

Links in Problem 2 updated

Chapters/topics covered

- Review from R introduction. (R sources module on Canvas)
- Importing data into R (Ch 1 lecture notes)
- Summaries of categorical and quantitative variables (Ch 1 lecture notes).

In this lab handout, we will learn different ways to import data set. We use :

- `read.csv` for csv (comma separated value)
- `read.table` for txt files.

The data set, `movies2009`, contains information of 129 movies from 2009 ;

Total.Gross : Domestic total box office (in millions of dollars)

Opening : Box office during the opening weekend (in millions of dollars)

Screens: Number of screens played

RT : Rotten Tomatoes ratings

International : International box office (in millions of dollars)

DVD : Estimated sales from DVD

Budget : in millions of dollars

Ratings, Genre

0. Open R script and save using a proper name.

From page 6 and 7 from [R reference guide](#)

1. Import .csv using url address

Load and attach the data set using the following R command:

```
movies1<-read.csv("http://users.stat.umn.edu/~parky/movies2009.csv", header=T)
```

You might have to delete and type ~ again in R studio.

2. Import .csv from local computer

Download `movies2009` from Canvas or `http://users.stat.umn.edu/~parky/movies2009.csv`

Use the following to import :

```
movies2<-read.csv(file="~/Downloads/movies2009.csv", header=TRUE)
movies3<-read.csv(file.choose(), header=TRUE)
```

You might have to delete and type ~ again in R studio.

(Optional) 3. Checking and setting working directory then import the data by its name

You may check your working directory using :

```
getwd() #shows your working directory
setwd("your folder location")
```

Another (probably easier) way to change your working directory is

(i) click "Session" on the menu bar -> "Set working directory" -> "Choose working directory"

OR

(ii) Press [ctrl]+[shift]+[h], then set new working directory folder.

You can copy the command from console and paste it into R script. You can rerun the command next time.

You can import data files in the working directory by its name.

```
movies4<-read.csv("movies2009.csv")
```

4. Explore data set

Click the Data name under Environment to make sure the data file is imported correctly.

```
names(movies1) #names of all variables
head(movies1) #top 6 rows
tail(movies1) #bottom 6 rows
```

5. Explore distributions Total.Gross

Use the following commands to describe the distribution of total gross. What is the shape of the distribution?

Any outliers?

```
hist(movies1$Total.Gross) #histogram of Domestic box office
hist(movies1$Total.Gross, main="Histogram of Total Gross", xlab="in millions USD")
```

```
boxplot(movies1$Total.Gross)
```

6. Other useful commands to understand this data

```
table(movies1$Rating) #frequency table of Rating
mean(movies1$Total.Gross)
#####
# If there is a missing observation NA, use na.rm=TRUE (remove missing observations)
# to learn more, type '?mean' on Console
#####
mean(movies1$Total.Gross, trim=0.1) #drop top 10% and bottom 10%, then comput
median()
summary()
IQR()
sd()
max()
min()
```

Save your R script before closing R studio.