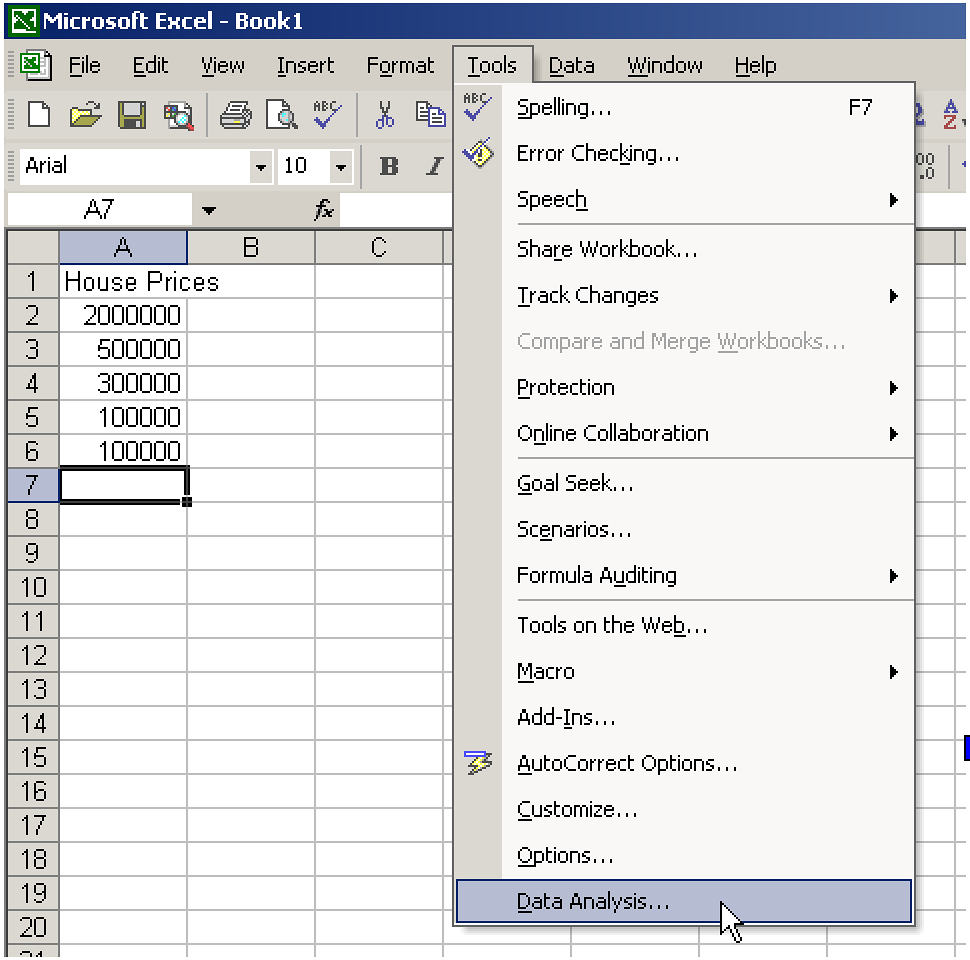
Both stocks have the same standard deviation, but stock B is less variable relative to its price

Using Microsoft Excel

- Descriptive Statistics can be obtained from Microsoft® Excel
  - Use menu choice:

    tools / data analysis / descriptive statistics
  - Enter details in dialog box

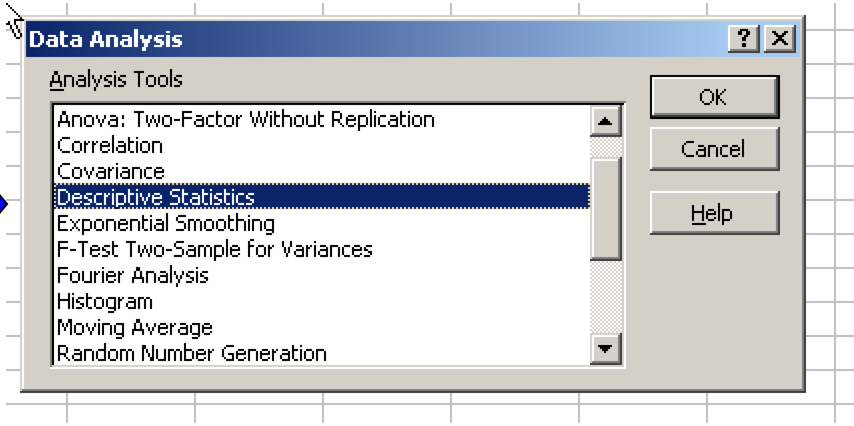
Using Excel



	A
1	House Prices
2	2000000
3	500000
4	300000
5	100000
6	100000
7	

- Use menu choice:

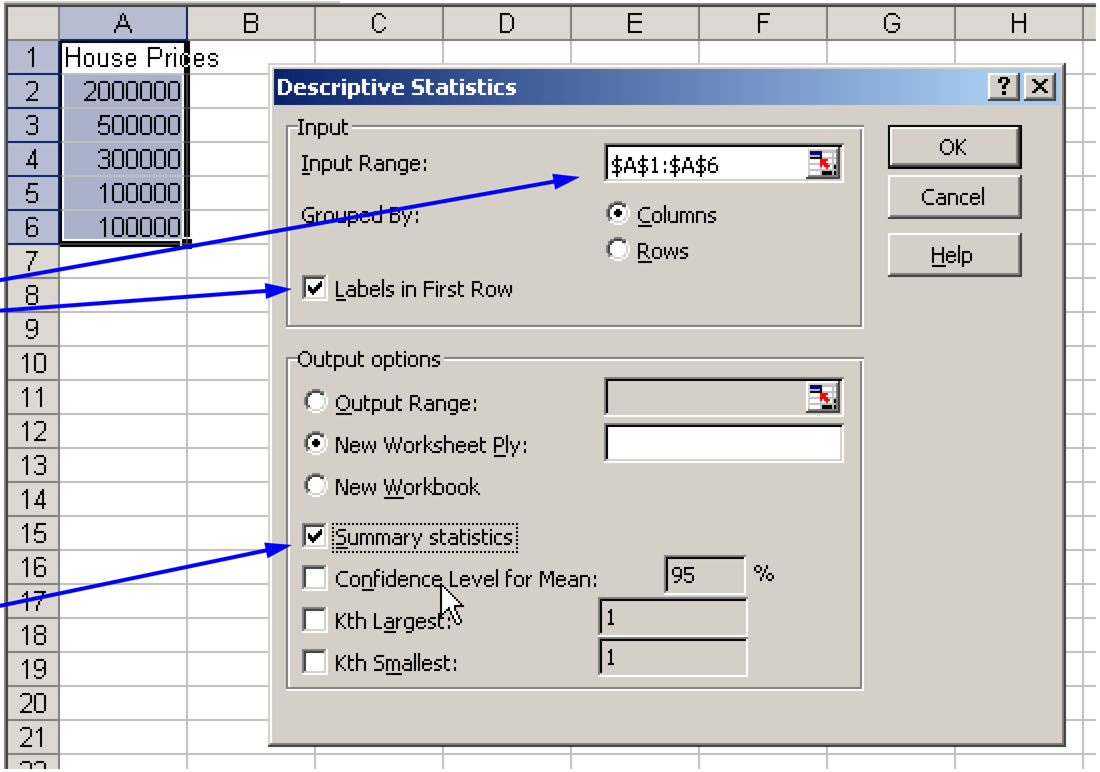
tools / data analysis / descriptive statistics



Using Excel

(continued)

- Enter dialog box details
- Check box for summary statistics
- Click OK



	A
1	House Prices
2	2000000
3	500000
4	300000
5	100000
6	100000

Descriptive Statistics

Input

Input Range: $A$1:$A$6

Grouped By: ● Columns ○ Rows

☑ Labels in First Row

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

☑ Summary statistics

□ Confidence Level for Mean: 95 %

□ Kth Largest: 1

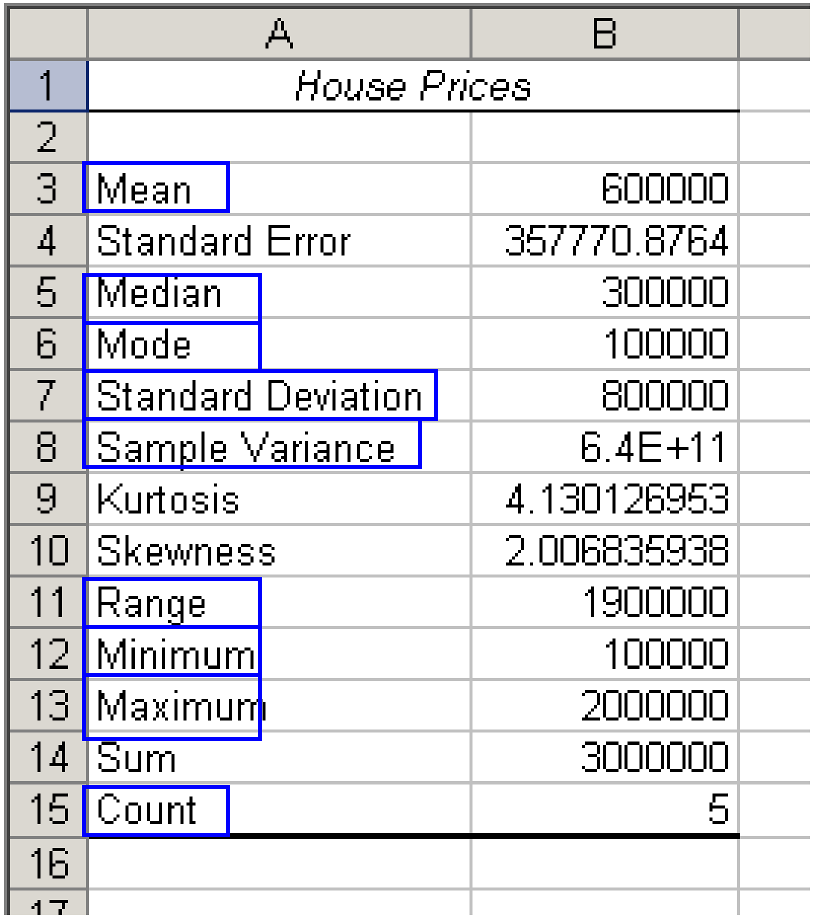
□ Kth Smallest: 1

Excel output

Microsoft Excel descriptive statistics output, using the house price data:

House Prices:

$2,000,000
500,000
300,000
100,000
100,000



	A	B
1	*House Prices*	
2		
3	Mean	600000
4	Standard Error	357770.8764
5	Median	300000
6	Mode	100000
7	Standard Deviation	800000
8	Sample Variance	6.4E+11
9	Kurtosis	4.130126953
10	Skewness	2.006835938
11	Range	1900000
12	Minimum	100000
13	Maximum	2000000
14	Sum	3000000
15	Count	5
16		
17		

Range

The range can then be easily computed, as you have guessed, by subtracting the minimum from the maximum:

```
max(dat$Sepal.Length) - min(dat$Sepal.Length)
```

```
## [1] 3.6
```

Interquartile range

The interquartile range (i.e., the difference between the first and third quartile) can be computed with the `IQR()` function:

```
IQR(dat$Sepal.Length)
```

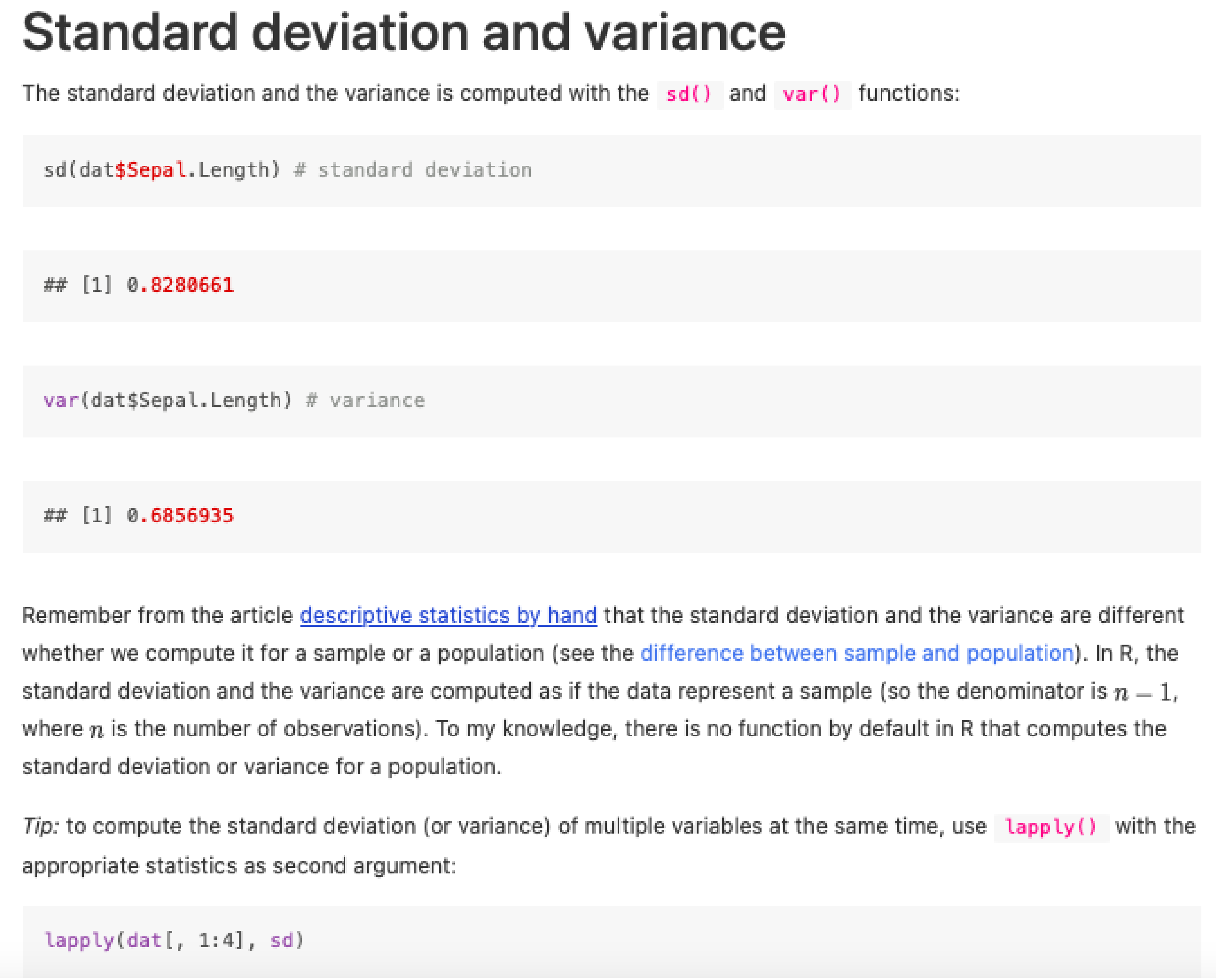
```
## [1] 1.3
```

or alternatively with the `quantile()` function again:

```
quantile(dat$Sepal.Length, 0.75) - quantile(dat$Sepal.Length, 0.25)
```

```
## 75%
## 1.3
```

As mentioned earlier, when possible it is usually recommended to use the shortest piece of code to arrive at the result. For this reason, the `IQR()` function is preferred to compute the interquartile range.



Standard deviation and variance

The standard deviation and the variance is computed with the `sd()` and `var()` functions:

```
sd(dat$Sepal.Length) # standard deviation
```

```
## [1] 0.8280661
```

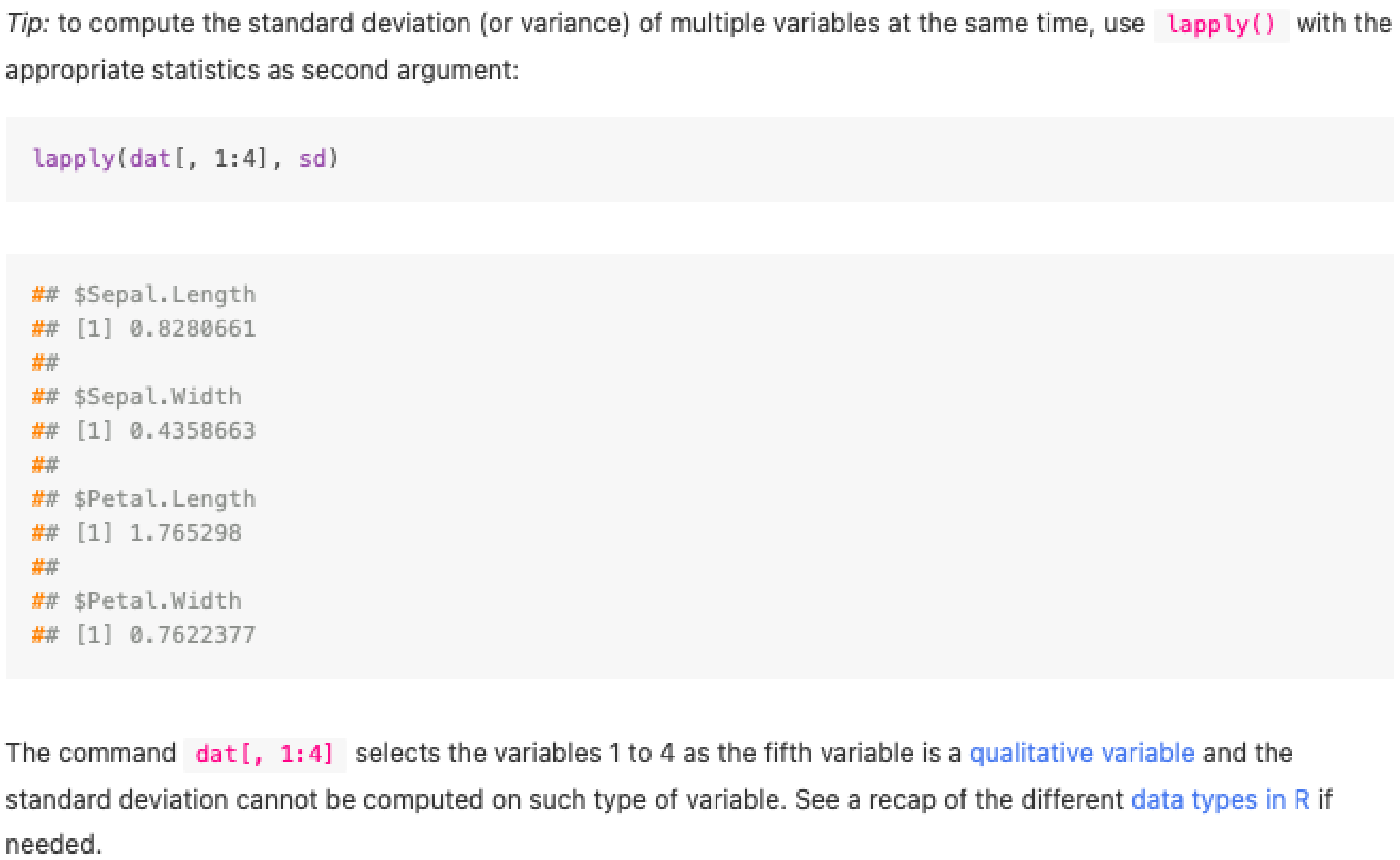
```
var(dat$Sepal.Length) # variance
```

```
## [1] 0.6856935
```

Remember from the article descriptive statistics by hand that the standard deviation and the variance are different whether we compute it for a sample or a population (see the difference between sample and population). In R, the standard deviation and the variance are computed as if the data represent a sample (so the denominator is $n - 1$, where n is the number of observations). To my knowledge, there is no function by default in R that computes the standard deviation or variance for a population.

Tip: to compute the standard deviation (or variance) of multiple variables at the same time, use `lapply()` with the appropriate statistics as second argument:

```
lapply(dat[, 1:4], sd)
```



Tip: to compute the standard deviation (or variance) of multiple variables at the same time, use `lapply()` with the appropriate statistics as second argument:

```
lapply(dat[, 1:4], sd)
```

```
## $Sepal.Length
## [1] 0.8280661
##
## $Sepal.Width
## [1] 0.4358663
##
## $Petal.Length
## [1] 1.765298
##
## $Petal.Width
## [1] 0.7622377
```

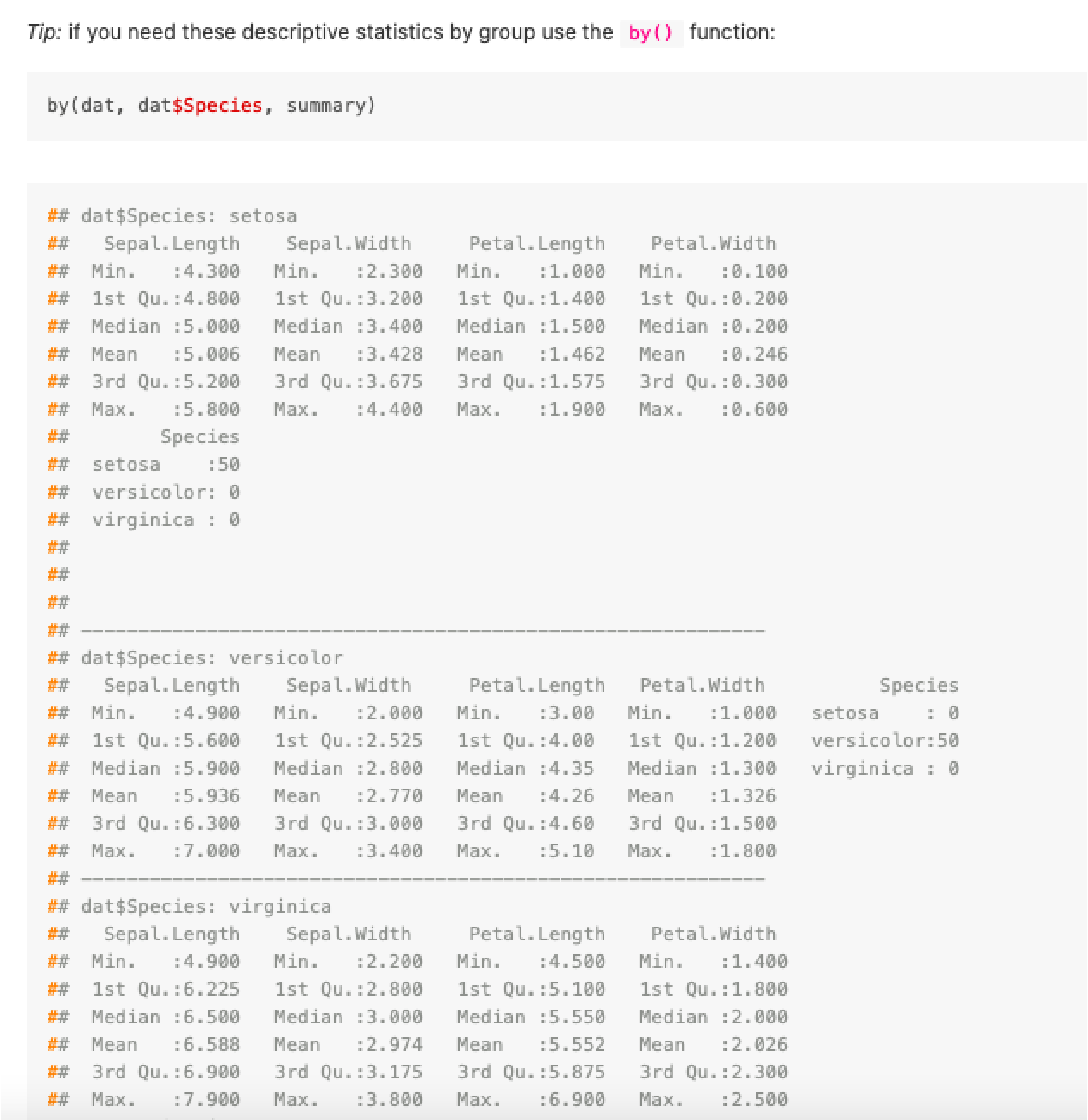
The command `dat[, 1:4]` selects the variables 1 to 4 as the fifth variable is a qualitative variable and the standard deviation cannot be computed on such type of variable. See a recap of the different data types in R if needed.

Summary

You can compute the minimum, 1^{st} quartile, median, mean, 3^{rd} quartile and the maximum for all numeric variables of a dataset at once using `summary()` :

```
summary(dat)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width   
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100  
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300  
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300  
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199  
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800  
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500  
##        Species  
##  setosa    :50  
##  versicolor:50  
##  virginica :50  
##                 
##                 
## 
```



Tip: if you need these descriptive statistics by group use the `by()` function:

```
by(dat, dat$Species, summary)
```

```
## dat$Species: setosa
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width   
##  Min.   :4.300   Min.   :2.300   Min.   :1.000   Min.   :0.100  
##  1st Qu.:4.800   1st Qu.:3.200   1st Qu.:1.400   1st Qu.:0.200  
##  Median :5.000   Median :3.400   Median :1.500   Median :0.200  
##  Mean   :5.006   Mean   :3.428   Mean   :1.462   Mean   :0.246  
##  3rd Qu.:5.200   3rd Qu.:3.675   3rd Qu.:1.575   3rd Qu.:0.300  
##  Max.   :5.800   Max.   :4.400   Max.   :1.900   Max.   :0.600  
##        Species  
##  setosa    :50  
##  versicolor: 0  
##  virginica : 0  
##                 
##                 
##                 
## ------------------------------------------------------------ 
## dat$Species: versicolor
##   Sepal.Length    Sepal.Width     Petal.Length   Petal.Width          Species  
##  Min.   :4.900   Min.   :2.000   Min.   :3.00   Min.   :1.000   setosa    : 0  
##  1st Qu.:5.600   1st Qu.:2.525   1st Qu.:4.00   1st Qu.:1.200   versicolor:50  
##  Median :5.900   Median :2.800   Median :4.35   Median :1.300   virginica : 0  
##  Mean   :5.936   Mean   :2.770   Mean   :4.26   Mean   :1.326                  
##  3rd Qu.:6.300   3rd Qu.:3.000   3rd Qu.:4.60   3rd Qu.:1.500                  
##  Max.   :7.000   Max.   :3.400   Max.   :5.10   Max.   :1.800                  
## ------------------------------------------------------------ 
## dat$Species: virginica
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width   
##  Min.   :4.900   Min.   :2.200   Min.   :4.500   Min.   :1.400  
##  1st Qu.:6.225   1st Qu.:2.800   1st Qu.:5.100   1st Qu.:1.800  
##  Median :6.500   Median :3.000   Median :5.550   Median :2.000  
##  Mean   :6.588   Mean   :2.974   Mean   :5.552   Mean   :2.026  
##  3rd Qu.:6.900   3rd Qu.:3.175   3rd Qu.:5.875   3rd Qu.:2.300  
##  Max.   :7.900   Max.   :3.800   Max.   :6.900   Max.   :2.500  
```

Coefficient of variation

The coefficient of variation can be found with `stat.desc()` (see the line `coef.var` in the table above) or by computing manually (remember that the coefficient of variation is the standard deviation divided by the mean):

```
sd(dat$Sepal.Length) / mean(dat$Sepal.Length)
```

```
## [1] 0.1417113
```