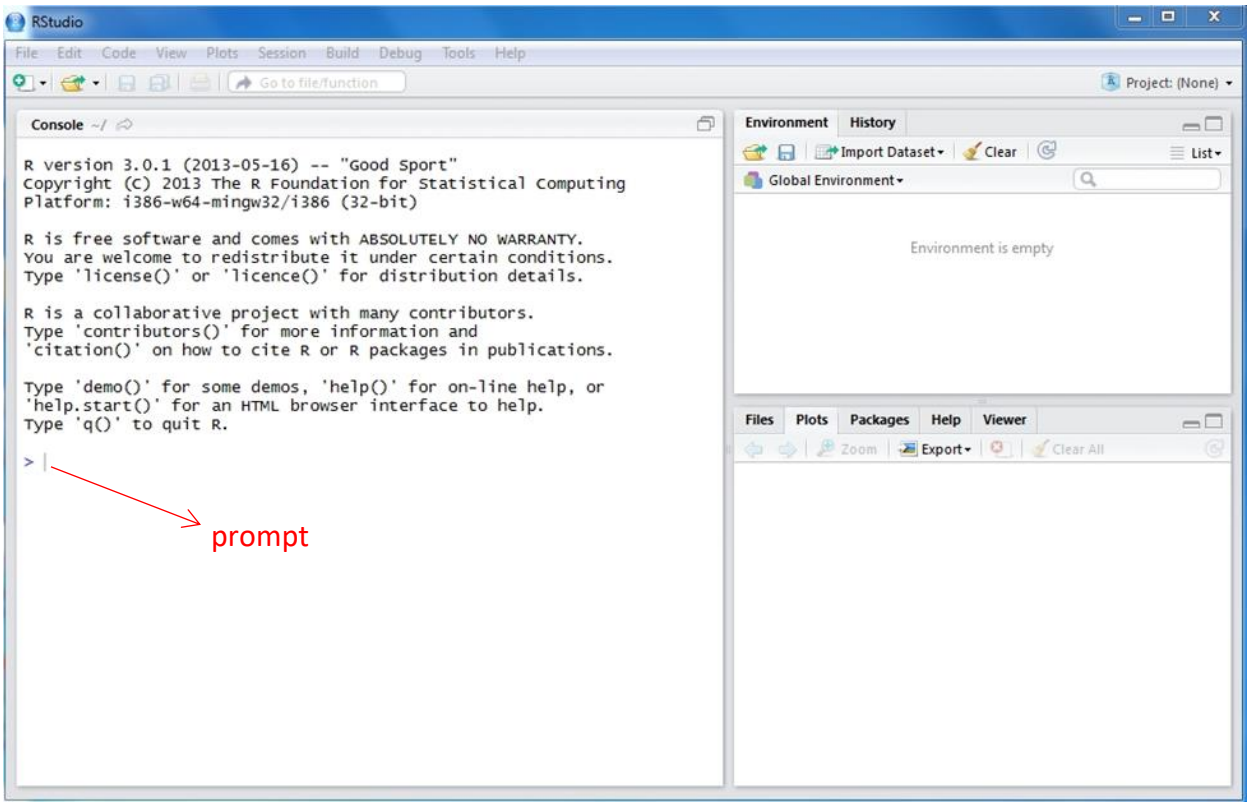
R Handout #1: Data Import and Basics

1. Start and Quit RStudio

- To start RStudio, just double-click on the RStudio icon.
- To quit, click on 'x' at top-right corner.



2. Simple Math Operations

Typing numbers with arithmetic operators will give the result of the operation. No space or many spaces can be used before or after arithmetic operators.

Arithmetic Operators

- Addition: +
- Subtraction: -
- Multiplication: *
- Division: /
- Exponent: ^

```
> 2 + 5
[1] 7

> 2^4*5
[1] 80

> (2 + 5)^2
[1] 49

> 3/40 + 2 -   5
[1] -2.925
```

Built-in Functions

- Square Root: `sqrt()`
- Absolute Value: `abs()`

Note: Function names are **case sensitive**.

```
> sqrt(4.2) + 7 / (-1)
[1] -4.95061

> abs(-3) * 5
[1] 15

> SQRT(4)
Error: could not find function "SQRT"
```

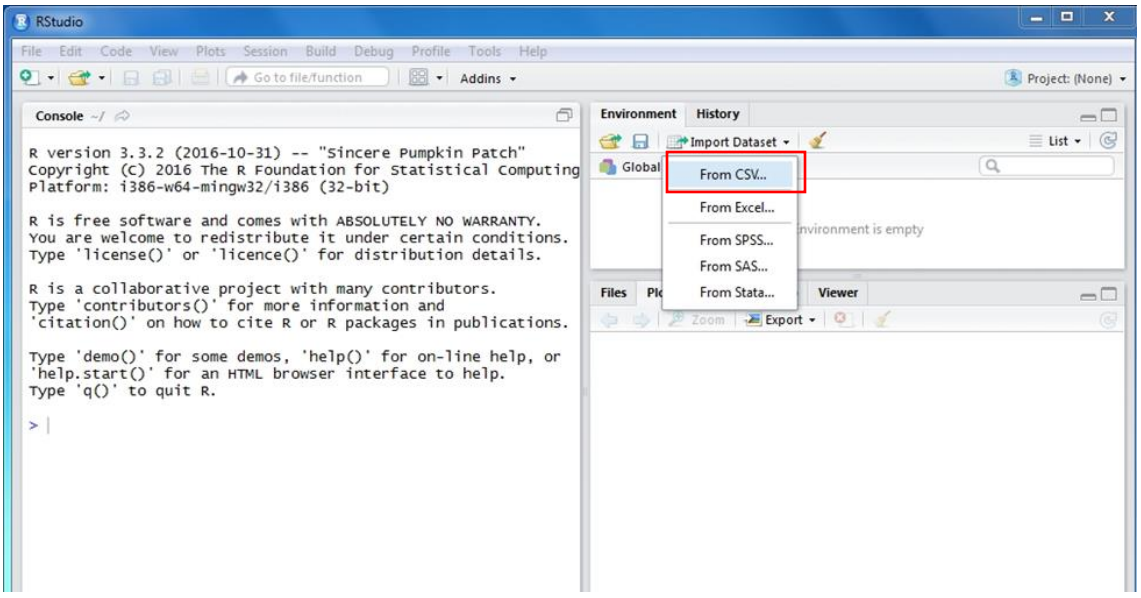
3. Data Import

< Ex > Data File: student (CSV file) A portion of the data file is shown below.

	A	B	C	D	E	F	G	H	I	J
1	GENDER	AGE	WORKHR	TATTOO	SHOES	EXERCISE	GPA	CLASS	CARAGE	MARRIED
2	FEMALE	18	20	NO	23	5	2.9	FRESHMAI	1	NO
3	FEMALE	18	16	NO	12	3	3.9	FRESHMAI	2	NO
4	FEMALE	18	20	NO	3	2	3.6	SOPHOMC	16	NO
5	FEMALE	18	20	NO	6	2	3.81	FRESHMAI	7	NO
6	FEMALE	18	16	NO	41	0	3.4	FRESHMAI	14	NO
7	FEMALE	18	12	NO	9	0	3.3	SOPHOMC	7	NO
8	FEMALE	18	20	NO	40	0	3	FRESHMAI	10	NO
9	FEMALE	18	12	NO	8	2	3.75	FRESHMAI	13	NO

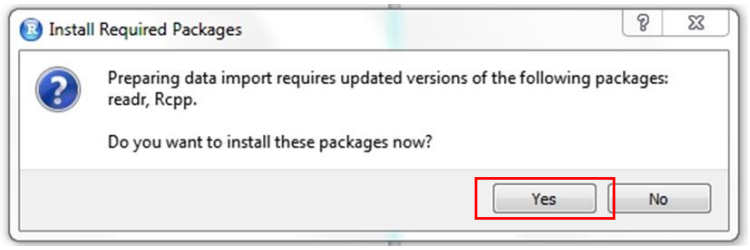
(1) First, save the file you want to work with **into your computer** to a directory of your choice.

(2) Click '**Import Dataset**' on the top right panel and then select '**From text (base)** Note: Depending on the version, it could be '**From CSV…**' or '**From Local File…**'

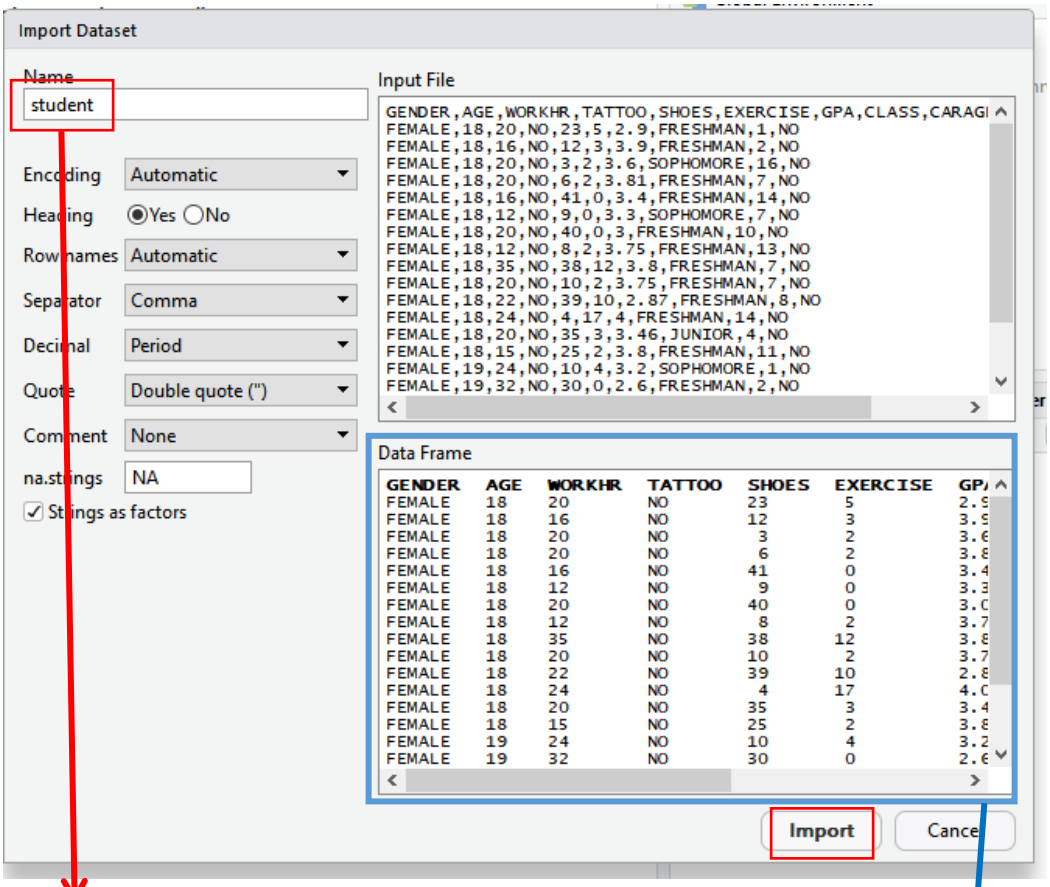


(3) Navigate the pup-up window to locate the data file and **double-click on the file**. Then, the following will appear. Click on the **"Import" button.**

(4) You may see the following message. Then, click "**Yes**" **Note**: This should appear only once when you use RStudio for the first time.



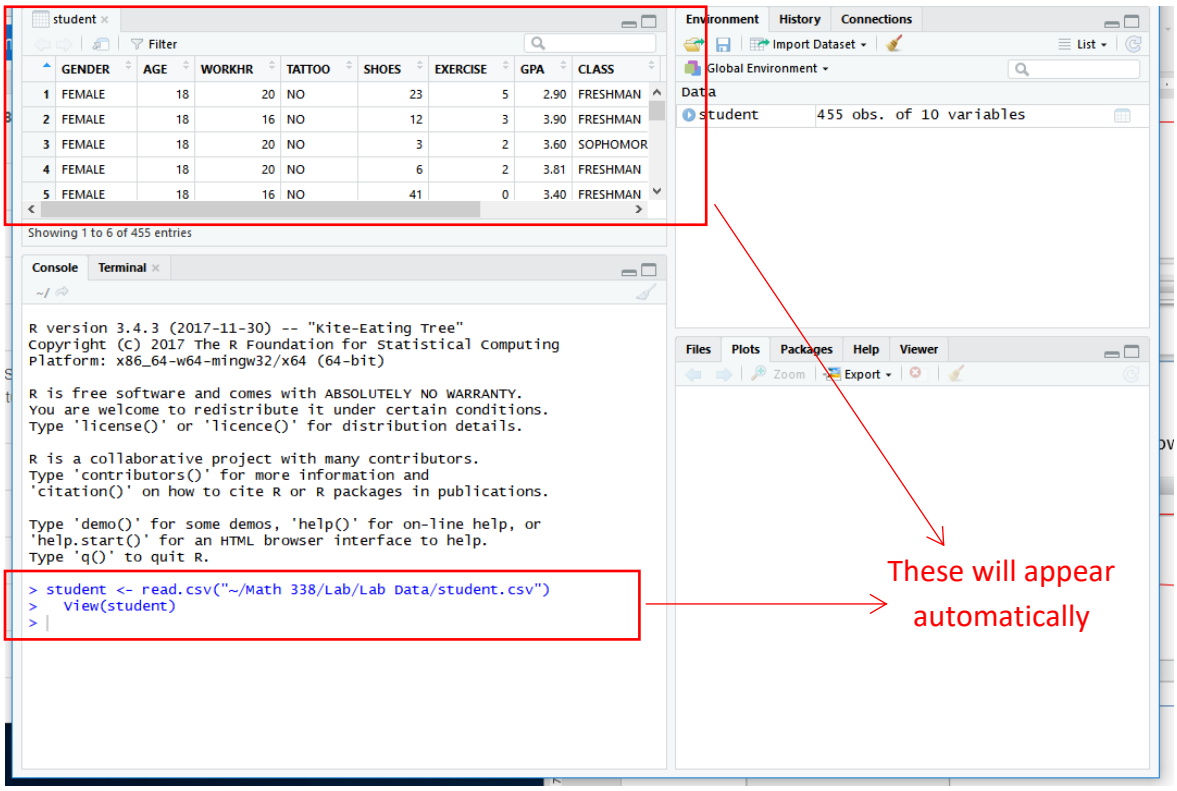
(5) When another window appears as shown below, you can do the following. Click "**Import**"



You can change it if you want to. This is the name R will use.

This will and should look like the Excel file.

(6) Now, you'll see the data on the top left panel. R calls this data a "**data frame**"



4. Access the Individual Variables (Vectors) from the Imported Data

(1) Variables (GENDER, AGE, etc) are NOT directly accessible by their names in R.

```
> GPA
Error: object 'GPA' not found
```

(2) First, we need to **attach the data set (data frame) using `attach()`**. Type the following.

```
> attach(student)
```

→ Name of data you imported

(3) Now, we can use the variable names**.**

```
> GPA
[1] 3.2 2.8 3.7 3.6 3.5
> MAJOR
[1] "Math" "Math" "Bio"  "CS"   "Bio"
> GENDER
[1] "F" "M" "M" "M" "F"
```

Note: Each **variable is a vector** (character vector or numeric vector). We've actually created 10 vectors (10 variables: GENDER, AGE, etc) by importing the data file.

(4) Variable names are **case sensitive**, so the variable 'GPA' would be different from 'gpa'

```
> gpa
Error: object 'gpa' not found
```

(5) We can compute some statistics using the variable names,

```
> mean(GPA)
[1] 3.36
> median(GPA)
[1] 3.5
> var(GPA)
[1] 0.133
> sd(GPA)
[1] 0.3646917
```

mean()	sample mean or average
median()	sample median
var()	sample variance
sd()	sample standard deviation
summary()	five number summary and mean
table()	Frequencies/counts

```
> summary(GPA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.80    3.20    3.50    3.36    3.60    3.70
> table(MAJOR)
MAJOR
 Bio   CS Math
   2    1    2
```

(6) We can compute some statistics by group,

```
tapply(X, INDEX, fun)
```

- X : vector (numerical variable)
- INDEX : a list of one or more factors (categorical variable), each of same length as X
- X : the function to be applied.

```
> tapply(GPA, GENDER, mean)
       F        M
3.350000 3.366667
> tapply(GPA, GENDER, median)
   F    M
3.35 3.60
> tapply(GPA, GENDER, sd)
        F         M
0.2121320 0.4932883
> tapply(GPA, GENDER, summary)
$F
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.200   3.275   3.350   3.350   3.425   3.500

$M
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.800   3.200   3.600   3.367   3.650   3.700
```

(7) When we are done, it's better to detach it using **detach()**

```
> detach(student)
```