Descriptive Statistics

Erasmus Q-Intelligence B.V.

Data Science and Business Analytics
Programming

Content

Software requirements

$\longrightarrow$ Data from package `datasets` are used

```
R> data("iris")
R> ?iris
R> forsale <- readRDS('../../data/forsale.Rds')
```

```
R> library(dplyr)


Attaching package:  'dplyr'
The following objects are masked from 'package:stats':
   filter, lag
The following objects are masked from 'package:base':
   intersect, setdiff, setequal, union

R> library(ggplot2)
```

Basics

Special values

`NA`	Not available (represents missing value)
`NaN`	Not a number (usually result of division 0/0)
`Inf`	Positive infinity
`-Inf`	Negative infinity
`NULL`	Represents undefined value

Basic math

Operator or function	Operation	Example
`+`	addition	`x + y`
`-`	subtraction	`x - y`
	univariate minus	`-x`
`*`	multiplication	`x * y`
`/`	division	`x / y`
`^`	exponentiation	`x ^ y`
`abs()`	absolute value	`abs(x)`
`sqrt()`	square root	`sqrt(x)`
`log()`	logarithm	`log(x)`
`exp()`	exponential function	`exp(x)`

$\longrightarrow$ Vectorized arithmetic: operations are performed elementwise

First data interpretation

View data

⟶ Get overview of what the data look like

⟶ Using `?`

⟶ Using `head()`

⟶ Using `tail()`

⟶ Using `summary()`

⟶ Using `View()`

Data dimensions

Number of observations and columns together:

```
R> dim(iris)
[1] 150   5
```

Number of observations and columns separately:

```
R> nrow(iris)
[1] 150
R> ncol(iris)
[1] 5
```

Names

Variable names:

```
R> colnames(iris)
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"
[5] "Species"
```

Row names:

```
R> rownames(iris)
```

Frequency

The frequencies can be counted with function `table()`:

```
R> table(iris$Species)

    setosa versicolor  virginica
        50         50         50
```

Descriptive statistics

Some useful statistical functions

Minimum and maximum separate:

```
R> min(iris$Sepal.Length)
[1] 4.3
R> max(iris$Sepal.Length)
[1] 7.9
```

Minimum and maximum together:

```
R> range(iris$Sepal.Length)
[1] 4.3 7.9
```

Quantiles

Default quantiles:

```
R> quantile(iris$Sepal.Length)
  0%  25%  50%  75% 100%
 4.3  5.1  5.8  6.4  7.9
```

Quantiles for specified probabilities:

```
R> quantile(iris$Sepal.Length,
+           probs = c(0.05, 0.25, 0.5, 0.75, 0.95))
   5%   25%   50%   75%   95%
4.600 5.100 5.800 6.400 7.255
```

Center and dispersion

Mean and median:

```
R> mean(iris$Sepal.Length)
[1] 5.843333
R> median(iris$Sepal.Length)
[1] 5.8
```

Standard deviation and variance:

```
R> sd(iris$Sepal.Length)
[1] 0.8280661
R> var(iris$Sepal.Length)
[1] 0.6856935
```

Covariance and correlation

Only numerical variables from data

```
R> iris_num <- iris %>% select(where(is.numeric))
```

Covariance:

```
R> cov(iris_num)
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    0.6856935  -0.0424340    1.2743154   0.5162707
Sepal.Width    -0.0424340   0.1899794   -0.3296564  -0.1216394
Petal.Length    1.2743154  -0.3296564    3.1162779   1.2956094
Petal.Width     0.5162707  -0.1216394    1.2956094   0.5810063
```

Correlation:

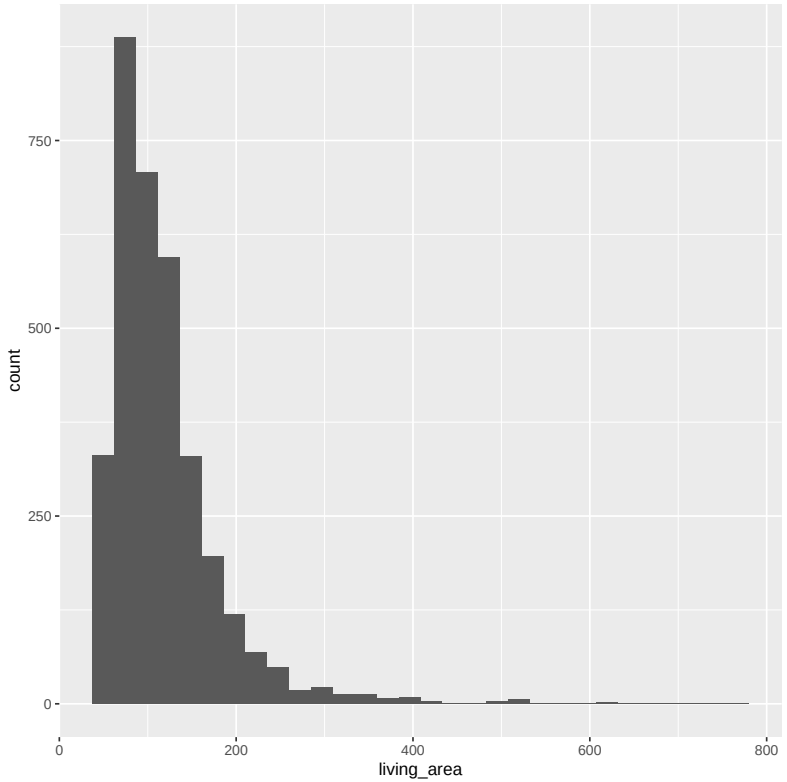
```
R> cor(iris_num)
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

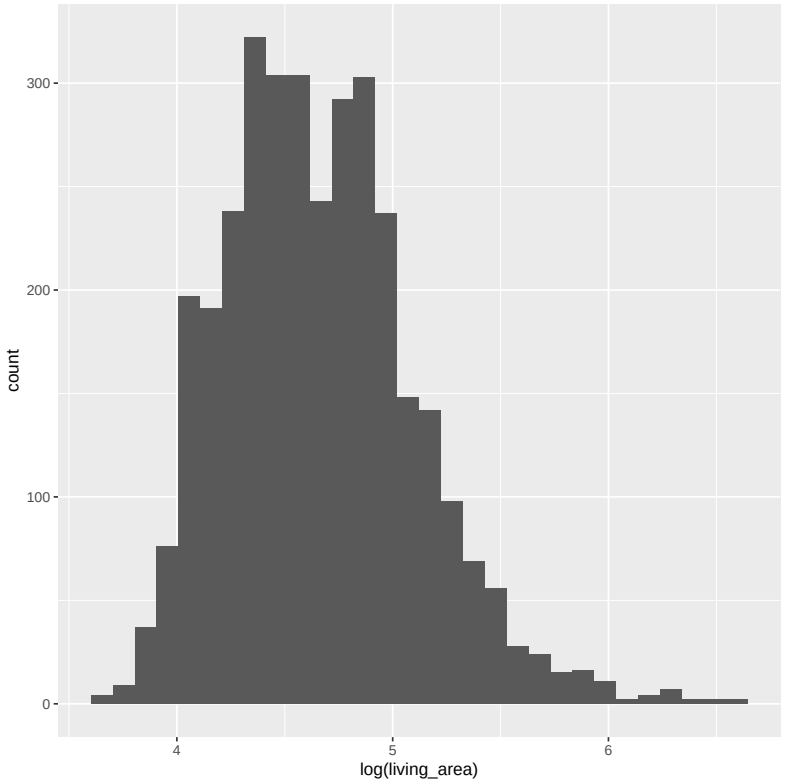
Distribution of a variable

```
R> summary(forsale$living_area)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   40.0    75.0   102.0   117.6   139.5   758.0
```

```
R> summary(log(forsale$living_area))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.689   4.317   4.625   4.661   4.938   6.631
```

Distribution of a variable





Exercises

Download and open the *Descriptives_Exercises.pdf* file from Canvas, and do the Exercises