R Lesson 5

Importing Data

- See help on read.table for more info
- R recommends converting Excel, etc. to delimited text if possible
- Know thy data
- Use \\ or / rather than \ in Windows path
- Remember to use the entire file path and not just the file name.
- read.csv – for comma separated data
- read.delim – for tab delimited data

General Form of read Function

read.csv(*file*, *header* = TRUE, *sep* = ",", *quote*="\"", *dec*=".")

- *file* – file path or URL to data
- *header* – does first line contain field names
- *sep* – field separator character
- *quote* – set of quoting characters
- *dec* – character used for decimal points

Example of read function

- Read csv-like file that uses pipe as delimiter and * for missing data
  read.csv("c:/data/pipedata.csv", sep="|", na.strings="*")

Character Strings

- Use single (') or double (") quotes to mark strings, but don’t mix:

```
x <- 'good'
y <- "no'
z <- "it's working"
```

String Functions

- Create a vector of character strings
  s <- c('apple','bee','cars','danish','egg')
- Get the number of characters in each string
  nchar(s)
- Convert all letters to upper case
  toupper(s)

String Functions

- Extract or replace substrings in a character vector.
- General forms:

  substr(*x, start, stop*)

  substr(*x, start, stop*) <- *value*

  - *x* – a character vector
  - *start* – first element of substring
  - *stop* – last element of substring
  - *value* – character vector to replace original values

Substring Function

- Examples:
- Extract the first to third characters
  substr(s,1,3)
- Replace first and second characters
  substr(s,1,2) <- ‘BU’
- How can we use this in our logical test?
  names(cpi)>'y2005’
  substr(names(cpi),2,5) > ‘2005’

String Functions

- Replace (substitute) specific values
- sub – replaces first occurrence in string
- gsub – replaces all occurrences (globally)
- General form:

  sub(*pattern, replacement, x*)

  - *pattern* – string to be replaced
  - *replacement* – new values for matched pattern
  - *x* – character vector where matches are sought

Substitute Functions

- Examples:
- Replace first 'e' with '_ '
  sub('e', '_', s)
- Globally replace every 'e' with '_ '
  gsub('e', '_', s)

Working with Text

- Find characters in a string
- grep – return indices or values of strings where matches are found
- grepl – (logic) return TRUE or FALSE to indicate where matches are found

General Form of grep Functions

grep(*pattern*, *x*, *ignore.case* = FALSE, *value* = FALSE)

grepl(*pattern*, *x*, *ignore.case* = FALSE)

- *pattern* – what you are looking for (add ^ to search only at beginning)
- *x* – vector being searched
- *ignore.case* – control case sensitivity
- *value* – returns indices of matches instead of actual value

Examples of grep Functions

- Search s for instances of 'e'
  grep('e', s)
- Return list of TRUE/FALSE instead of indices
  grepl('e', s)
- Search using a regular expression
  grep('^e', s)
- Return values instead of indices
  grep("Z", cpi$Country, value=TRUE)

Reordering Values

- sort – returns actual values in desired order
- order – returns vector of indices in new order
- order is most useful when working with tabular data like data frames

General Form of sort/order

order(..., *na.last* = TRUE, *decreasing* = FALSE)
sort(..., *na.last* = NA, *decreasing* = FALSE)

- … a sequence of numeric, complex, character or logical vectors, all of the same length, or a classed **R** object.
- *na.last*= – where to put missing values (NA removes them)
- *decreasing*= – increasing order by default

Reordering Values - Examples

- Show country names in ascending order
  sort(cpi$Country)
- Return index values for descending order
  order(cpi$Country, decreasing=TRUE)
- Use ordered index values with other data
  cpi[order(cpi$AvgCPI),c(1,12)]

Date Values

- Date objects in R have class
- Date – days since January 1, 1970
- **`POSIXct`** – seconds since January 1, 1970 with time and time zone
- **`POSIXlt`** – list of date and time components
- Default input = year-month-day hour: minute:second
- difftime – time interval between two dates

Date Functions

- as.Date – converts to date
- as.POSIXct – converts to date/time
- as.POSIXlt – converts to date/time list
- strptime – removes time from date/time
- strftime – converts date/time to character

Selected Format Codes

- strptime('14Sep2021 05:30:00 PM',
  format='%d%b%Y %I:%M:%S %p')
- %d – day number
- %m – month number
- %b – abbreviated month
- %B – month name
- %y – 2 digit year
- %Y – 4 digit year

Selected Format Codes

- `%H – hours (24)`
- `%I – hours (12)`
- `%M – minutes`
- `%S – seconds`
- `%p – AM/PM`
- slash, dash, space, colon are literals