Intro to R

Fall 2016

Intro to R

- R is a language and environment that allows:
  - – Data management
  - – Graphs and tables
  - – Statistical analyses
  - – You will need: some basic statistics
    - We will discuss these
- R is freeware that runs on Windows, Mac, Linux systems

R Environment

- R is an integrated software suite that includes:
  - Effective data handling
  - A suite of operators for array/matrix calculations
  - Intermediate tools for data analysis
  - Graphical facilities
  - Simple and effective programming language which includes conditionals, loops, functions, I/O

R

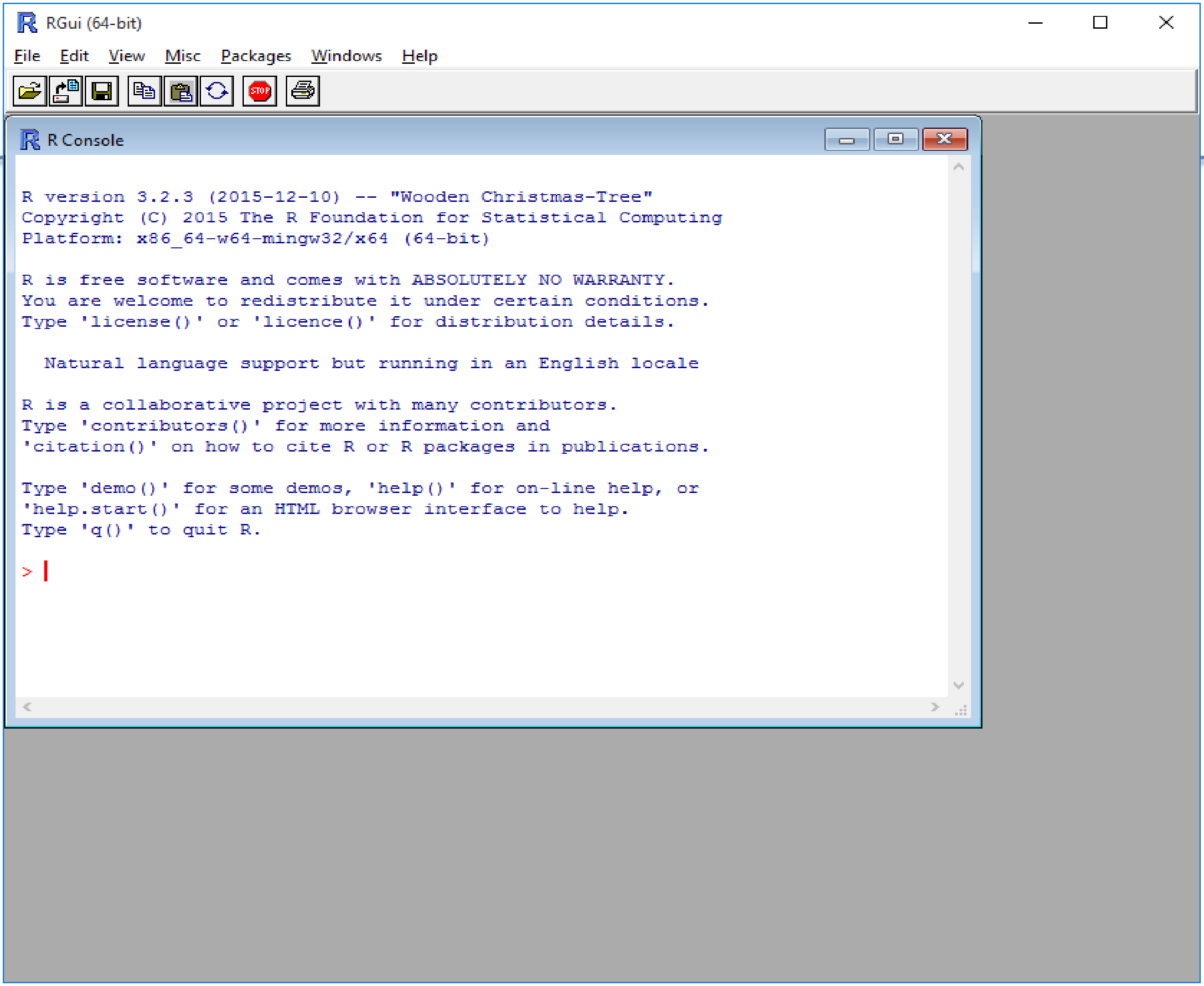
- Goals for this section of the course include:
  - Becoming familiar with Statistical Packages
  - Creating new Datasets
  - Importing & exporting Datasets
  - Manipulating data in a Dataset
  - Basic analysis of data (mainly descriptive statistics with some inferential statistics)
  - An overview of R's advanced features

Note: This is not a statistics course such as Math 207. We will only concentrate on basic statistical concepts.

R Resources

- Web site resources:
  - R console application only
    - https://cran.r-project.org/
  - Rstudio IDE
    - https://www.rstudio.com/products/rstudio/download/
  - R documentation
    - http://www.tutorialspoint.com/r/index.htm
    - http://www.cyclismo.org/tutorial/R/index.html

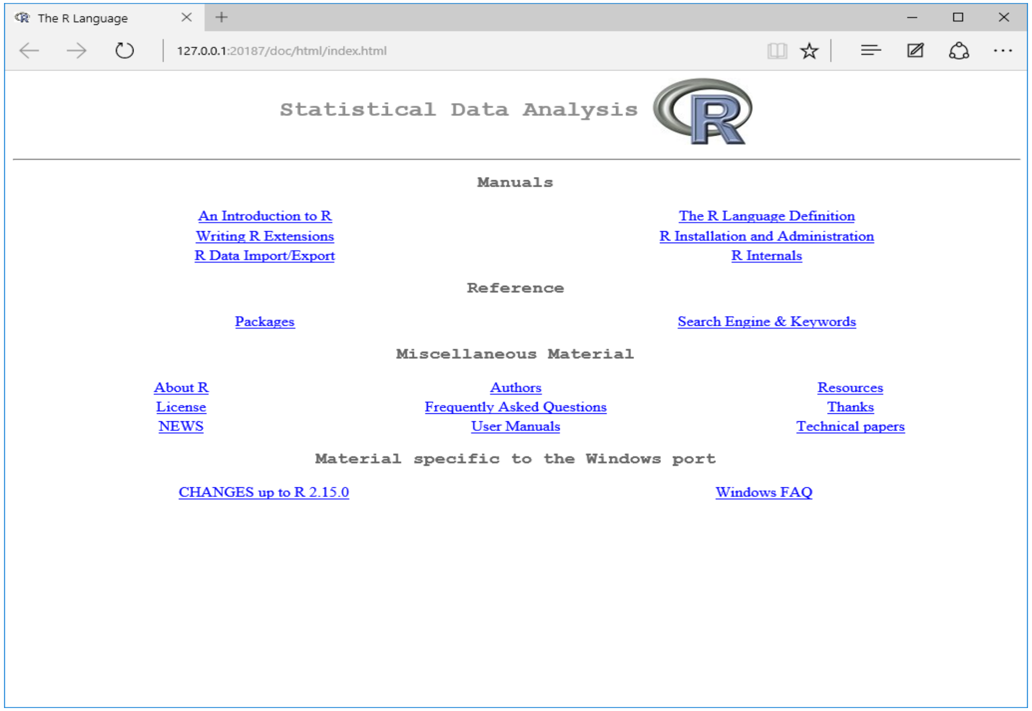
Open R Console Version



```
R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

R Help

- Type help.start() at the prompt in in R console



Statistical Data Analysis

Manuals

An Introduction to R
Writing R Extensions
R Data Import/Export
The R Language Definition
R Installation and Administration
R Internals

Reference

Packages
Search Engine & Keywords

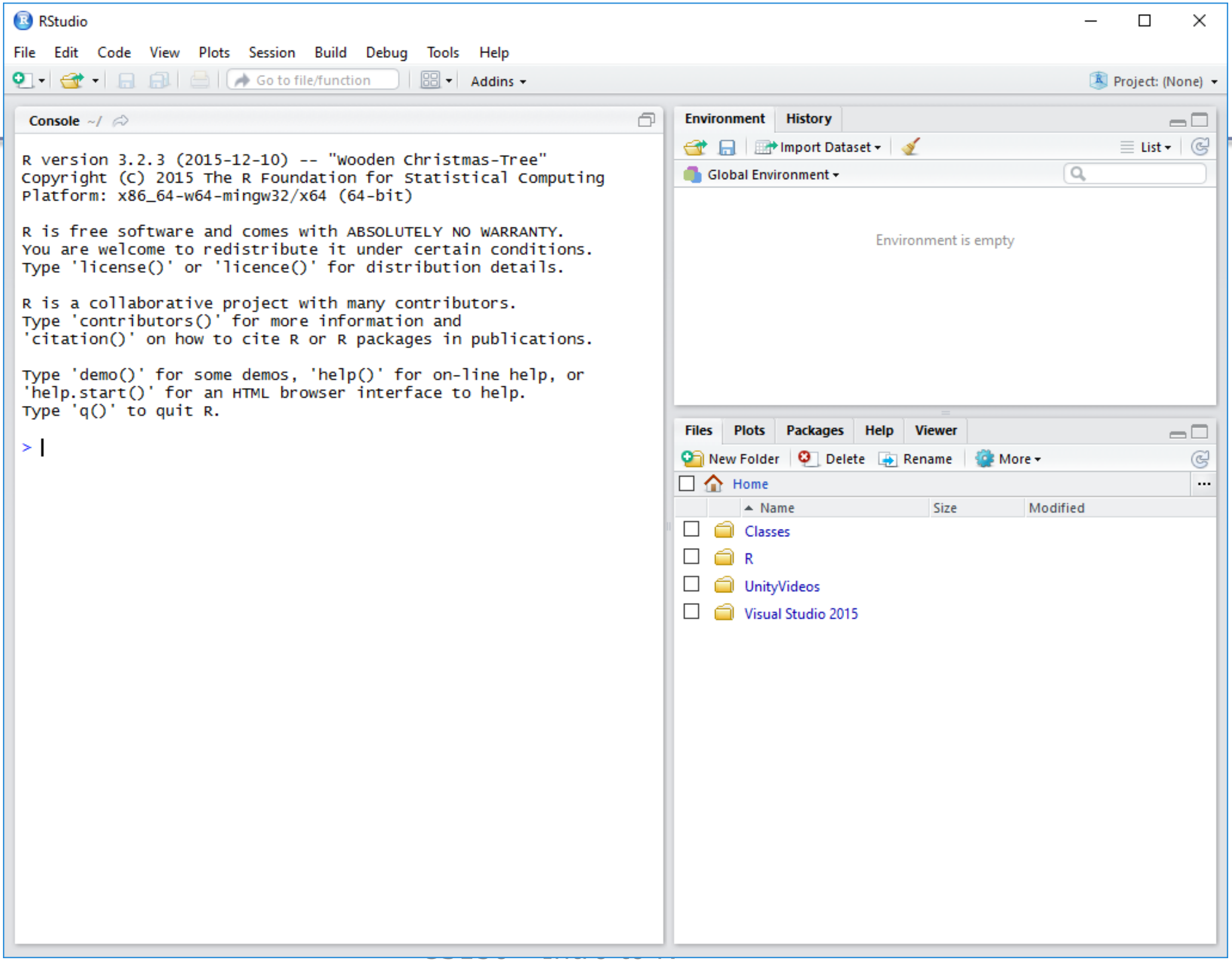
Miscellaneous Material

About R
License
NEWS
Authors
Frequently Asked Questions
User Manuals
Resources
Thanks
Technical papers

Material specific to the Windows port

CHANGES up to R 2.15.0
Windows FAQ

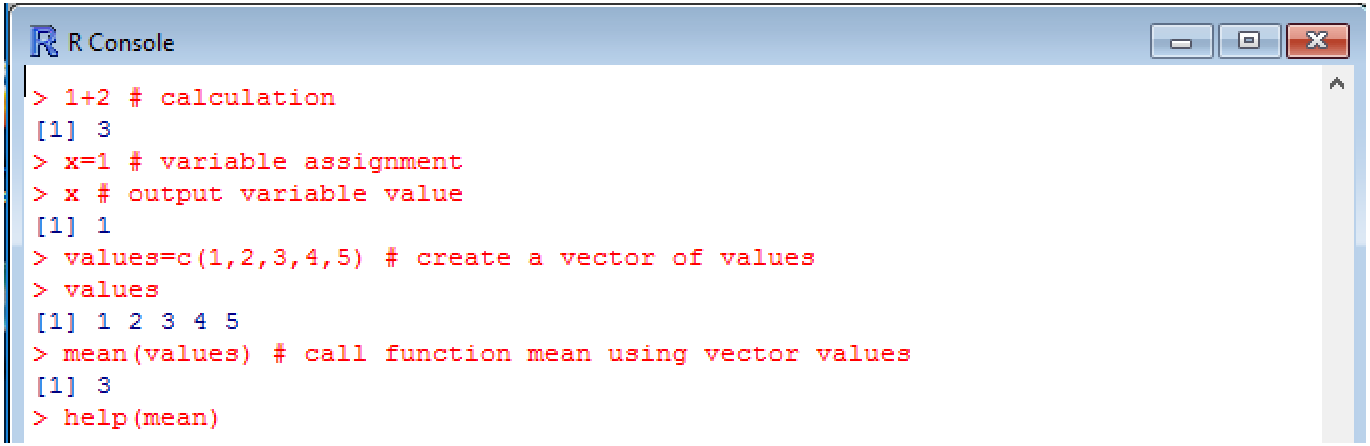
Open RStudio



```
R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

R Session

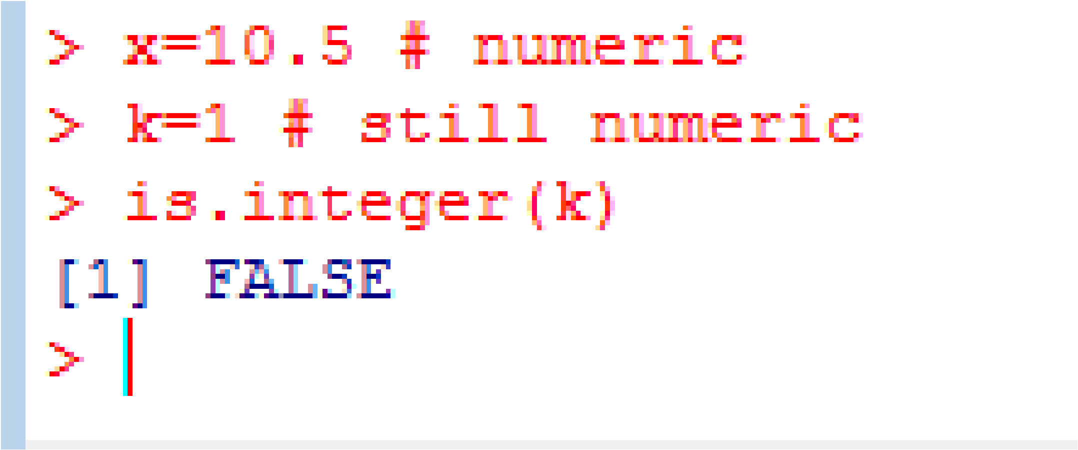
- Start an RStudio session
- We will use the console window of RStudio



```
> 1+2 # calculation
[1] 3
> x=1 # variable assignment
> x # output variable value
[1] 1
> values=c(1,2,3,4,5) # create a vector of values
> values
[1] 1 2 3 4 5
> mean(values) # call function mean using vector values
[1] 3
> help(mean)
```

Basic Datatypes
Numeric

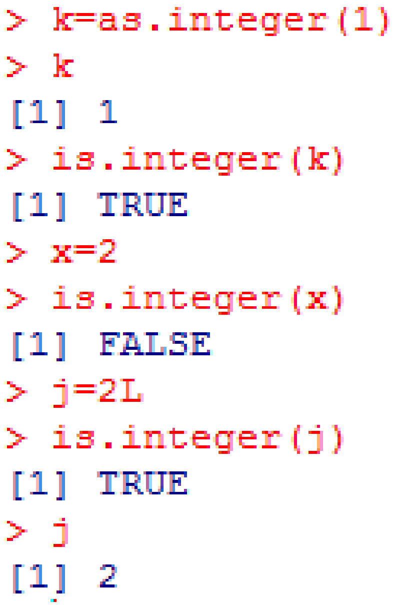
- Numeric – the “default” datatype where a value includes a decimal point



```
> x=10.5 # numeric
> k=1 # still numeric
> is.integer(k)
[1] FALSE
> 
```

Basic Datatypes
Integer

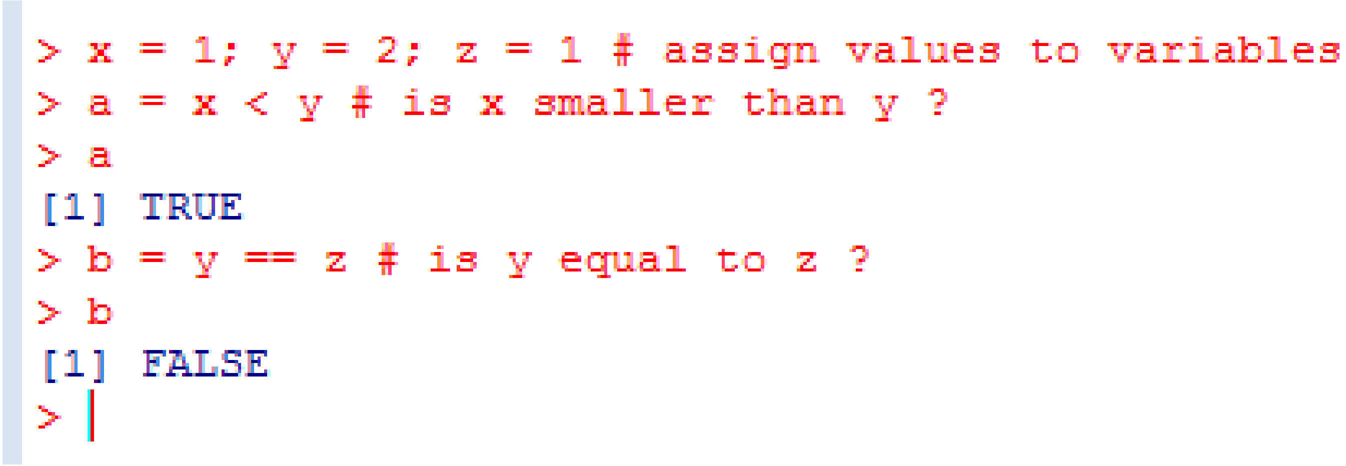
- Integer – does not include a decimal point and is created using as.integer () function or L as in 2L



```
> k=as.integer(1)
> k
[1] 1
> is.integer(k)
[1] TRUE
> x=2
> is.integer(x)
[1] FALSE
> j=2L
> is.integer(j)
[1] TRUE
> j
[1] 2
```

Basic Datatypes
Logical

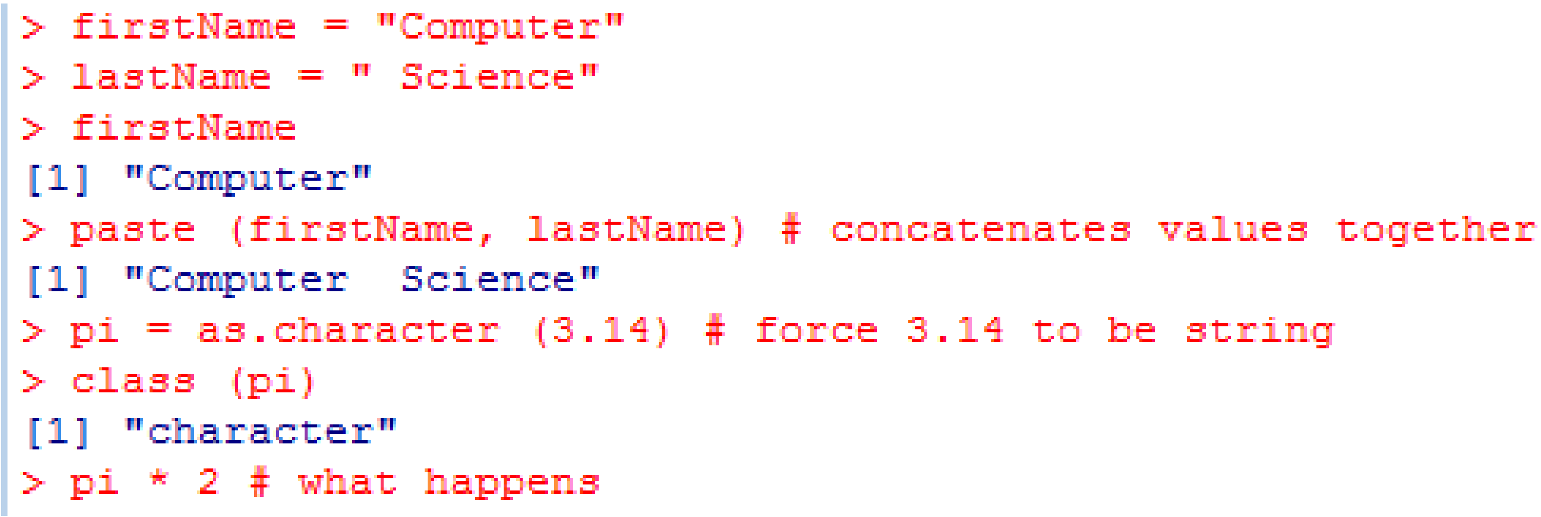
- Logical – is either TRUE or FALSE



```
> x = 1; y = 2; z = 1 # assign values to variables
> a = x < y # is x smaller than y ?
> a
[1] TRUE
> b = y == z # is y equal to z ?
> b
[1] FALSE
> |
```

Basic Datatypes
Character

- Character – is used to represent string values



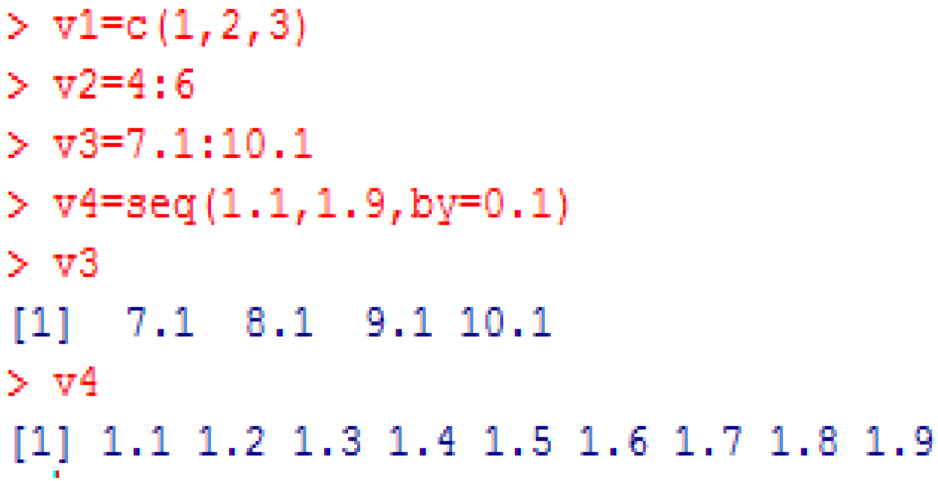
```
> firstName = "Computer"
> lastName = " Science"
> firstName
[1] "Computer"
> paste (firstName, lastName) # concatenates values together
[1] "Computer  Science"
> pi = as.character (3.14) # force 3.14 to be string
> class (pi)
[1] "character"
> pi * 2 # what happens
```

Measures of Central Tendency

- Used to describe the center of a distribution
- Define each of the following:
  - Mean
  - Median
  - Mode

Vector

- The most basic R data objects are called vectors.
- Six types of atomic vectors
  1. Logical
  2. Integer
  3. Double (Numeric)
  4. Character
  5. Complex
  6. Raw



```
> v1=c(1,2,3)
> v2=4:6
> v3=7.1:10.1
> v4=seq(1.1,1.9,by=0.1)
> v3
[1]  7.1  8.1  9.1 10.1
> v4
[1] 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9
```

- For now we will concern ourselves with 1-4.

Problems

- 1) Create a vector of ages in a variable called age with the following integer values: 18, 19, 18, 21, 22, 23, 19, 18

- 2) Compute the mean and median of the age values

- 3) Compute the mean of the first 1000 natural numbers

Problem

- Given the following dataset, find the mean, median, and mode of the Age variable using R

Breed	Age	Weight
Collie	2	23.2
Collie	3	35.7
Setter	5	45.4
Shepard	1	65.9
Setter	2	72.2

An R Solution

- First of all, what do we expect the answers to be?

- Let’s use R to check expected results:

1. Create a vector **age** with the Age values
2. Call function mean
3. Call function median
4. Call function mode

Did we get our expected results?

Data Frame

- A data frame is a two-dimensional (2D) structure where
  - column data refers to a variable
  - row data refers to an observation or a case
- Column names are to be unique non-empty.
- Row names are optional but should be unique.
- Allowable types of variable info: numeric, factor or character type.

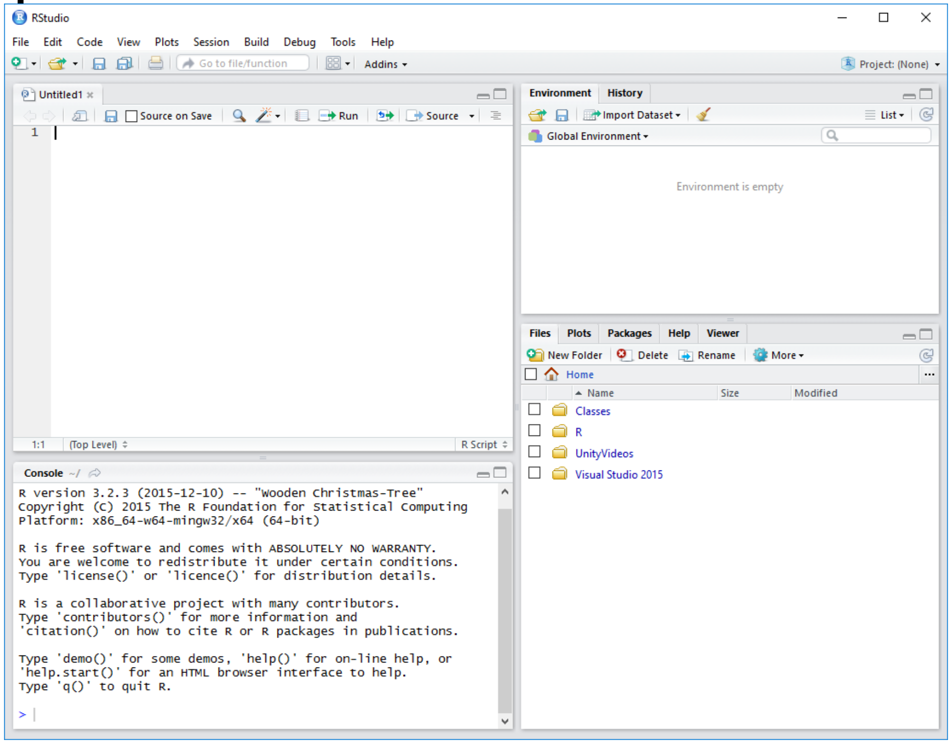
Dog Data Frame Example

- What type is Breed? Age? Weight?

Breed	Age	Weight
Collie	2	23.2
Collie	3	35.7
Setter	5	45.4
Shepard	1	65.9
Setter	2	72.2

Dog Data Frame

- We are going to start creating scripts in Rstudio
- File->New File->R Script



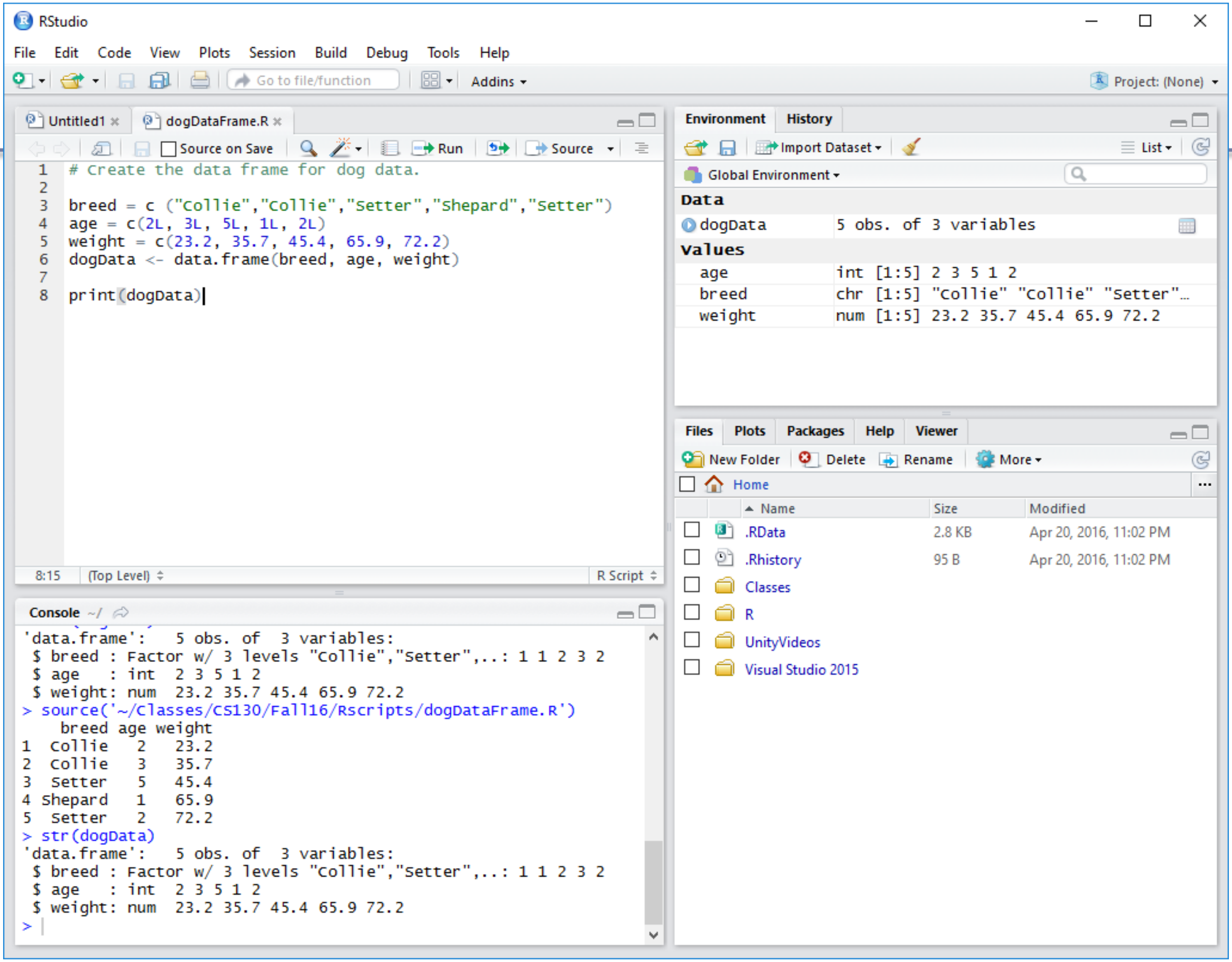
```
R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

Dog Data Frame

- In the Untitled script window, type the following R script

```
# Create the data frame for dog data.

breed = c("Collie","Collie","Setter","Shepard","Setter")
age = c(2L, 3L, 5L, 1L, 2L)
weight = c(23.2, 35.7, 45.4, 65.9, 72.2)
dogData <- data.frame(breed, age, weight)

print(dogData)
```

Execute the script



```
# Create the data frame for dog data.

breed = c ("Collie","Collie","Setter","Shepard","Setter")
age = c(2L, 3L, 5L, 1L, 2L)
weight = c(23.2, 35.7, 45.4, 65.9, 72.2)
dogData <- data.frame(breed, age, weight)

print(dogData)
```

```
Data
dogData     5 obs. of 3 variables
Values
  age       int [1:5] 2 3 5 1 2
  breed     chr [1:5] "Collie" "Collie" "Setter"…
  weight    num [1:5] 23.2 35.7 45.4 65.9 72.2
```

```
'data.frame':   5 obs. of  3 variables:
 $ breed : Factor w/ 3 levels "Collie","Setter",..: 1 1 2 3 2
 $ age   : int  2 3 5 1 2
 $ weight: num  23.2 35.7 45.4 65.9 72.2
> source('~/Classes/CS130/Fall16/Rscripts/dogDataFrame.R')
    breed age weight
1  Collie   2   23.2
2  Collie   3   35.7
3  Setter   5   45.4
4 Shepard   1   65.9
5  Setter   2   72.2
> str(dogData)
'data.frame':   5 obs. of  3 variables:
 $ breed : Factor w/ 3 levels "Collie","Setter",..: 1 1 2 3 2
 $ age   : int  2 3 5 1 2
 $ weight: num  23.2 35.7 45.4 65.9 72.2
>
```

	Name	Size	Modified
□	.RData	2.8 KB	Apr 20, 2016, 11:02 PM
□	.Rhistory	95 B	Apr 20, 2016, 11:02 PM
□	Classes		
□	R		
□	UnityVideos		
□	Visual Studio 2015		

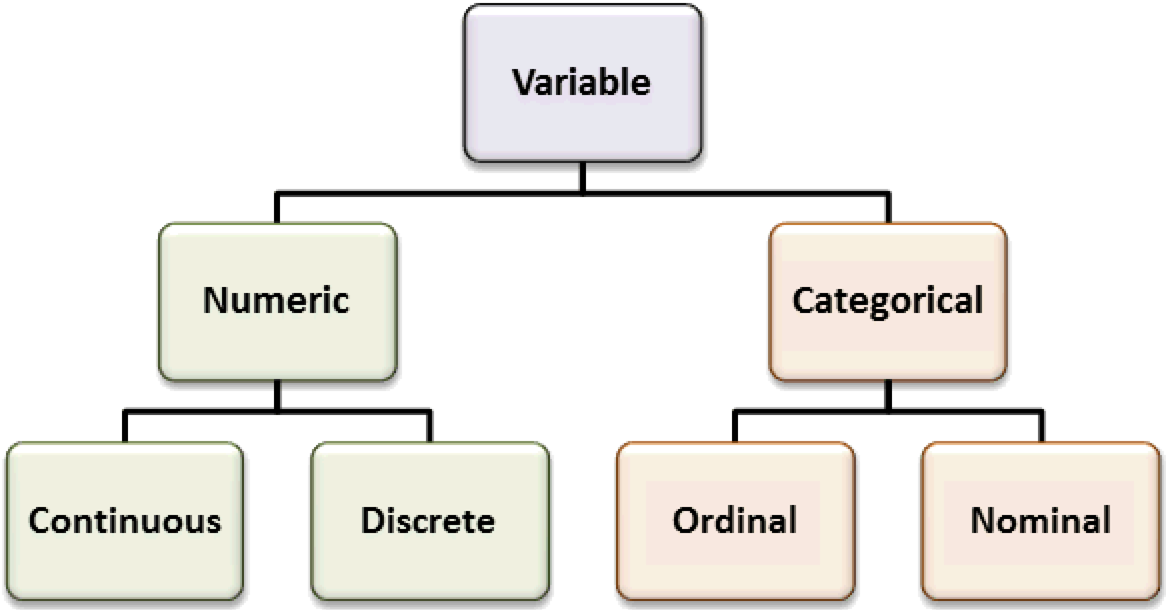
Problems

- Find the mean and median of the age and weight variables. Use the console window to do this. Hint: Variables of a Data Frame can be specified as dataframe$variable (e.g. dogData$age)

Variables in R

- Different statistical packages identify variables a little differently.
- Let's define the following terms used in statistics and give an example of each term
  - Variable
  - Categorical (or Qualitative) Variable
    - Nominal
    - Ordinal
  - Quantitative Variables
    - Numeric
      - Discrete
      - Continuous

Variables in R



Qualitative vs. Quantitative

- Qualitative: classify individuals into categories
- Quantitative: tell how much or how many of something there is

- Which are qualitative and which are quantitative?
  - Person’s Age
  - Person’s Gender
  - Mileage (in miles per gallon) of a car
  - Color of a car

Qualitative: Ordinal vs. Nominal

- Ordinal variables:
  - One whose categories have a natural ordering
  - Example: grades

- Nominal variables:
  - One whose categories have no natural ordering
  - Example: state of residence

Quantitative

- Discrete variables: Variables whose possible values can be listed
  - Example: number of children

- Continuous variables: Variables that can take any value in an interval
  - Example: height of a person

Problem

- Using the command str(dogData), identify:
  - variable name
  - quantitative or qualitative
  - discrete, continuous, neither
  - nominal, ordinal, neither
- A specific variable can be selected and passed to the class function. Pass the variable age of dogData to class. What does the result tell us?