The problem

Perform a `t-test` to compare the MPG of 4-cylinder cars versus 6-cylinder cars. Before conducting the `t-test`, visualize the distribution of MPG for both groups using box plots or violin plots. Provide an interpretation of the results of the `t-test` in the context of the visualizations.

Tasks & Objectives

In order to solve the problem, it can be useful to divide it into multiple, smaller tasks (as per the "*divide-et-impera*" approach):

1) **Plotting** the **distribution** (with either *box plots* or *violin plots*);
2) **Performing** the **t-test** on the dataset;
3) **Provide** an **interpretation** of the `t-test`.

Tasks & Objectives

In order to solve the problem, it can be useful to divide it into multiple, smaller tasks (as per the "*divide-et-impera*" approach):

1) **Plotting** the **distribution** (with either *box plots* or *violin plots*);
2) **Performing** the **t-test** on the dataset;
3) **Provide** an **interpretation** of the `t-test`.

Tasks & Objectives

In order to solve the problem, it can be useful to divide it into multiple, smaller tasks (as per the "*divide-et-impera*" approach):

1) **Plotting** the **distribution** (with either *box plots* or *violin plots*);
2) **Performing** the **t-test** on the dataset;
3) **Provide** an **interpretation** of the `t-test`.

Tasks & Objectives

In order to solve the problem, it can be useful to divide it into multiple, smaller tasks (as per the "*divide-et-impera*" approach):

1) **Plotting** the **distribution** (with either *box plots* or *violin plots*);
2) **Performing** the **t-test** on the dataset;
3) **Provide** an **interpretation** of the `t-test`.

Installing the packages

```r
# Install the packages
install.packages("datasets")
install.packages("ggplot2")

# Load the libraries
library(datasets)
library(ggplot2)
```

Viewing the dataset

```
head(mtcars)        # In Out[1]
summary(mtcars)     # In Out[2]
```

↓

```
Out[1]                       mpg cyl disp  hp drat    wt  qsec vs am gear carb
     |  Mazda RX4           21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
     |  Mazda RX4 Wag       21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
     |  Datsun 710          22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
     |  Hornet 4 Drive      21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
     |  Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
     |  Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
Out[2]     mpg             cyl             disp             hp
      |  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
      |  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
      |  Median :19.20   Median :6.000   Median :196.3   Median :123.0
      |  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
      |  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
      |  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
      |       drat             wt             qsec             vs
      |  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
      |  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
      |  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
      |  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
      |  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
      |  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
      |        am              gear            carb
      |  Min.   :0.0000   Min.   :3.000   Min.   :1.000
      |  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
      |  Median :0.0000   Median :4.000   Median :2.000
      |  Mean   :0.4062   Mean   :3.688   Mean   :2.812
      |  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
      |  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

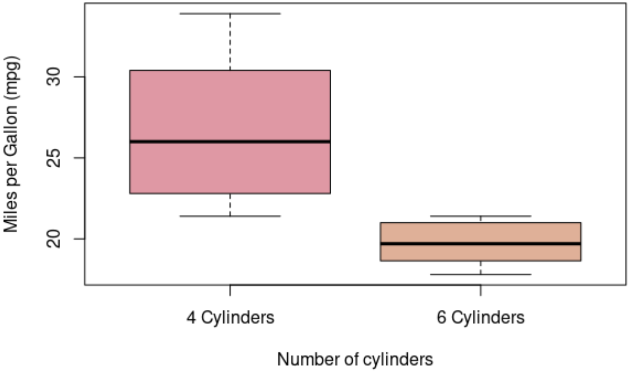
First task

The distribution can be **plotted** with either a *box plot* or a *violin plot*, and to do so the **`boxplot()`** function can be used. In order to plot both the two variables, it's enough to pass them into the function as a vector of variables. The following code uses also some "*eyecandy*" commands in order to easily differentiate the two datasets:

```
boxplot(mtcars$mpg[mtcars$cyl == 4],
        mtcars$mpg[mtcars$cyl == 6],

        # Eyecandy
        names = c("4 Cylinders", "6 Cylinders"),
        col = c("#DF98A4", "#DFB098"),
        xlab = "Cylinders",
        ylab = "Miles per Gallon (mpg)")
```

First task - Result of the Box Plot

The result of the code of the previous slide is the following:



Second task

Now that the box plot has been done, we can pass to the second step: **performing** the **t-test** on the data with the `t.test()` function.

```
# Selecting from both variables the values where ‘cyl‘ is equal
# either to 4 or 6
mtcars_mpg_corr <- mtcars$mpg[mtcars$cyl == 4 | mtcars$cyl == 6]
mtcars_cyl_corr <- mtcars$cyl[mtcars$cyl == 4 | mtcars$cyl == 6]

# Performing the t.test() assuming that the variance is equal
# for both variables
t.test(mtcars_mpg_corr ~ mtcars_cyl_corr, var.equal = TRUE)
```

Second task

The output of the previous code is the following:

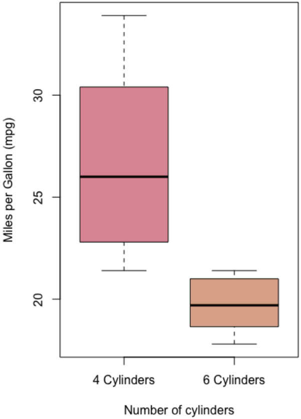
```

	Two Sample t-test

data:  mtcars$mpg[mtcars$cyl == 4 | mtcars$cyl == 6] by mtcars$cyl[
     mtcars$cyl == 4 | mtcars$cyl == 6]
t = 3.8952, df = 16, p-value = 0.001287
alternative hypothesis: true difference in means between group 4 and
      group 6 is not equal to 0
95 percent confidence interval:
  3.154286 10.687272
sample estimates:
mean in group 4 mean in group 6
       26.66364        19.74286
```

Third task

Some observations regarding the result of the `t.test()` and the plot can be drawn:

- by directly looking at the obtained data, it's easy to notice that the means of the two groups are very different from each other. It can be seen both from the plot and from the last part of the `t-test`:

```
mean in group 4 mean in group 6
       26.66364        19.74286
```

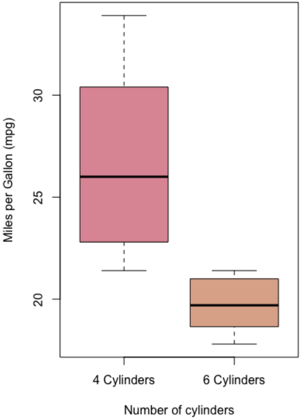


Third task

Some observations regarding the result of the `t.test()` and the plot can be drawn:

- by directly looking at the obtained data, it's easy to notice that the means of the two groups are very different from each other. It can be seen both from the plot and from the last part of the `t-test`:

```
mean in group 4 mean in group 6
       26.66364        19.74286
```



Third task

- more explicitly, the `mpg` value for the cars having `cyl == 4` is higher than the one for the cars having `cyl == 6`;
- this highlights a possible correlation: **on average**, the higher the number of cylinders in a car, the less miles per gallon the car will do. It **doesn't (necessarily) imply causation** though;
- why "*on average*"? Because we are looking at the mean value of each category.

Third task

- more explicitly, the `mpg` value for the cars having `cyl == 4` is higher than the one for the cars having `cyl == 6`;
- this highlights a possible correlation: **on average**, the higher the number of cylinders in a car, the less miles per gallon the car will do. It **doesn't (necessarily) imply causation** though;
- why "*on average*"? Because we are looking at the mean value of each category.

Third task

- more explicitly, the `mpg` value for the cars having **`cyl == 4`** is higher than the one for the cars having **`cyl == 6`**;
- this highlights a possible correlation: **on average**, the higher the number of cylinders in a car, the less miles per gallon the car will do. It **doesn't (necessarily) imply causation** though;
- why "*on average*"? Because we are looking at the mean value of each category.

The Whole Code

```r
# Install the packages needed for the exercise
install.packages(c("datasets", "ggplot2"))

# Call the packages
library(datasets)
library(ggplot2)

# Perform a summary and head in order to have an idea of the dataset
summary(mtcars)
head(mtcars)

# Prepare the box plot, which compares the mpg of the cars with
# cyl == 4 and with cyl == 6
boxplot(mtcars$mpg[mtcars$cyl == 4],
        mtcars$mpg[mtcars$cyl == 6],
        names = c("4 Cylinders", "6 Cylinders"), # Eyecandy
        col = c("#DF98A4", "#DFB098"),
        xlab = "Number of cylinders",
        ylab = "Miles per Gallon (mpg)")
```

The Whole Code

```r
# Performing the t-test of the cars’ mpg with cyl equal to 4 or 6
t.test(
    mtcars$mpg[mtcars$cyl == 4 | mtcars$cyl == 6] ~
    mtcars$cyl[mtcars$cyl == 4 | mtcars$cyl == 6],
    var.equal=TRUE)


# This is a more verbose version of the previous t-test, where the
# two examined variables are first stored in two separate R
# variables and then compared with the t-test
mtcars_mpg_corr <- mtcars$mpg[mtcars$cyl == 4 | mtcars$cyl == 6]
mtcars_cyl_corr <- mtcars$cyl[mtcars$cyl == 4 | mtcars$cyl == 6]

t.test(mtcars_mpg_corr ~ mtcars_cyl_corr, var.equal = TRUE)
```