ANOVA example

The file sbux_anova.csv contains nutrition information on a bunch of items sold at a sample of Starbucks locations in the U.S. The variables are clearly labeled in the file (the units are grams for fat, carbohydrates, fiber and protein). We will carry out an ANOVA analysis to determine whether the mean calories in each type of food item (bakery vs. bistro vs. etc) differs significantly from each other.

Follow the instructions below to download the file, and to carry out the steps in an ANOVA analysis.

In [23]:
```
# Read sbux_anova.csv and store in a dataframe:
#
sbdat = read.csv(file="https://cs.earlham.edu/~pardhan/sage_and_r/sbux_anova.csv", header=TRUE, sep=",")
head(sbdat)
```

A data.frame: 6 × 7

	item	calories	fat	carb	fiber	protein	type
	<fct>	<int>	<dbl>	<int>	<int>	<int>	<fct>
1	8-Grain Roll	350	8	67	5	10	bakery
2	Apple Bran Muffin	350	9	64	7	6	bakery
3	Apple Fritter	420	20	59	0	5	bakery
4	Banana Nut Loaf	490	19	75	4	7	bakery
5	Birthday Cake Mini Doughnut	130	6	17	0	0	bakery
6	Blueberry Oat Bar	370	14	47	5	6	bakery

In [24]:
```
# Check how many and what types of food items are in the data:
#
levels(sbdat$type)
```

'bakery' · 'bistro box' · 'hot breakfast' · 'parfait' · 'petite' · 'sandwich'

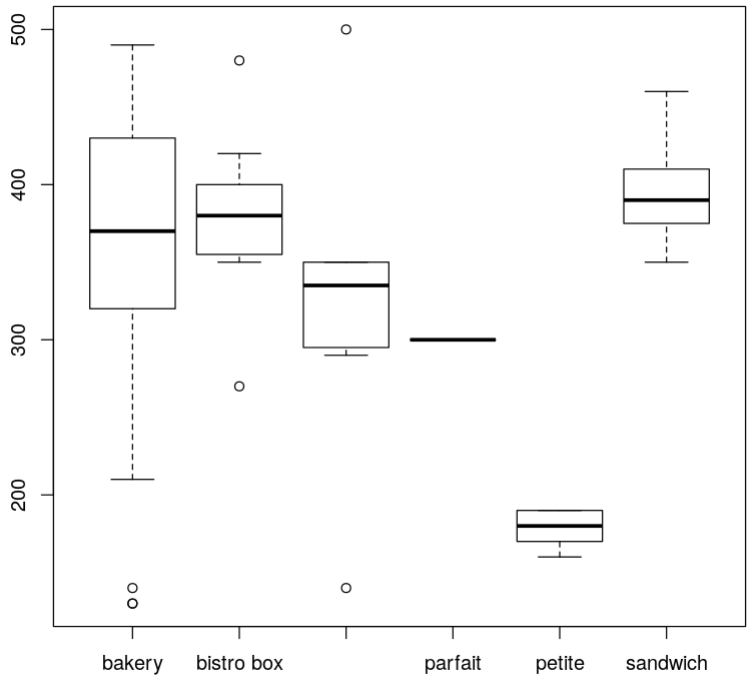
In [33]:
```r
# Group the data by type of food item, and compute the mean and SD
# of each group:
#
library(dplyr)
sbdat %>%
    group_by(type) %>%
    summarise(
        count_tyep = n(),
        mean_cal = mean(calories, na.rm = TRUE),
        sd_cal = sd(calories, na.rm = TRUE)
    )
```

```
`summarise()` ungrouping output (override with `.groups` argument)
```

A tibble: 6 × 4

type	count_tyep	mean_cal	sd_cal
<fct>	<int>	<dbl>	<dbl>
bakery	41	368.7805	95.29415
bistro box	8	377.5000	59.70164
hot breakfast	8	325.0000	98.99495
parfait	3	300.0000	0.00000
petite	9	177.7778	10.92906
sandwich	7	395.7143	36.90399

In [29]:
```
# To check the conditions we will look at side-by-side boxplots.
# If you are a serious data scientist, you should also check histog
rams and
# normal probability plots.
#
boxplot(sbdat$calories ~ sbdat$type)
```



In [34]:
```
# Compute and print the ANOVA table:
#
aout = aov(calories ~ type, data=sbdat )
summary (aout)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
type         5 310004   62001   9.315 7.79e-07 ***
Residuals   70 465916    6656
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In [35]:
```
# Since the results shows a statistically significant difference,
# let's compare the pairwise differences in means:
#
TukeyHSD(aout)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = calories ~ type, data = sbdat)

$type
                                diff        lwr        upr     p adj
bistro box-bakery           8.719512  -83.67655  101.11557 0.9997720
hot breakfast-bakery      -43.780488 -136.17655   48.61557 0.7337624
parfait-bakery            -68.780488 -211.75747   74.19649 0.7211541
petite-bakery            -191.002710 -278.99896 -103.00646 0.0000003
sandwich-bakery            26.933798  -70.82863  124.69623 0.9652776
hot breakfast-bistro box  -52.500000 -172.02599   67.02599 0.7910770
parfait-bistro box        -77.500000 -239.33895   84.33895 0.7249942
petite-bistro box        -199.722222 -315.88060  -83.56384 0.0000506
sandwich-bistro box        18.214286 -105.50686  141.93544 0.9980248
parfait-hot breakfast     -25.000000 -186.83895  136.83895 0.9975152
petite-hot breakfast     -147.222222 -263.38060  -31.06384 0.0052506
sandwich-hot breakfast     70.714286  -53.00686  194.43544 0.5528039
petite-parfait           -122.222222 -281.59020   37.14576 0.2299886
sandwich-parfait           95.714286  -69.24725  260.67582 0.5363603
sandwich-petite           217.936508   97.46564  338.40738 0.0000185
```

In [ ]:
```
# Look at the P-values and determine which pairs of differences
# are statistically significant.
```