```r
# Importing Data

# The first step for data management and alaysis is to have data. You may
# enter data using the keyboard, but, in most cases, you will import data file
# (from Excel, Text, Stata, Sas, etc) into R.

# 1. csv: comman separated values
getwd()
setwd("C:/Users/Min Seong Kim/Dropbox/R_programming/lecture/elsect_main")
dir()
rev_exp0 <- read.csv("district_rev_exp.csv")

rev_exp <- read.csv(file.choose(), stringsAsFactors = FALSE)  # district_rev_exp.csv

rev_exp$STATE <- as.factor(rev_exp$STATE)

class(rev_exp)
head(rev_exp)
tail(rev_exp)

str(rev_exp)
summary(rev_exp)

rev_exp$TOTALREV[rev_exp$TOTALREV == "-"] <- NA
rev_exp$TOTALREV <- as.numeric(rev_exp$TOTALREV)

# Calculate the averages of enrollment, total revenue and total expenditure for each state.
colMeans(rev_exp[rev_exp$STATE == "California",c("ENROLL", "TOTALREV", "TOTALEXP")], na.rm=TRUE)
colMeans(rev_exp[rev_exp$STATE == "Connecticut",c("ENROLL", "TOTALREV", "TOTALEXP")])
colMeans(rev_exp[rev_exp$STATE == "Massachusetts",c("ENROLL", "TOTALREV", "TOTALEXP")])
colMeans(rev_exp[rev_exp$STATE == "Missouri",c("ENROLL", "TOTALREV", "TOTALEXP")])

aggregate(rev_exp[ ,c("ENROLL", "TOTALREV", "TOTALEXP")], list(rev_exp$STATE), mean, na.rm=TRUE)
# list() specifies the criterion to make groups
aggregate(rev_exp[ ,c("ENROLL", "TOTALREV", "TOTALEXP")],
          by=list(ST = rev_exp$STATE, EnR = rev_exp$ENROLL > 1000), FUN=mean, na.rm=TRUE)

a <- na.omit(rev_exp)  # eleminate the rows that contains NA.

# 2. text file:  tab-delimited file
rev_exp1 <- read.delim(file.choose(), stringsAsFactors = FALSE)   # district_rev_exp.txt

head(rev_exp1)
tail(rev_exp1)

# 3. read.table: read any tabular file as a data.frame

# Use district_rev_exp_readtable.txt
rev_exp2 <- read.table(file.choose(), sep="/", stringsAsFactors = FALSE)
names(rev_exp2)
rev_exp2 <- read.table(file.choose(), sep="/", header = TRUE, stringsAsFactors = FALSE)
names(rev_exp2)

# We can also read csv file and tab delimited txt file using read.table
rev_exp3 <- read.table(file.choose(), sep=",", header = TRUE, stringsAsFactors = FALSE)   # district_rev_exp.csv
rev_exp3 <- read.table(file.choose(), sep="\t", header = TRUE, stringsAsFactors = FALSE)  # district_rev_exp.csv

# You can save excel file with csv or tab delimited txt file. Then, you can use the functions above
# to read the file.
# You can read excel file directly.
# First install the package "readxl"
install.packages(readxl)
```

```
library(readxl)
excel_sheets(file.choose())                          # list different sheets
rev_exp4 <- read_excel(file.choose(), sheet = 1)     # actually import data into R

# You can also import data from Stata
# You first install the package "foreign".
install.packages(foreign)
library(foreign)
read.dta(file.choose())

# Exercise ################################################################################


# Download complete.csv from HustkyCT (lecture10-data) in your computer.
# Import this dataset to R.

# 1. Which league is the best in terms of wage (eur_wage) and overall?
# 2. Based on "eur_value", which team has the most players in top 100?
#    hint: 1. Sort based on eur_value, 2. Make sure that team is a factor
# 3. Present the distribution of average wage (eur_wage) based on age.
#    hint: 1. Use aggregate() function to obtain average wages for each age, 2. Use plot() function
# 4. Which team has the most players under 23?
#    hint: Make sure that team is a factor


sc <- read.csv(file.choose())
mean(sc$eur_wage)
```