```
## (C) (cc by-sa) Wouter van Atteveldt, file generated juni 01 2016
```

> Note on the data used in this howto: This data can be downloaded from http://piketty.pse.ens.fr/files/capital21c/en/xls/, but the excel format is a bit difficult to parse at it is meant to be human readable, with multiple header rows etc. For that reason, I've extracted csv files for some interesting tables that I've uploaded to https://github.com/vanatteveldt/learningr/tree/master/data. If you're accessing this tutorial from the github project, these files should be in your 'data' sub folder automatically.

Playing with data in R

To demonstrate R, we will use the data from Piketty's 'Capital in the 21st Century'

```
income = read.csv("data/income_topdecile.csv")
```

We've downloaded a csv file and read it into a new variable `income`, which should appear in your environment list. You can click on the file to inspect it visually, but we can also use the `head` command:

```
head(income, n=10)
```

```
##    Year U.S. U.K. Germany France Sweden Europe
## 1  1900 0.41 0.47    0.45   0.46   0.46   0.46
## 2  1901   NA   NA      NA     NA     NA     NA
## 3  1902   NA   NA      NA     NA     NA     NA
## 4  1903   NA   NA      NA     NA     NA     NA
## 5  1904   NA   NA      NA     NA     NA     NA
## 6  1905   NA   NA      NA     NA     NA     NA
## 7  1906   NA   NA      NA     NA     NA     NA
## 8  1907   NA   NA      NA     NA     NA     NA
## 9  1908   NA   NA      NA     NA     NA     NA
## 10 1909   NA   NA      NA     NA     NA     NA
```

As you can see, the values are NA (missing) for most rows, especially in the earlier period. Let's throw out all data containing missing values using the `na.omit` function:

```
income = na.omit(income)
head(income)
```

```
##    Year U.S. U.K. Germany France Sweden Europe
## 1  1900 0.41 0.47    0.45   0.46   0.46   0.46
## 11 1910 0.41 0.47    0.44   0.47   0.46   0.46
## 21 1920 0.45 0.41    0.39   0.42   0.36   0.39
## 31 1930 0.45 0.39    0.42   0.43   0.38   0.40
## 41 1940 0.36 0.34    0.34   0.33   0.33   0.34
## 51 1950 0.34 0.30    0.33   0.34   0.29   0.32
```

Much better. Now, we can list the variables in the file using `names` and get the numbers of rows or columns with `nrow` and `ncol`, respectively:

```
names(income)
```

```
## [1] "Year"    "U.S."    "U.K."    "Germany" "France"  "Sweden"  "Europe"
```

```
nrow(income)
```

```
## [1] 12
```

```
ncol(income)
```

```
## [1] 7
```

We can also ask for a summary of each of the variables in the file using the `summary` command:

```
summary(income)
```

```
##       Year           U.S.             U.K.           Germany
##  Min.   :1900   Min.   :0.3300   Min.   :0.2800   Min.   :0.3100
##  1st Qu.:1928   1st Qu.:0.3550   1st Qu.:0.3225   1st Qu.:0.3275
##  Median :1955   Median :0.4100   Median :0.3850   Median :0.3500
##  Mean   :1955   Mean   :0.4025   Mean   :0.3733   Mean   :0.3642
##  3rd Qu.:1982   3rd Qu.:0.4500   3rd Qu.:0.4125   3rd Qu.:0.3975
##  Max.   :2010   Max.   :0.4800   Max.   :0.4700   Max.   :0.4500
##      France           Sweden           Europe
##  Min.   :0.3100   Min.   :0.2200   Min.   :0.2900
##  1st Qu.:0.3300   1st Qu.:0.2675   1st Qu.:0.3200
##  Median :0.3350   Median :0.2950   Median :0.3400
##  Mean   :0.3692   Mean   :0.3217   Mean   :0.3575
##  3rd Qu.:0.4225   3rd Qu.:0.3650   3rd Qu.:0.3925
##  Max.   :0.4700   Max.   :0.4600   Max.   :0.4600
```

This lists the range, mean, etc. for each variable. We can select any column from a data frame using variable$column:

```
income$U.S.
```

```
##  [1] 0.41 0.41 0.45 0.45 0.36 0.34 0.34 0.33 0.37 0.42 0.47 0.48
```

This gives a vector of numbers representing the different cells in that column. We can use various functions such as `mean`, `sum`, and `length` to get information about a vector.

```
length(income$U.S.)
```

```
## [1] 12
```

```
mean(income$U.S.)
```

```
## [1] 0.4025
```

```
mean(income$Europe)
```

```
## [1] 0.3575
```

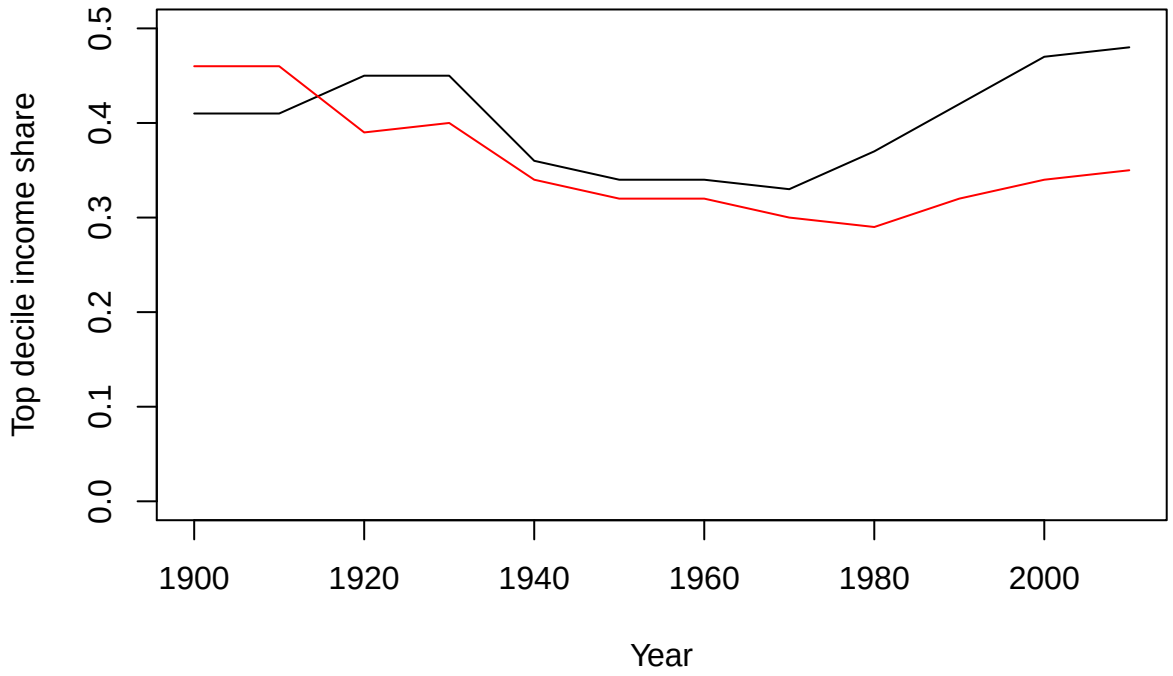
As perhaps expected, the mean income inequality in Europe is lower than than in the U.S.. Let's do a t-test to see if the difference is significant:

```
t.test(income$U.S., income$Europe, paired=T)
```

```
##
##  Paired t-test
##
## data:  income$U.S. and income$Europe
## t = 2.6146, df = 11, p-value = 0.02406
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.007119254 0.082880746
## sample estimates:
## mean of the differences
##                   0.045
```

So, with p<.05 we can conclude that the income distribution in the U.S. is more unequal than in Europe. Let's make a simple plot of the income inequality in the U.S. and Europe (reproducing fig 9.8 on page 324)

```
plot(x=income$Year, y=income$U.S., type="l", ylab="Top decile income share", xlab="Year", ylim=c(0, 0.5)
lines(x=income$Year, y=income$Europe, col="red")
```



As you can see, income distribution in pre-WWI Europe is actually more unequal than in the U.S., but this is reversed during the 1910's and inequality diverges after the 1970's. Still, the lines are probably correlated:

```r
cor.test(income$U.S., income$Europe)
```

```
##
##  Pearson's product-moment correlation
##
## data:  income$U.S. and income$Europe
## t = 1.4919, df = 10, p-value = 0.1666
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1949743  0.8037581
## sample estimates:
##     cor
## 0.42667
```

So, although the correlation is moderate at 0.43, it is not significant (due to a lack of data points)