Practical: What is R

BaselRBootcamp 2017

Slides

Here a link to the lecture slides for this session: **LINK** (https://therbootcamp.github.io/_sessions/D1S1_WhatIsR/What_is_R.html)

Overview

In this practical you'll get started with R. By the end of this practical you will:

1. Know your way around R Studio
2. Know how to run code
3. Have an impression of R basic functionality

Tasks

For this practical you will go through an existing analysis script chunck by chunk to experience how programming and analysing in R works. The idea is that you go through the code, copy the code chunks to the script editor, send the code to the console, and evaluate what happens.

While at it try to practice two very useful shortcuts: (1) `cmd + enter (MAC)` or `cntrl + enter (Windows)` for running the current line in the console. (2) `cmd + shift + p (MAC)` or `cntrl + shift + p (Windows)` for running the same block of code again in the console. The latter is really helpful because you can rerun the a chunk of code after you have made changes to it.

All of the code used in this tutorial is based on basic R functions. While they are already powerful, we will later in the course introduce you to more modern options for several of the steps.

Install and load the yarrr package

1. First we'll install and load the yarrr package. The yarrr package contains many datasets and functions (created by Nathaniel Phillips).

```
# Install and load the yarrr package
# 1install.packages('yarrr')
library(yarrr)
```

```
Loading required package: jpeg
```

```
Loading required package: BayesFactor
```

```
Loading required package: coda
```

```
Loading required package: Matrix
```

```
************
Welcome to BayesFactor 0.9.12-2. If you have questions, please contact Richard Morey
 (richarddmorey@gmail.com).

Type BFManual() to open the manual.
************
```

```
Loading required package: circlize
```

```
yarrr v0.1.5. Citation info at citation('yarrr'). Package guide at yarrr.guide()
```

```
Email me at Nathaniel.D.Phillips.is@gmail.com
```

Explore the pirates dataset

2. The pirates dataset contains data from a survey of 1,000 pirates. Inspect it one-by-one using the following functions.

```
# Get help for pirates data
?pirates

# Print the first few rows of the dataset
head(pirates)
```

```
  id    sex age height weight headband college tattoos tchests parrots
1  1   male  28 173.11   70.5      yes   JSSFP       9        0       0
2  2   male  31 209.25  105.6      yes   JSSFP       9       11       0
3  3   male  26 169.95   77.1      yes    CCCC      10       10       1
4  4 female  31 144.29   58.5       no   JSSFP       2        0       2
5  5 female  41 157.85   58.4      yes   JSSFP       9        6       4
6  6   male  26 190.20   85.4      yes    CCCC       7       19       0
  favorite.pirate sword.type eyepatch sword.time beard.length
1    Jack Sparrow    cutlass        1       0.58           16
2    Jack Sparrow    cutlass        0       1.11           21
3    Jack Sparrow    cutlass        1       1.44           19
4    Jack Sparrow   scimitar        1      36.11            2
5            Hook    cutlass        1       0.11            0
6    Jack Sparrow    cutlass        1       0.59           17
            fav.pixar grogg
1      Monsters, Inc.    11
2               WALL-E     9
3          Inside Out     7
4          Inside Out     9
5          Inside Out    14
6 Monsters University     7
```

```
# Show the structure of the dataset
str(pirates)
```

```
'data.frame':   1000 obs. of  17 variables:
 $ id             : int  1 2 3 4 5 6 7 8 9 10 ...
 $ sex            : chr  "male" "male" "male" "female" ...
 $ age            : num  28 31 26 31 41 26 31 31 28 30 ...
 $ height         : num  173 209 170 144 158 ...
 $ weight         : num  70.5 105.6 77.1 58.5 58.4 ...
 $ headband       : chr  "yes" "yes" "yes" "no" ...
 $ college        : chr  "JSSFP" "JSSFP" "CCCC" "JSSFP" ...
 $ tattoos        : num  9 9 10 2 9 7 9 5 12 12 ...
 $ tchests        : num  0 11 10 0 6 19 1 13 37 69 ...
 $ parrots        : num  0 0 1 2 4 0 7 7 2 4 ...
 $ favorite.pirate: chr  "Jack Sparrow" "Jack Sparrow" "Jack Sparrow" "Jack Sparrow"
 ...
 $ sword.type     : chr  "cutlass" "cutlass" "cutlass" "scimitar" ...
 $ eyepatch       : num  1 0 1 1 1 1 0 1 0 1 ...
 $ sword.time     : num  0.58 1.11 1.44 36.11 0.11 ...
 $ beard.length   : num  16 21 19 2 0 17 1 1 1 25 ...
 $ fav.pixar      : chr  "Monsters, Inc." "WALL-E" "Inside Out" "Inside Out" ...
 $ grogg          : num  11 9 7 9 14 7 9 12 16 9 ...
```

```
# Show the entire dataset in a new window
View(pirates)
```

3. Descriptives for numeric data and categorical data.

```
# What is the mean age?
mean(pirates$age)
```

```
[1] 27.36
```

```
# What was the height of the tallest pirate?
max(pirates$height)
```

```
[1] 209.25
```

```
# How many pirates are there of each sex?
table(pirates$sex)
```

```
female   male  other
   464    490     46
```

4. Descriptive tatistics as a function of another categorical variable.

```
# What was the mean age for each sex?
aggregate(formula = age ~ sex,
          data = pirates,
          FUN = mean)
```

```
     sex      age
1 female 29.92241
2   male 24.96735
3  other 27.00000
```

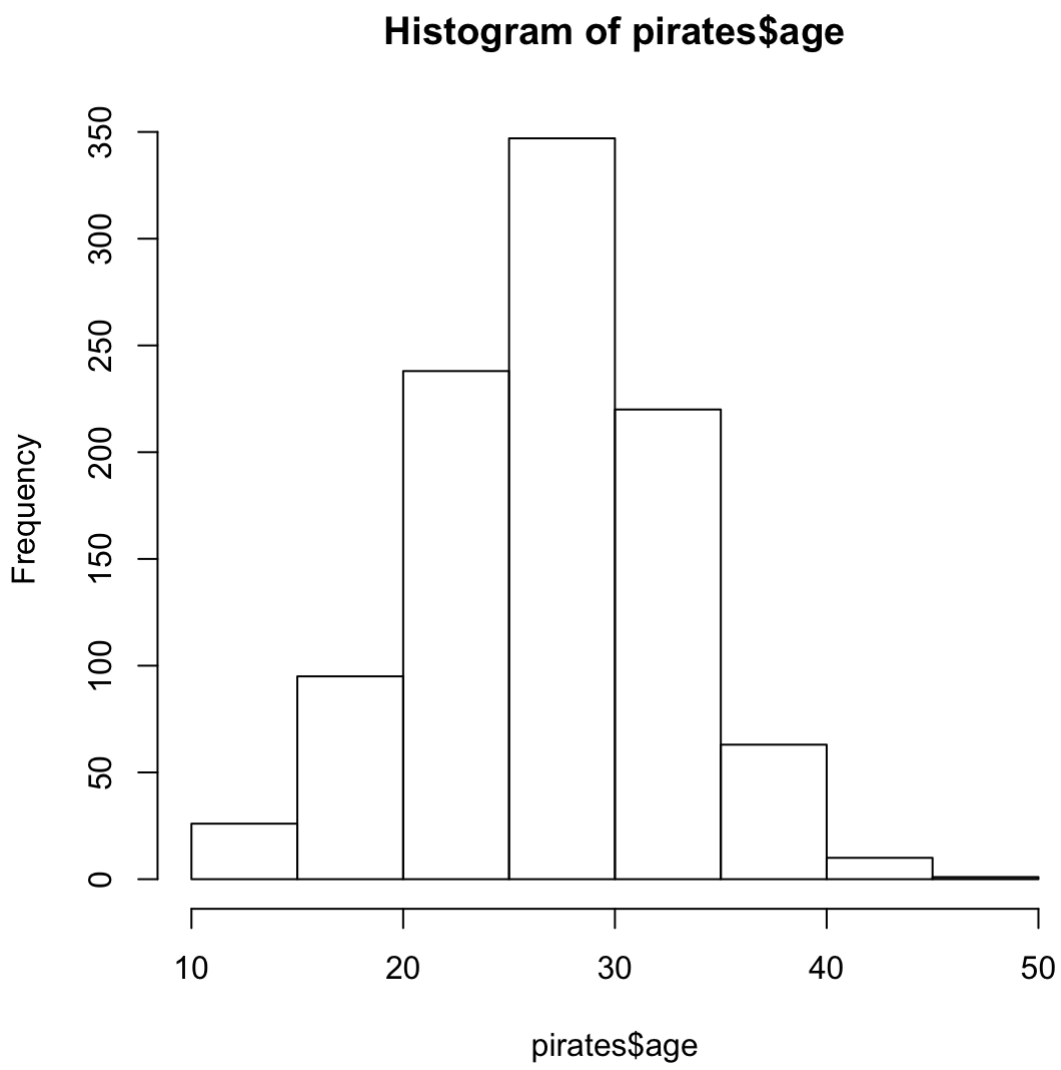
```
# What is the median age of pirates for each combination of sex and headband?
aggregate(formula = age ~ sex + headband,
          data = pirates,
          FUN = median)
```

```
     sex headband age
1 female       no  31
2   male       no  25
3  other       no  25
4 female      yes  30
5   male      yes  25
6  other      yes  27
```

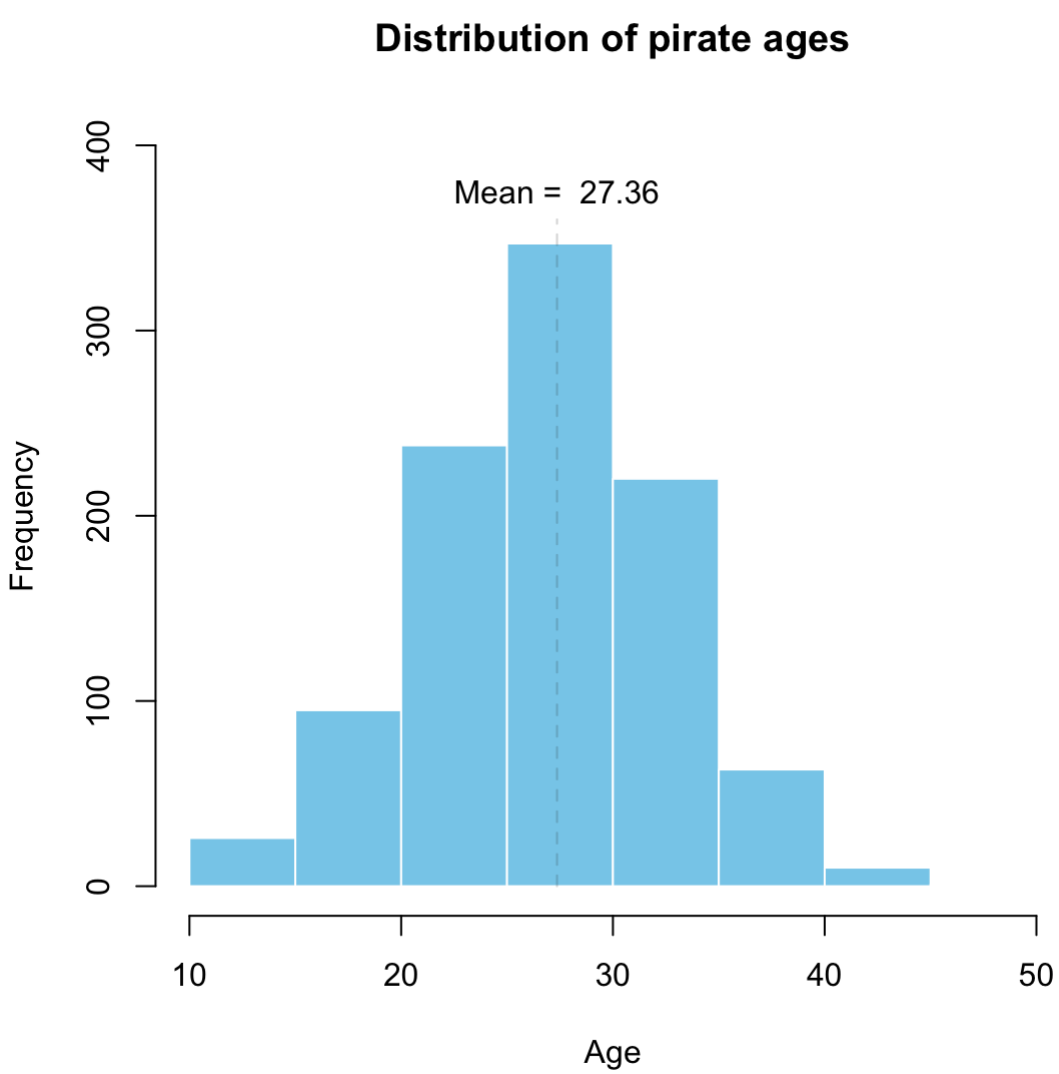
Base plotting (aka high-level plotting)

5. Creating a histograms of numeric variables.

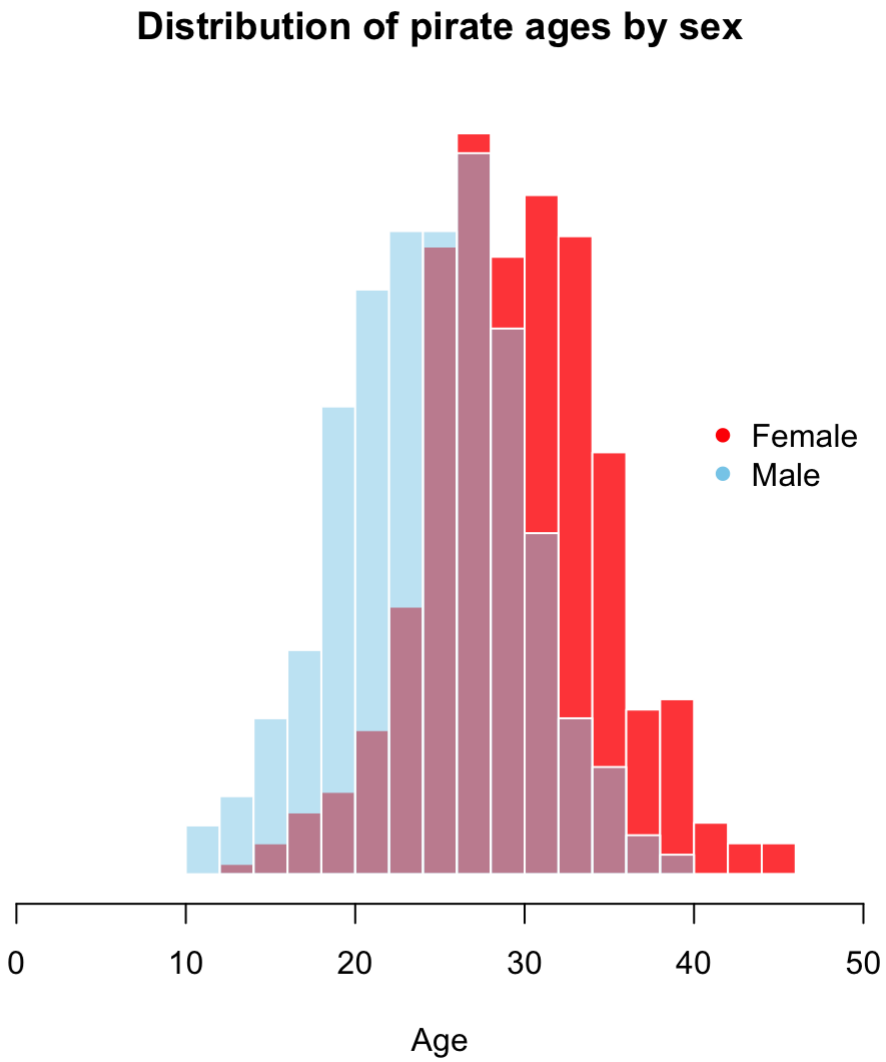
```
# --- A default histogram of pirate ages
hist(x = pirates$age)
```




```
# --- A customized histogram of pirate ages
hist(x = pirates$age,
     main = "Distribution of pirate ages",
     col = "skyblue",
     border = "white",
     xlab = "Age",
     ylim = c(0, 400))

# Add mean label
text(x = mean(pirates$age), y = 375,
     labels = paste("Mean = ", round(mean(pirates$age), 2)))

# Add dashed line at mean
segments(x0 = mean(pirates$age), y0 = 0,
         x1 = mean(pirates$age), y1 = 360,
         col = gray(.2, .2),
         lty = 2)
```

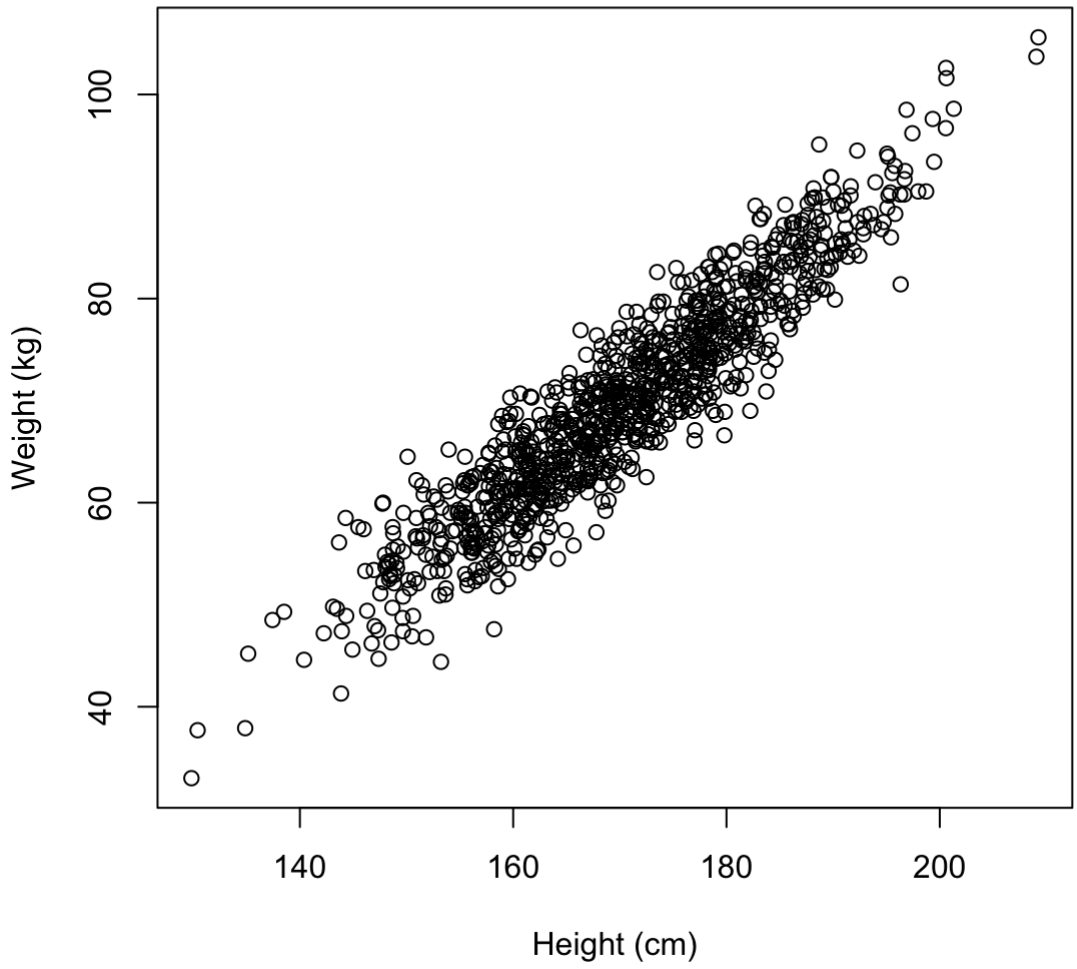

Distribution of pirate ages
Mean = 27.36
Frequency
0
100
200
300
400
10
20
30
40
50
Age

```
# ---- Overlapping histograms of pirate ages for females and males

# Start with the female data
hist(x = pirates$age[pirates$sex == "female"],
     main = "Distribution of pirate ages by sex",
     col = transparent("red", .2),
     border = "white",
     xlab = "Age",
     breaks = seq(0, 50, 2),
     probability = T,
     ylab = "",
     yaxt = "n")

# Add male data
hist(x = pirates$age[pirates$sex == "male"],
     add = T,
     probability = T,
     border = "white",
     breaks = seq(0, 50, 2),
     col = transparent("skyblue", .5))

# Add the legend
legend(x = 40,
       y = .05,
       col = c("red", "skyblue"),
       legend = c("Female", "Male"),
       pch = 16,
       bty = "n")
```

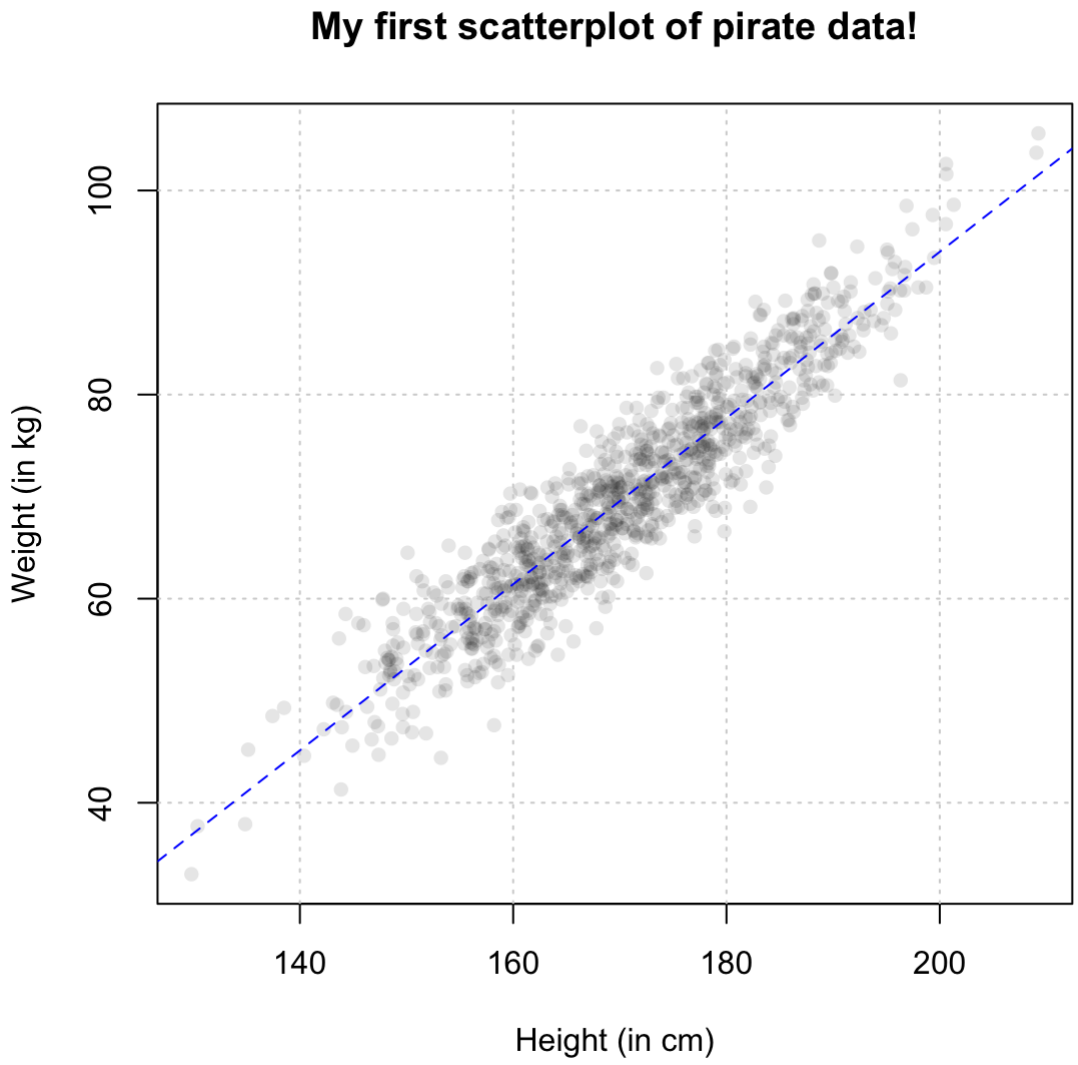


6. Creating scatterplots of two numerical variables.

```
# --- A simple scatterplot of pirate height and weight
plot(x = pirates$height,
     y = pirates$weight,
     xlab = "Height (cm)",
     ylab = "Weight (kg)")
```

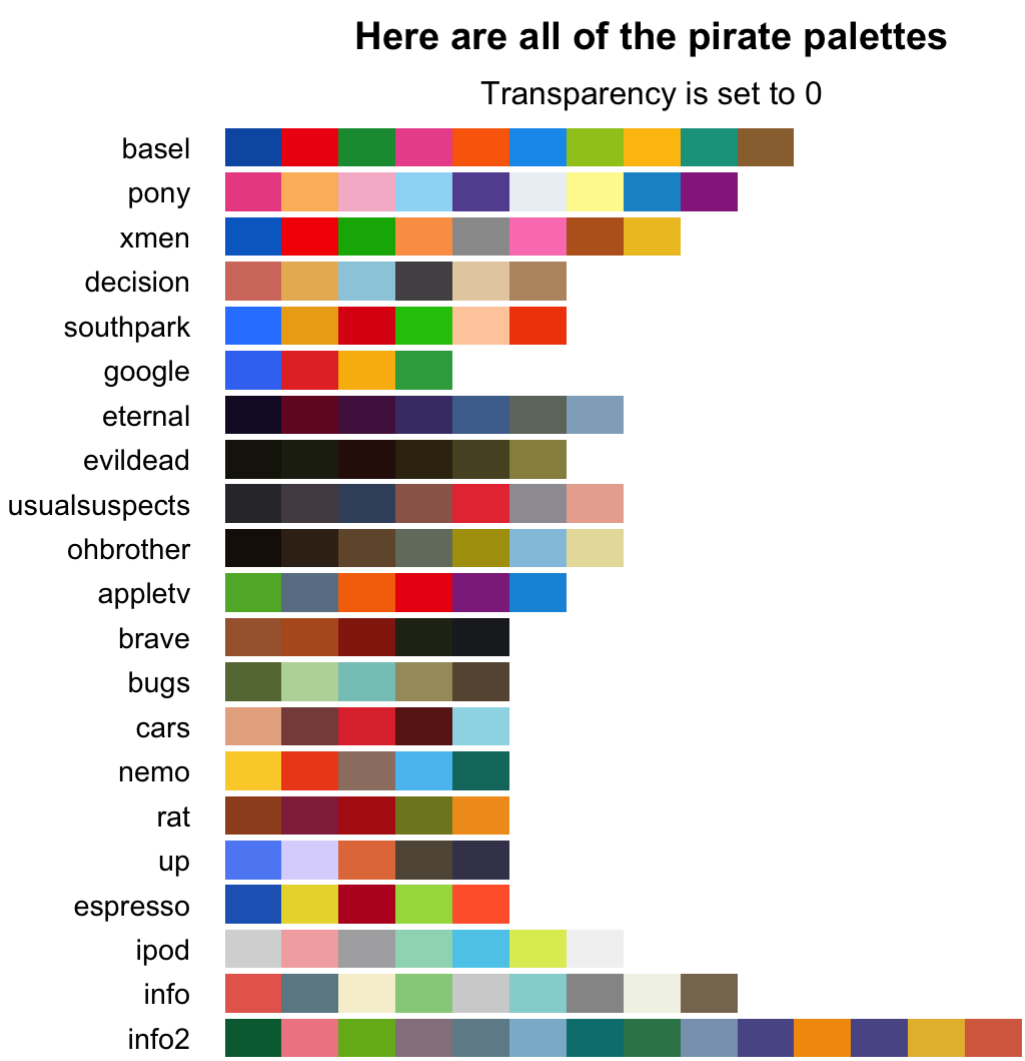



```
# --- A fancier scatterplot of the same data with some additional arguments

# Create main plot
plot(x = pirates$height,
     y = pirates$weight,
     main = 'My first scatterplot of pirate data!',
     xlab = 'Height (in cm)',
     ylab = 'Weight (in kg)',
     pch = 16,    # Filled circles
     col = gray(0, .1)) # Transparent gray

# Add gridlines
grid()

# Create a linear regression model
model <- lm(formula = weight ~ height,
            data = pirates)

# Add regression to plot
abline(model,
       col = 'blue', lty = 2)
```

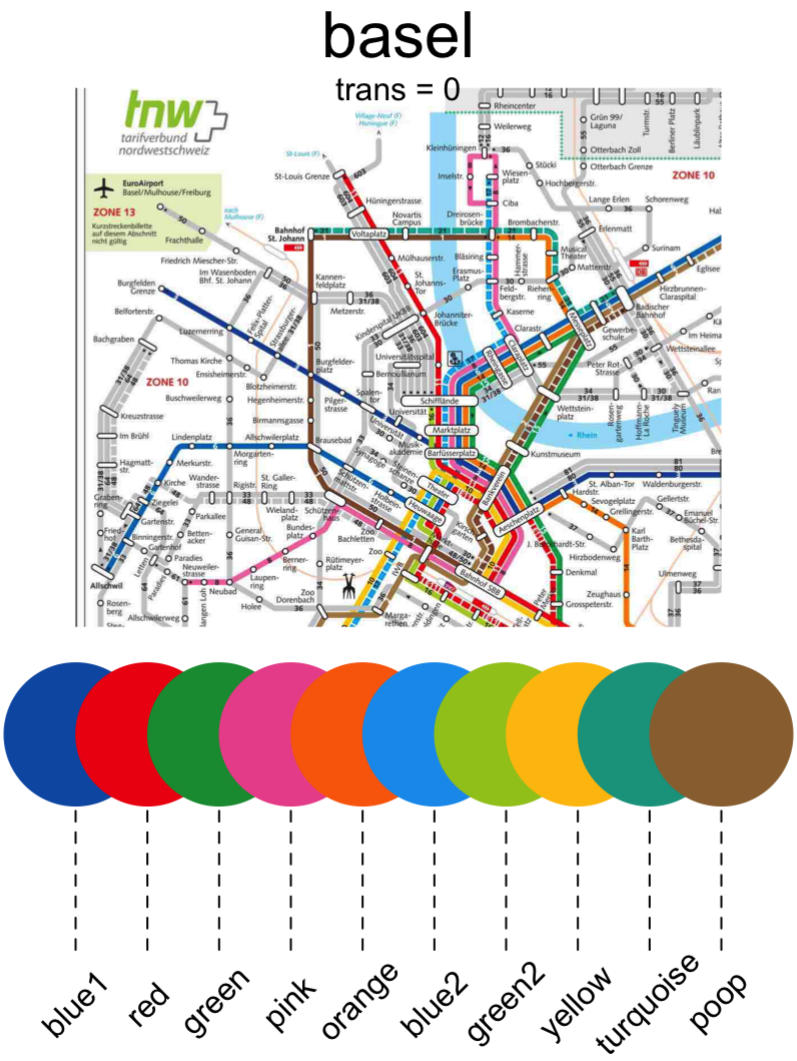


7. Changing colors

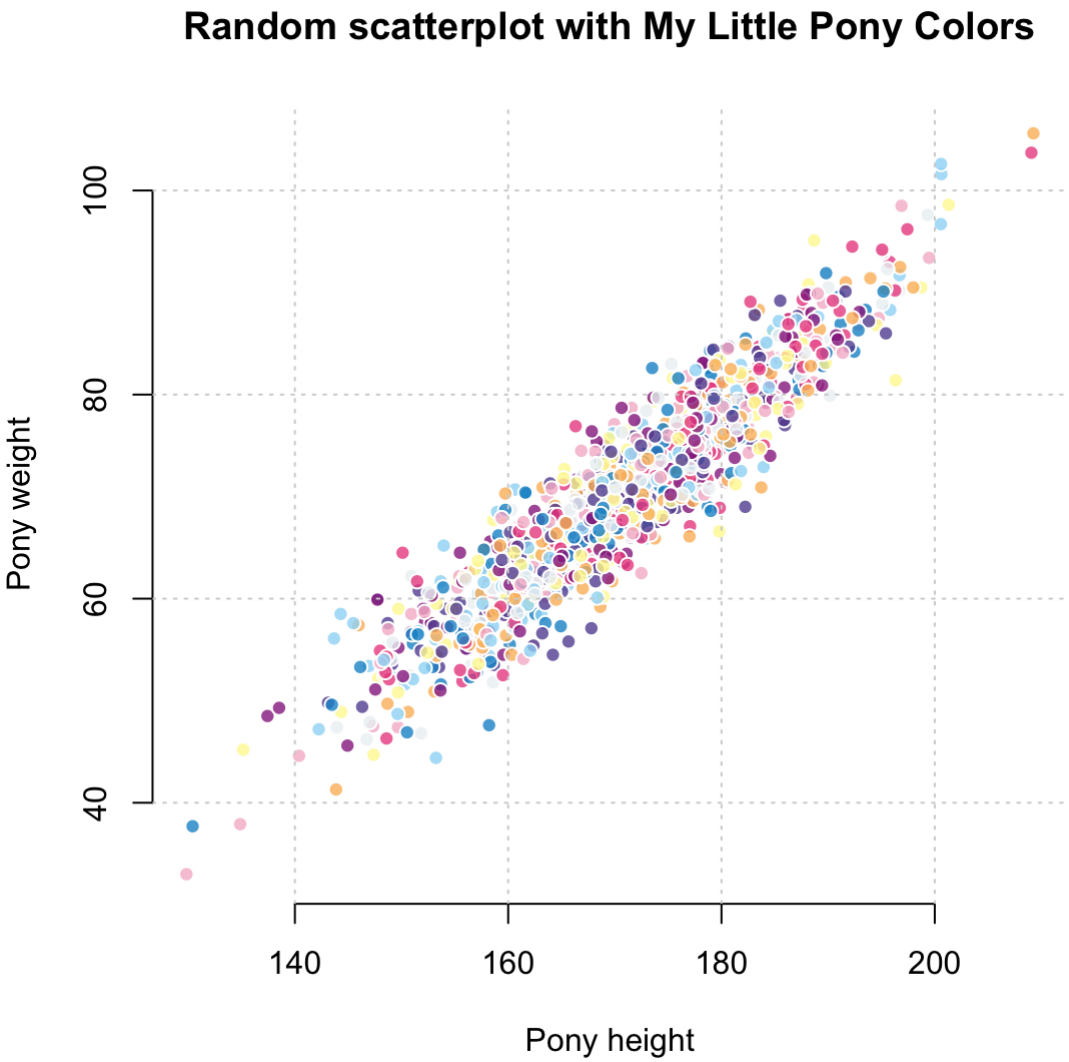
```
# --- Look at all the palettes from piratepal()
piratepal()
```




```
# Look at the basel palette in detail
piratepal(palette = "basel", plot.result = TRUE)
```

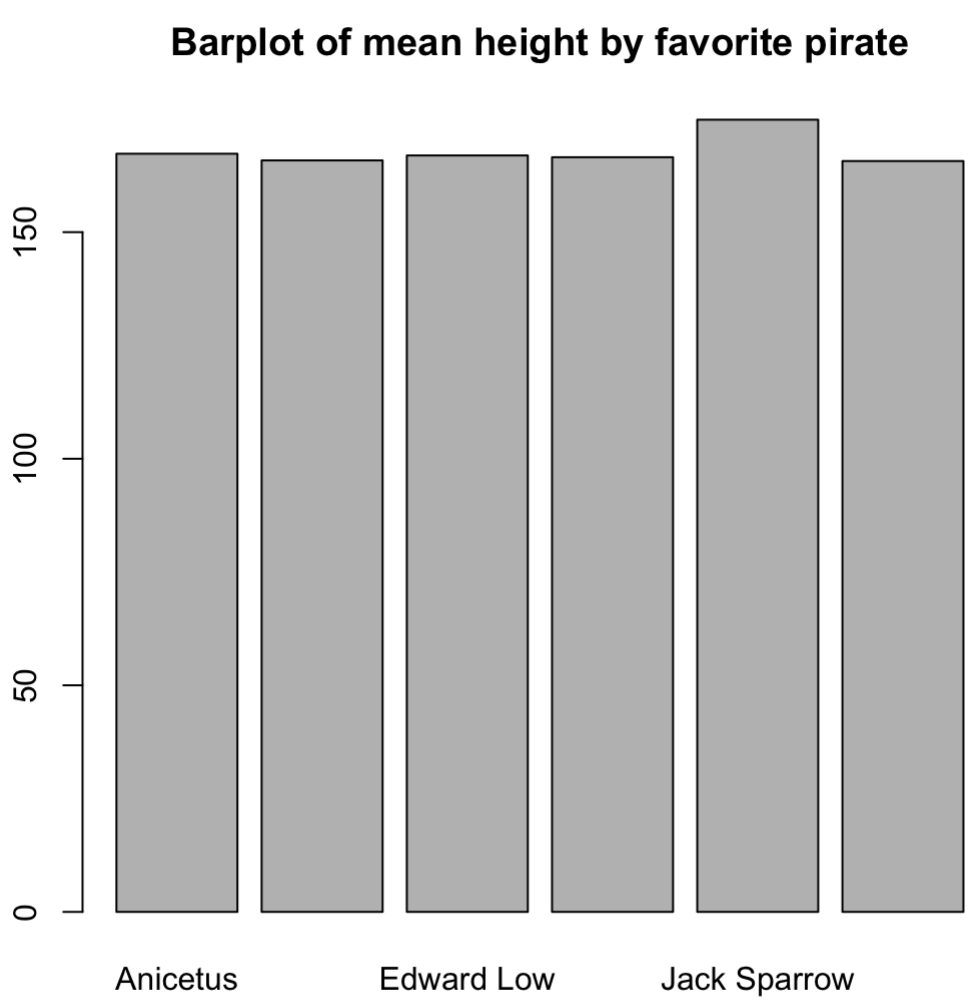



```
# --- Scatterplot of pirate height and weight using the pony palette
my.cols <- piratepal(palette = "pony",
                     trans = .2,
                     length.out = nrow(pirates))

# Create the plot
plot(x = pirates$height, y = pirates$weight,
     main = "Random scatterplot with My Little Pony Colors",
     xlab = "Pony height",
     ylab = "Pony weight",
     pch = 21,  # Round symbols with borders
     col = "white",  # White border
     bg = my.cols,  # Random colors
     bty = "n"  # No plot border
     )

# Add gridlines
grid()
```

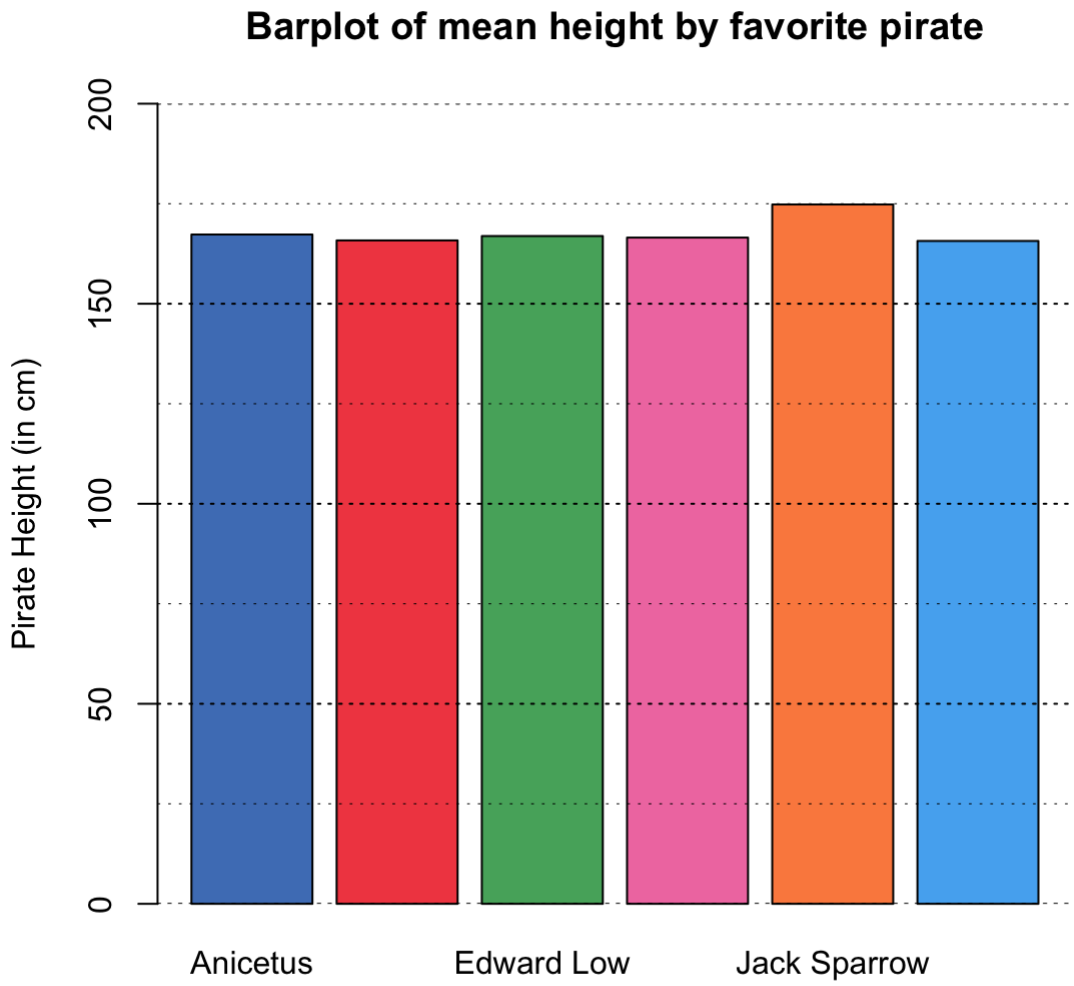


8. Create a barplot for stratified data

```
# --- Barpot of mean height by favorite.pirate

# Calculate mean height for each favorite.pirate
pirate.heights <- aggregate(height ~ favorite.pirate,
                     data = pirates,
                     FUN = mean)

barplot(pirate.heights$height,
        main = "Barplot of mean height by favorite pirate",
        names.arg = pirate.heights$favorite.pirate)
```




```
# --- Same, but with customizations
barplot(pirate.heights$height,
        ylim = c(0, 200),
        ylab = "Pirate Height (in cm)",
        main = "Barplot of mean height by favorite pirate",
        names.arg = pirate.heights$favorite.pirate,
        col = piratepal("basel", trans = .2))

abline(h = seq(0, 200, 25), lty = 3, lwd = c(1, .5))
```



Hypothesis testing

9. Run a group comparisons.

```
# Do pirates with eyepatches have longer beards than those without eyepatches?
t.test(formula = beard.length ~ eyepatch,
       data = pirates,
       alternative = 'two.sided')
```

```
    Welch Two Sample t-test

data:  beard.length by eyepatch
t = 1.4404, df = 674.27, p-value = 0.1502
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3625525  2.3593174
sample estimates:
mean in group 0 mean in group 1
       11.04094        10.04255
```

```
# ANOVA on beard.length as a function of sex and college

# Run the ANOVA
beard.aov <- aov(formula = beard.length ~ sex + college,
                 data = pirates)

# Print summary results
summary(beard.aov)
```

```
             Df Sum Sq Mean Sq  F value Pr(>F)
sex           2  87174   43587 2276.901 <2e-16 ***
college       1     12      12    0.641  0.424
Residuals   996  19067      19
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

10. Run a regression analysis.

```
# regression analysis showing if age, weight, and tattoos predict how many treasure c
hests a pirate has found

# Run the regression
chests.lm <- lm(formula = tchests ~ age + weight + tattoos,
                data = pirates)

# Print summary results
summary(chests.lm)
```

```

Call:
lm(formula = tchests ~ age + weight + tattoos, data = pirates)

Residuals:
    Min      1Q  Median      3Q     Max
-33.302 -15.832  -6.860   8.407 119.966

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.19084    7.18437   0.723     0.47
age          0.78177    0.13438   5.818 8.03e-09 ***
weight      -0.09013    0.07183  -1.255     0.21
tattoos      0.25398    0.22550   1.126     0.26
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.99 on 996 degrees of freedom
Multiple R-squared:  0.04056,   Adjusted R-squared:  0.03767
F-statistic: 14.04 on 3 and 996 DF,  p-value: 5.751e-09
```

Additional reading

- For more details on all steps of data analysis check out Hadley Wickham's R for Data Science (http://r4ds.had.co.nz/).

- For more on pirates and data analysis check out the respective chapters in YaRrr! The Pirate's Guide to R YaRrr! Chapter Link (https://bookdown.org/ndphillips/YaRrr/htests.html)