RStudio and Basic R

UW Tacoma

Charles Lanfear

Apr 18, 2020

Updated: Apr 17, 2020

Overview

1. R and RStudio Orientation
2. Packages
3. Creating and Using Objects
4. Dataframes and Indexing
5. Basic Analyses
6. Resources for Further Learning

R and RStudio

A quick orientation

Why R?

R is a programming language built for statistical computing.

If one already knows Excel or Stata, why use R?

- R is *free*, so you don't need a terminal server or license.
- R has a *very* large community for support and packages.
- R can handle virtually any data format.
- R makes replication *easy*.
- R is a *language* so it can do *everything*.[1]
- R is similar to other programming languages.

[1] Including generate these slides (using RMarkdown)!

R Studio

R Studio is a "front-end" or integrated development environment (IDE) for R that can make your life *easier*.

RStudio can:

- Organize your code, output, and plots.
- Auto-complete code and highlight syntax.
- Help view data and objects.
- Enable easy integration of R code into documents.

Getting Started

Open up RStudio now and choose *File* > *New File* > *R Script*.

Then, let's get oriented with the interface:

- *Top Left*: Code **editor** pane, data viewer (browse with tabs)
- *Bottom Left*: **Console** for running code (`>` prompt)
- *Top Right*: List of objects in **environment**, code **history** tab.
- *Bottom Right*: Tabs for browsing files, viewing plots, managing packages, and viewing help files.

You can change the layout in *Preferences* > *Pane Layout*

Editing and Running Code

There are several ways to run R code in RStudio:

- Highlight lines in the **editor** window and click *Run* at the top or hit `Ctrl+Enter` or `⌘+Enter` to run them all.
- With your **caret** on a line you want to run, hit `Ctrl+Enter` or `⌘+Enter`. Note your caret moves to the next line, so you can run code sequentially with repeated presses.
- Type individual lines in the **console** and press `Enter`.

The console will show the lines you ran followed by any printed output.

Incomplete Code

If you mess up (e.g. leave off a parenthesis), R might show a `+` sign prompting you to finish the command:

```
> (11-2
+
```

Finish the command or hit `Esc` to get out of this.

R as a Calculator

In the **console**, type `123 + 456 + 789` and hit `Enter`.

```
123 + 456 + 789
```

```
## [1] 1368
```

The `[1]` in the output indicates the numeric **index** of the first element on that line.

Now in your blank R document in the **editor**, try typing the line `sqrt(400)` and either clicking *Run* or hitting `Ctrl+Enter` or `⌘+Enter`.

```
sqrt(400)
```

```
## [1] 20
```

Functions and Help

`sqrt()` is an example of a **function** in R.

If we didn't have a good guess as to what `sqrt()` will do, we can type `?sqrt` in the console and look at the **Help** panel on the right.

```
?sqrt
```

Arguments are the *inputs* to a function. In this case, the only argument to `sqrt()` is `x` which can be a number or a vector of numbers.

Help files provide documentation on how to use functions and what functions produce.

Creating Objects

R stores *everything* as an **object**, including data, functions, models, and output.

Creating an object can be done using the **assignment operator**: `<-`

```
new.object <- 144
```

Operators like `<-` are functions that look like symbols but typically sit between their arguments (e.g. numbers or objects) instead of having them inside `()` like in `sqrt(x)`[1].

We do math with operators, e.g., `x + y`. `+` is the addition operator!

[1] We can actually call operators like other functions by stuffing them between backticks: `+(x,y)`

Calling Objects

You can display or "call" an object simply by using its name.

```
new.object
```

```
## [1] 144
```

Object names can contain `_` and `.` in them, but cannot *begin* with numbers. Try to be consistent in naming objects. RStudio auto-complete means *long names are better than vague ones*!

Good names[1] *save confusion later.*

[1] "There are only two hard things in Computer Science: cache invalidation and naming things." - Phil Karlton

Using Objects

An object's **name** represents the information stored in that **object**, so you can treat the object's name as if it were the values stored inside.

```
new.object + 10
```

```
## [1] 154
```

```
new.object + new.object
```

```
## [1] 288
```

```
sqrt(new.object)
```

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

Creating Vectors

A **vector** is a series of **elements**, such as numbers.

You can create a vector and store it as an object in the same way. To do this, use the function `c()` which stands for "combine" or "concatenate".

```
new.object <- c(4, 9, 16, 25, 36)
new.object
```

```
## [1]  4  9 16 25 36
```

If you name an object the same name as an existing object, *it will overwrite it*.

You can provide a vector as an argument for many functions.

```
sqrt(new.object)
```

```
## [1] 2 3 4 5 6
```

Character Vectors

We often work with data that are categorical. To create a vector of text elements—**strings** in programming terms—we must place the text in quotes:

```
string.vector <- c("Atlantic", "Pacific", "Arctic")
string.vector
```

```
## [1] "Atlantic" "Pacific"  "Arctic"
```

Categorical data can also be stored as a **factor**, which has an underlying numeric representation. Models will convert factors to dummies.[1]

```
factor.vector <- factor(string.vector)
factor.vector
```

```
## [1] Atlantic Pacific  Arctic
## Levels: Arctic Atlantic Pacific
```

[1] Factors have **levels** which you can use to set a reference category in models using `relevel()`.

Saving and Loading Objects

You can save an R object on your computer as a file to open later:

```
save(new.object, file="new_object.RData")
```

You can open saved files in R as well:

```
load("new_object.RData")
```

But where are these files being saved and loaded from?

Working Directories

R saves files and looks for files to open in your current **working directory**[1]. You can ask R what this is:

```
getwd()
```

```
## [1] "C:/Users/cclan/OneDrive/GitHub/r_exposure_workshop/lectures/r1"
```

Similarly, we can set a working directory like so:

```
setwd("C:/Users/cclan/Documents")
```

[1] For a simple R function to open an Explorer / Finder window at your working directory, see this StackOverflow response.

More Complex Objects

The same principles shown with vectors can be used with more complex objects like **matrices**, **arrays**, **lists**, and **dataframes** (lists which look like matrices but can hold multiple data types at once).

Most data sets you will work with will be read into R and stored as a **dataframe**, so the remainder of this workshop will mainly focus on using these objects.

Loading Dataframes

Delimited Text Files

The easiest way to work with external data—that isn't in R format—is for it to be stored in a *delimited* text file, e.g. comma-separated values (**.csv**) or tab-separated values (**.tsv**).

R has a variety of built-in functions for importing data stored in text files, like `read.table()` and `read.csv()`.[1]

By default, these functions will read *character* (string) columns in as a *factor*.

To disable this, use the argument `stringsAsFactors = FALSE`, like so:

```r
new_df <- read.csv("some_spreadsheet.csv", stringsAsFactors = FALSE)
```

[1] Use "write" versions (e.g. `write.csv()`) to create these files from R objects.

Data from Other Software

Working with **Stata**, **SPSS**, or **SAS** users? You can use a **package** to bring in their saved data files:

- `foreign`
  - Part of base R
  - Functions: `read.spss()`, `read.dta()`, `read.xport()`
  - Less complex but sometimes loses some metadata
- `haven`
  - Part of the `tidyverse` family
  - Functions: `read_spss()`, `read_dta()`, `read_sas()`
  - Keeps metadata like variable labels

For less common formats, Google it. I've yet to encounter a data format without an R package to handle it (or at least a clever hack).

If you encounter an ambiguous file extension (e.g. `.dat`), try opening it with a good text editor first (e.g. Atom, Sublime); there's a good chance it is actually raw text with a delimiter or fixed format that R can handle!

Installing Packages

Packages contain functions (and sometimes data) created by the community. The real power of R is found in add-on packages!

This workshop focuses on using packages from the [`tidyverse`].

The `tidyverse` is a collection of R packages which share a design philosophy, syntax, and data structures.

The `tidyverse` includes the most used packages in the R world: [`dplyr`] and [`ggplot2`]

You can install the *entire* `tidyverse` with the following:

```
install.packages("tidyverse")
```

We will also use the `gapminder` and `nycflights13` datasets:

```
install.packages("gapminder")
install.packages("nycflights13")
```

Loading Packages

To load a package, use `library()`:

```
library(gapminder)
```

Once a package is loaded, you can call on functions or data inside it.

```
data(gapminder) # Places data in your global environment
head(gapminder) # Displays first six elements of an object
```

```
## # A tibble: 6 x 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       1952    28.8  8425333      779.
## 2 Afghanistan Asia       1957    30.3  9240934      821.
## 3 Afghanistan Asia       1962    32.0 10267083      853.
## 4 Afghanistan Asia       1967    34.0 11537966      836.
## 5 Afghanistan Asia       1972    36.1 13079460      740.
## 6 Afghanistan Asia       1977    38.4 14880372      786.
```

Indexing and Subsetting

Base R

Indices and Dimensions

In base R, there are two main ways to access elements of objects: square brackets (`[]` or `[[]]`) and `$`. How you access an object depends on its *dimensions*.

Dataframes have *2* dimensions: **rows** and **columns**. Square brackets allow us to numerically **subset** in the format of `object[row, column]`. Leaving the row or column place empty selects *all* elements of that dimension.

```
gapminder[1,] # First row
```

```
## # A tibble: 1 x 6
##   country     continent  year lifeExp     pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>   <int>     <dbl>
## 1 Afghanistan Asia       1952    28.8 8425333      779.
```

```
gapminder[1:3, 3:4] # First three rows, third and fourth column
```

```
## # A tibble: 3 x 2
##    year lifeExp
##   <int>   <dbl>
## 1  1952    28.8
## 2  1957    30.3
## 3  1962    32.0
```

The **colon operator** (`:`) generates a vector using the sequence of integers from its first argument to its second. `1:3` is equivalent to `c(1,2,3)`.

Dataframes and Names

Columns in dataframes can also be accessed using their names with the `$` extract operator. This will return the column as a vector:

```
gapminder$gdpPercap[1:10]
```

```
##  [1] 779.4453 820.8530 853.1007 836.1971 739.9811 786.1134 978.0114 852.39
```

Note here I *also* used brackets to select just the first 10 elements of that column.

You can mix subsetting formats! In this case I provided only a single value (no column index) because **vectors** have *only one dimension* (length).

If you try to subset something and get a warning about "incorrect number of dimensions", check your subsetting!

Indexing by Expression

We can also index using expressions—logical *tests*.

```
gapminder[gapminder$year==1952, ]
```

```
## # A tibble: 142 x 6
##    country     continent  year lifeExp      pop gdpPercap
##    <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
##  1 Afghanistan Asia       1952    28.8  8425333      779.
##  2 Albania     Europe     1952    55.2  1282697     1601.
##  3 Algeria     Africa     1952    43.1  9279525     2449.
##  4 Angola      Africa     1952    30.0  4232095     3521.
##  5 Argentina   Americas   1952    62.5 17876956     5911.
##  6 Australia   Oceania    1952    69.1  8691212    10040.
##  7 Austria     Europe     1952    66.8  6927772     6137.
##  8 Bahrain     Asia       1952    50.9   120447     9867.
##  9 Bangladesh  Asia       1952    37.5 46886859      684.
## 10 Belgium     Europe     1952    68    8730405     8343.
## # ... with 132 more rows
```

How Expressions Work

What does `gapminder$year==1952` actually do?

```
head(gapminder$year==1952, 50) # display first 50 elements
```

```
##  [1]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FAL
## [23] FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FAL
## [45] FALSE FALSE FALSE FALSE  TRUE FALSE
```

It returns a vector of `TRUE` or `FALSE` values.

When used with the subset operator (`[]`), elements for which a `TRUE` is given are returned while those corresponding to `FALSE` are dropped.

Logical Operators

We used `==` for testing "equals": `gapminder$year==1952`.

There are many other logical operators:

- `!=`: not equal to
- `>`, `>=`, `<`, `<=`: less than, less than or equal to, etc.
- `%in%`: used with checking equal to one of several values

Or we can combine multiple logical conditions:

- `&`: both conditions need to hold (AND)
- `|`: at least one condition needs to hold (OR)
- `!`: inverts a logical condition (`TRUE` becomes `FALSE`, `FALSE` becomes `TRUE`)

Logical operators are one of the foundations of programming. You should experiment with these to become familiar with how they work!

Sidenote: Missing Values

Missing values are coded as `NA` entries without quotes:

```
vector_w_missing <- c(1, 2, NA, 4, 5, 6, NA)
```

Even one `NA` "poisons the well": You'll get `NA` out of your calculations unless you remove them manually or use the extra argument `na.rm = TRUE` in some functions:

```
mean(vector_w_missing)
```

```
## [1] NA
```

```
mean(vector_w_missing, na.rm=TRUE)
```

```
## [1] 3.6
```

Finding Missing Values

WARNING: You can't test for missing values by seeing if they "equal" (`==`) `NA`:

```
vector_w_missing == NA
```

```
## [1] NA NA NA NA NA NA NA
```

But you can use the `is.na()` function:

```
is.na(vector_w_missing)
```

```
## [1] FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE
```

We can use subsetting to get the equivalent of `na.rm=TRUE`:

```
mean(vector_w_missing[!is.na(vector_w_missing)])
```

```
## [1] 3.6
```

`!` *reverses* a logical condition. Read the above as "subset *not* `NA`"

Subsetting Data with `dplyr`

`dplyr`

`dplyr` is a Tidyverse package for working with data frames.

It provides an intuitive, powerful, and consistent alternative to base R for subsetting data.

It also provides functions for summarizing and joining data which are more straightforward than base R.

While I recommend all users be familiar with base R methods I've just covered, `dplyr` is the dominant platform for data manipulation in R, so we will focus on it for the remainder of this unit.

But First, Pipes: `%>%`

`dplyr` uses the `magrittr` forward pipe operator, usually called simply a **pipe**. We write pipes like `%>%` (`Ctrl+Shift+M` or `⌘ +Shift+M`).

Pipes take the object on the *left* and apply the function on the *right*: `x %>% f(y) = f(x, y)`. Read out loud: "and then..."

```
library(dplyr)
gapminder %>% filter(country == "Canada") %>% head(2)
```

```
## # A tibble: 2 x 6
##   country continent  year lifeExp      pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Canada  Americas   1952    68.8 14785584    11367.
## 2 Canada  Americas   1957    70.0 17010154    12490.
```

Pipes save us typing, make code readable, and allow chaining like above, so we use them *all the time* when manipulating data frames.

Using Pipes

Pipes are clearest to read when you have each function on a separate line.

```
take_this_data %>%
    do_first_thing(with = this_value) %>%
    do_next_thing(using = that_value) %>% ...
```

Stuff to the left of the pipe is passed to the *first argument* of the function on the right. Other arguments go on the right in the function.

If you ever find yourself piping a function where data are not the first argument, use `.` in the data argument instead.

```
gapminder %>% lm(pop ~ year, data = .)
```

Pipe Assignment

When creating a new object from the output of piped functions, you place the assignment operator *at the beginning*.

```
lm_pop_year <- gapminder %>%
  lm(pop ~ year, data = .)
```

No matter how long the chain of functions is, assignment is always done *at the top*.

`filter()` Data Frames

I used `filter()` earlier. We subset *rows* of data using logical conditions with `filter()`!

```
gapminder %>% filter(country == "Oman") %>% head(8)
```

```
## # A tibble: 8 x 6
##   country continent  year lifeExp     pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>   <int>     <dbl>
## 1 Oman    Asia       1952    37.6  507833     1828.
## 2 Oman    Asia       1957    40.1  561977     2243.
## 3 Oman    Asia       1962    43.2  628164     2925.
## 4 Oman    Asia       1967    47.0  714775     4721.
## 5 Oman    Asia       1972    52.1  829050    10618.
## 6 Oman    Asia       1977    57.4 1004533    11848.
## 7 Oman    Asia       1982    62.7 1301048    12955.
## 8 Oman    Asia       1987    67.7 1593882    18115.
```

What is this doing?

Multiple Conditions Example

Let's say we want observations from Oman after 1980 and through 2000.

```
gapminder %>%
    filter(country == "Oman" &
           year > 1980 &
           year <= 2000 )
```

```
## # A tibble: 4 x 6
##   country continent  year lifeExp     pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>   <int>     <dbl>
## 1 Oman    Asia       1982    62.7 1301048    12955.
## 2 Oman    Asia       1987    67.7 1593882    18115.
## 3 Oman    Asia       1992    71.2 1915208    18617.
## 4 Oman    Asia       1997    72.5 2283635    19702.
```

`%in%` Operator

Common use case: Filter rows to things in some *set*.

We can use `%in%` like `==` but for matching *any element* in the vector on its right[1].

```
former_yugoslavia <- c("Bosnia and Herzegovina", "Croatia",
                       "Macedonia", "Montenegro", "Serbia", "Slovenia")
yugoslavia <- gapminder %>% filter(country %in% former_yugoslavia)
tail(yugoslavia, 2)
```

```
## # A tibble: 2 x 6
##   country  continent  year lifeExp     pop gdpPercap
##   <fct>    <fct>     <int>   <dbl>   <int>     <dbl>
## 1 Slovenia Europe     2002    76.7 2011497    20660.
## 2 Slovenia Europe     2007    77.9 2009245    25768.
```

[1] The `c()` function is how we make **vectors** in R, which are an important data type.

Sorting: `arrange()`

Along with filtering the data to see certain rows, we might want to sort it:

```
yugoslavia %>% arrange(year, desc(pop))
```

```
## # A tibble: 60 x 6
##    country                continent  year lifeExp     pop gdpPercap
##    <fct>                  <fct>     <int>   <dbl>   <int>     <dbl>
##  1 Serbia                 Europe     1952    58.0 6860147     3581.
##  2 Croatia                Europe     1952    61.2 3882229     3119.
##  3 Bosnia and Herzegovina Europe     1952    53.8 2791000      974.
##  4 Slovenia               Europe     1952    65.6 1489518     4215.
##  5 Montenegro             Europe     1952    59.2  413834     2648.
##  6 Serbia                 Europe     1957    61.7 7271135     4981.
##  7 Croatia                Europe     1957    64.8 3991242     4338.
##  8 Bosnia and Herzegovina Europe     1957    58.4 3076000     1354.
##  9 Slovenia               Europe     1957    67.8 1533070     5862.
## 10 Montenegro             Europe     1957    61.4  442829     3682.
## # ... with 50 more rows
```

The data are sorted by ascending `year` and descending `pop`.

Keeping Columns: `select()`

Not only can we subset rows, but we can include specific columns (and put them in the order listed) using `select()`.

```
yugoslavia %>% select(country, year, pop) %>% head(4)
```

```
## # A tibble: 4 x 3
##   country                 year     pop
##   <fct>                  <int>   <int>
## 1 Bosnia and Herzegovina  1952 2791000
## 2 Bosnia and Herzegovina  1957 3076000
## 3 Bosnia and Herzegovina  1962 3349000
## 4 Bosnia and Herzegovina  1967 3585000
```

Dropping Columns: `select()`

We can instead drop only specific columns with `select()` using `-` signs:

```
yugoslavia %>% select(-continent, -pop, -lifeExp) %>% head(4)
```

```
## # A tibble: 4 x 3
##   country                 year gdpPercap
##   <fct>                  <int>     <dbl>
## 1 Bosnia and Herzegovina  1952      974.
## 2 Bosnia and Herzegovina  1957     1354.
## 3 Bosnia and Herzegovina  1962     1710.
## 4 Bosnia and Herzegovina  1967     2172.
```

Helper Functions for `select()`

`select()` has a variety of helper functions like `starts_with()`, `ends_with()`, and `contains()`, or can be given a range of continguous columns `startvar:endvar`. See `?select` for details.

These are very useful if you have a "wide" data frame with column names following a pattern or ordering.

```
# A tibble: 6 x 292
  married10 married11 married12 married13 married14 married15 married16 married17 married18 married19 married20
      <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
1        NA        NA         0         0         0         0         0         0         0         0         0
2        NA        NA        NA        NA         0         0         0         0         1         1        NA
3        NA        NA         0        NA         0         0         0         0         0         0        NA
4        NA        NA        NA        NA         0         0         0         0         0         0         0
5        NA        NA         0         0         0         0         0         0         0         0         0
6        NA        NA         0         0         0         0         0         0         0         0         0
# ... with 281 more variables: married21 <dbl>, married22 <dbl>, married23 <dbl>, married24 <dbl>,
#   married25 <dbl>, married26 <dbl>, in_school10 <dbl>, in_school11 <dbl>, in_school12 <dbl>, in_school13 <dbl>,
#   in_school14 <dbl>, in_school15 <dbl>, in_school16 <dbl>, in_school17 <dbl>, in_school18 <dbl>,
#   in_school19 <dbl>, in_school20 <dbl>, in_school21 <dbl>, in_school22 <dbl>, in_school23 <dbl>,
```

```r
DYS %>% select(starts_with("married"))
DYS %>% select(ends_with("18"))
```

Renaming Columns with `select()`

We can rename columns using `select()`, but that drops everything that isn't mentioned:

```
yugoslavia %>%
    select(Life_Expectancy = lifeExp) %>%
    head(4)
```

```
## # A tibble: 4 x 1
##   Life_Expectancy
##             <dbl>
## 1            53.8
## 2            58.4
## 3            61.9
## 4            64.8
```

Safer: Rename Columns with `rename()`

`rename()` renames variables using the same syntax as `select()` without dropping unmentioned variables.

```
yugoslavia %>%
    select(country, year, lifeExp) %>%
    rename(Life_Expectancy = lifeExp) %>%
    head(4)
```

```
## # A tibble: 4 x 3
##   country                 year Life_Expectancy
##   <fct>                  <int>           <dbl>
## 1 Bosnia and Herzegovina  1952            53.8
## 2 Bosnia and Herzegovina  1957            58.4
## 3 Bosnia and Herzegovina  1962            61.9
## 4 Bosnia and Herzegovina  1967            64.8
```

Creating Variables

mutate()

In `dplyr`, you can add new columns to a data frame using **`mutate()`**.

```
yugoslavia %>% filter(country == "Serbia") %>%
    select(year, pop, lifeExp) %>%
    mutate(pop_million = pop / 1000000,
           life_exp_past_40 = lifeExp - 40) %>%
    head(5)
```

```
## # A tibble: 5 x 5
##    year     pop lifeExp pop_million life_exp_past_40
##   <int>   <int>   <dbl>       <dbl>            <dbl>
## 1  1952 6860147    58.0        6.86             18.0
## 2  1957 7271135    61.7        7.27             21.7
## 3  1962 7616060    64.5        7.62             24.5
## 4  1967 7971222    66.9        7.97             26.9
## 5  1972 8313288    68.7        8.31             28.7
```

Note you can create multiple variables in a single `mutate()` call by separating the expressions with commas.

`ifelse()`

A common function used in `mutate()` (and in general in R programming) is **`ifelse()`**. It returns a vector of values depending on a logical test.

```
ifelse(test = x==y, yes = first_value , no = second_value)
```

Output from `ifelse()` if `x==y` is...

- `TRUE`: `first_value` - the value for `yes =`
- `FALSE`: `second_value` - the value for `no =`
- `NA`: `NA` - because you can't test for NA with an equality!

For example:

```
example <- c(1, 0, NA, -2)
ifelse(example > 0, "Positive", "Not Positive")
```

```
## [1] "Positive"     "Not Positive" NA             "Not Positive"
```

`ifelse()` Example

```
yugoslavia %>% mutate(short_country =
                ifelse(country == "Bosnia and Herzegovina",
                       "B and H", as.character(country))) %>%
    select(short_country, year, pop) %>%
    arrange(year, short_country) %>%
    head(3)
```

```
## # A tibble: 3 x 3
##   short_country  year     pop
##   <chr>         <int>   <int>
## 1 B and H        1952 2791000
## 2 Croatia        1952 3882229
## 3 Montenegro     1952  413834
```

Read this as "For each row, if country equals 'Bosnia and Herzegovina', make `short_country` equal to 'B and H', otherwise make it equal to that row's value of `country`."

This is a simple way to change some values but not others!

`case_when()`

`case_when()` performs multiple `ifelse()` operations at the same time. `case_when()` allows you to create a new variable with values based on multiple logical statements. This is useful for making categorical variables or variables from combinations of other variables.

```
gapminder %>%
  mutate(gdpPercap_ordinal =
    case_when(
      gdpPercap <  700 ~ "low",
      gdpPercap >= 700 & gdpPercap < 800 ~ "moderate",
      TRUE ~ "high" )) %>% # Value when all other statements are FALSE
  slice(6:9) # get rows 6 through 9
```

```
## # A tibble: 4 x 7
##   country     continent  year lifeExp      pop gdpPercap gdpPercap_ordinal
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl> <chr>
## 1 Afghanistan Asia       1977    38.4 14880372      786. moderate
## 2 Afghanistan Asia       1982    39.9 12881816      978. high
## 3 Afghanistan Asia       1987    40.8 13867957      852. high
## 4 Afghanistan Asia       1992    41.7 16317921      649. low
```

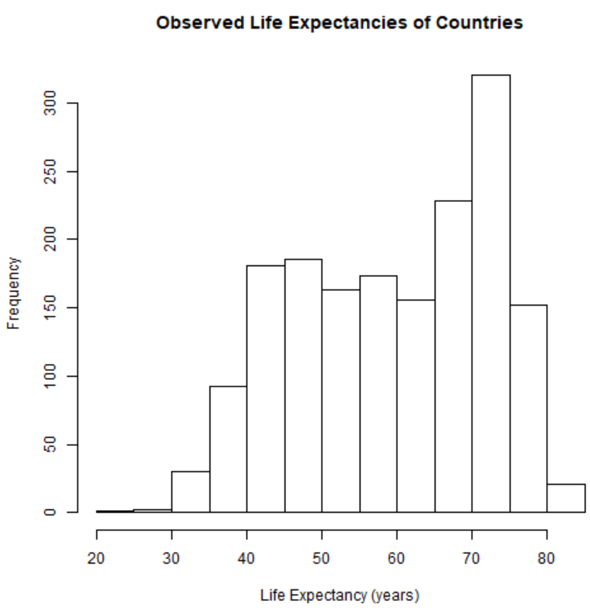
Analyses

Basic Graphics and Models

Histograms

We can use the `hist()` function to generate a histogram of a vector:

```
hist(gapminder$lifeExp,
     xlab = "Life Expectancy (years)",
     main = "Observed Life Expectancies of Countries")
```



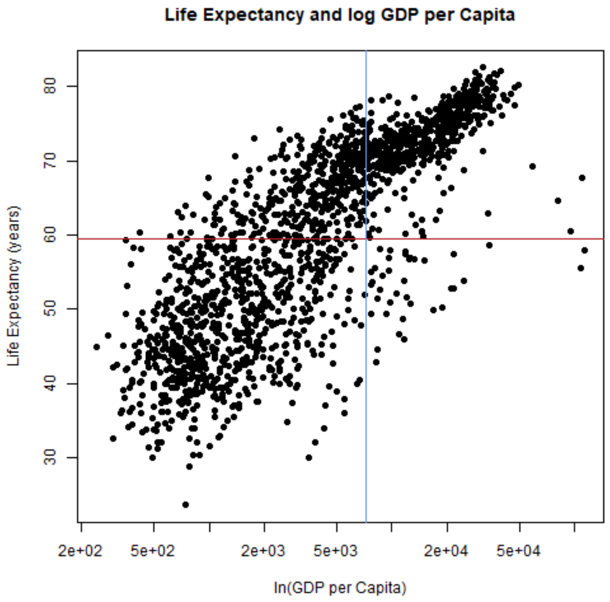
`xlab =` is used to set the label of the x-axis of a plot.

`main =` is used to set the title of a plot.

Use `?hist` to see additional options available for customizing a histogram.

Scatter Plots

```
plot(lifeExp ~ gdpPercap, data = gapminder,
     xlab = "ln(GDP per Capita)",
     ylab = "Life Expectancy (years)",
     main = "Life Expectancy and log GDP per Capita",
     pch = 16, log="x") # log="x" sets x axis to log scale!
abline(h = mean(gapminder$lifeExp), col = "firebrick")
abline(v = mean(gapminder$gdpPercap), col = "cornflowerblue")
```



Note that `lifeExp ~ gdpPercap` is a **formula** of the type `y ~ x`. The first element (`lifeExp`) gets plotted on the y-axis and the second (`gdpPercap`) goes on the x-axis.

The `abline()` calls place horizontal (`h =`) or vertical (`v =`) lines at the means of the variables used in the plot.

Formulae

Most modeling functions in R use a common formula format—the same seen with the previous plot:

```
new_formula <- y ~ x1 + x2 + x3
new_formula
```

```
## y ~ x1 + x2 + x3
## <environment: 0x0000023edf097488>
```

```
class(new_formula)
```

```
## [1] "formula"
```

The dependent variable goes on the left side of `~` and independent variables go on the right.

See here for more on formulae.

Simple Tables

`table()` creates basic cross-tabulations of vectors.

```
table(mtcars$cyl, mtcars$am)
```

```
##
##       0  1
##   4   3  8
##   6   4  3
##   8  12  2
```

Chi-Square

We can give the output from `table()` to `chisq.test()` to perform a Chi-Square test of assocation.

```
chisq.test(table(mtcars$cyl, mtcars$am))
```

```
## Warning in chisq.test(table(mtcars$cyl, mtcars$am)): Chi-squared approxima

##
##     Pearson's Chi-squared test
##
## data:  table(mtcars$cyl, mtcars$am)
## X-squared = 8.7407, df = 2, p-value = 0.01265
```

Note the warning here. You can use rescaled (`rescale.p=TRUE`) or simulated p-values (`simulate.p.value=TRUE`) if desired.

T Tests

T tests for mean comparisons are simple to do.

```
gapminder$post_1980 <- ifelse(gapminder$year > 1980, 1, 2)
t.test(lifeExp ~ post_1980, data=gapminder)
```

```
##
##  Welch Two Sample t-test
##
## data:  lifeExp by post_1980
## t = 17.174, df = 1694.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   8.791953 11.059068
## sample estimates:
## mean in group 1 mean in group 2
##        64.43719        54.51168
```

Linear Models

We can run an ordinary least squares linear regression using `lm()`:

```
lm(lifeExp~pop + gdpPercap + year + continent, data=gapminder)
```

```
##
## Call:
## lm(formula = lifeExp ~ pop + gdpPercap + year + continent, data = gapminde
##
## Coefficients:
##       (Intercept)                pop          gdpPercap               year
##        -5.185e+02          1.791e-09          2.985e-04          2.863e-01
##  continentOceania
##         2.056e+01
```

Note we get a lot less output here than you may have expected! This is because we're only viewing a tiny bit of the information produced by `lm()`. We need to expore the object `lm()` creates!

Model Summaries

The `summary()` function provides Stata-like regression output:

```
lm_out <- lm(lifeExp~pop + gdpPercap + year + continent, data=gapminder)
summary(lm_out)
```

```
##
## Call:
## lm(formula = lifeExp ~ pop + gdpPercap + year + continent, data = gapminder)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.4051  -4.0550   0.2317   4.5073  20.0217
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -5.185e+02  1.989e+01 -26.062   <2e-16 ***
## pop                 1.791e-09  1.634e-09   1.096    0.273
## gdpPercap           2.985e-04  2.002e-05  14.908   <2e-16 ***
## year                2.863e-01  1.006e-02  28.469   <2e-16 ***
## continentAmericas   1.429e+01  4.946e-01  28.898   <2e-16 ***
## continentAsia       9.375e+00  4.719e-01  19.869   <2e-16 ***
## continentEurope     1.936e+01  5.182e-01  37.361   <2e-16 ***
## continentOceania    2.056e+01  1.469e+00  13.995   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.883 on 1696 degrees of freedom
## Multiple R-squared:  0.7172,    Adjusted R-squared:  0.716
## F-statistic: 614.5 on 7 and 1696 DF,  p-value: < 2.2e-16
```

Model Objects

`lm()` produces a lot more information than what is shown by `summary()` however. We can see the **str**ucture of `lm()` output using `str()`:

```
str(lm_out)
```

```
## List of 13
##  $ coefficients : Named num [1:8] -5.18e+02 1.79e-09 2.98e-04 2.86e-01 1.43e+01 ...
##   ..- attr(*, "names")= chr [1:8] "(Intercept)" "pop" "gdpPercap" "year" ...
##  $ residuals    : Named num [1:1704] -21.1 -21.1 -20.8 -20.2 -19.6 ...
##   ..- attr(*, "names")= chr [1:1704] "1" "2" "3" "4" ...
##  $ effects      : Named num [1:1704] -2455.1 34.6 312.1 162.6 100.6 ...
##   ..- attr(*, "names")= chr [1:1704] "(Intercept)" "pop" "gdpPercap" "year" ...
##  $ rank         : int 8
##  $ fitted.values: Named num [1:1704] 49.9 51.4 52.8 54.3 55.7 ...
##   ..- attr(*, "names")= chr [1:1704] "1" "2" "3" "4" ...
##  $ assign       : int [1:8] 0 1 2 3 4 4 4 4
##  $ qr           :List of 5
##   ..$ qr   : num [1:1704, 1:8] -41.2795 0.0242 0.0242 0.0242 0.0242 ...
##   .. ..- attr(*, "dimnames")=List of 2
##   .. ..- attr(*, "assign")= int [1:8] 0 1 2 3 4 4 4 4
##   .. ..- attr(*, "contrasts")=List of 1
##   ..$ qraux: num [1:8] 1.02 1 1.02 1.01 1.01 ...
##   ..$ pivot: int [1:8] 1 2 3 4 5 6 7 8
##   ..$ tol  : num 1e-07
##   ..$ rank : int 8
##   ..- attr(*, "class")= chr "qr"
##  [list output truncated]
##  - attr(*, "class")= chr "lm"
```

`lm()` actually has an enormous quantity of output! This is a type of object called a **list**.

Model Objects

We can access parts of `lm()` output using `$` like with dataframe names:

```
lm_out$coefficients
```

```
##       (Intercept)                 pop           gdpPercap                year continentAmericas     conti
##     -5.184555e+02        1.790640e-09        2.984892e-04        2.862583e-01        1.429204e+01       9.37
##   continentOceania
##      2.055921e+01
```

We can also do this with `summary()`, which provides additional statistics:

```
summary(lm_out)$coefficients
```

```
##                        Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)       -5.184555e+02 1.989299e+01 -26.062215 3.248472e-126
## pop                1.790640e-09 1.634107e-09   1.095791  2.733256e-01
## gdpPercap          2.984892e-04 2.002178e-05  14.908225  2.522143e-47
## year               2.862583e-01 1.005523e-02  28.468586 4.800797e-146
## continentAmericas  1.429204e+01 4.945645e-01  28.898241 1.183161e-149
## continentAsia      9.375486e+00 4.718629e-01  19.869087  3.798275e-79
## continentEurope    1.936120e+01 5.182170e-01  37.361177 2.025551e-223
## continentOceania   2.055921e+01 1.469070e+00  13.994707  3.390781e-42
```

ANOVA

ANOVAs can be fit and summarized just like `lm()`

```
summary(aov(lifeExp ~ continent, data=gapminder))
```

```
##                   Df Sum Sq Mean Sq F value Pr(>F)    
## continent          4 139343   34836   408.7 <2e-16 ***
## Residuals       1699 144805      85                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

More Complex Models

R supports many more complex models, for example:

- `glm()` has syntax similar to `lm()` but adds a `family =` argument to specify model families and link functions like logistic regression
  - ex: `glm(x~y, family=binomial(link="logit"))`
- The `lme4` package adds hierarchical (multilevel) GLM models.
- `lavaan` fits structural equation models with intuitive syntax.
- `plm` and `tseries` fit time series models.

Most of these other packages support mode summaries with `summary()` and all create output objects which can be accessed using `$`.

Because R is the dominant environment for statisticians, the universe of modeling tools in R is *enormous*. If you need to do it, it is probably in a package somewhere.

End of Unit 1