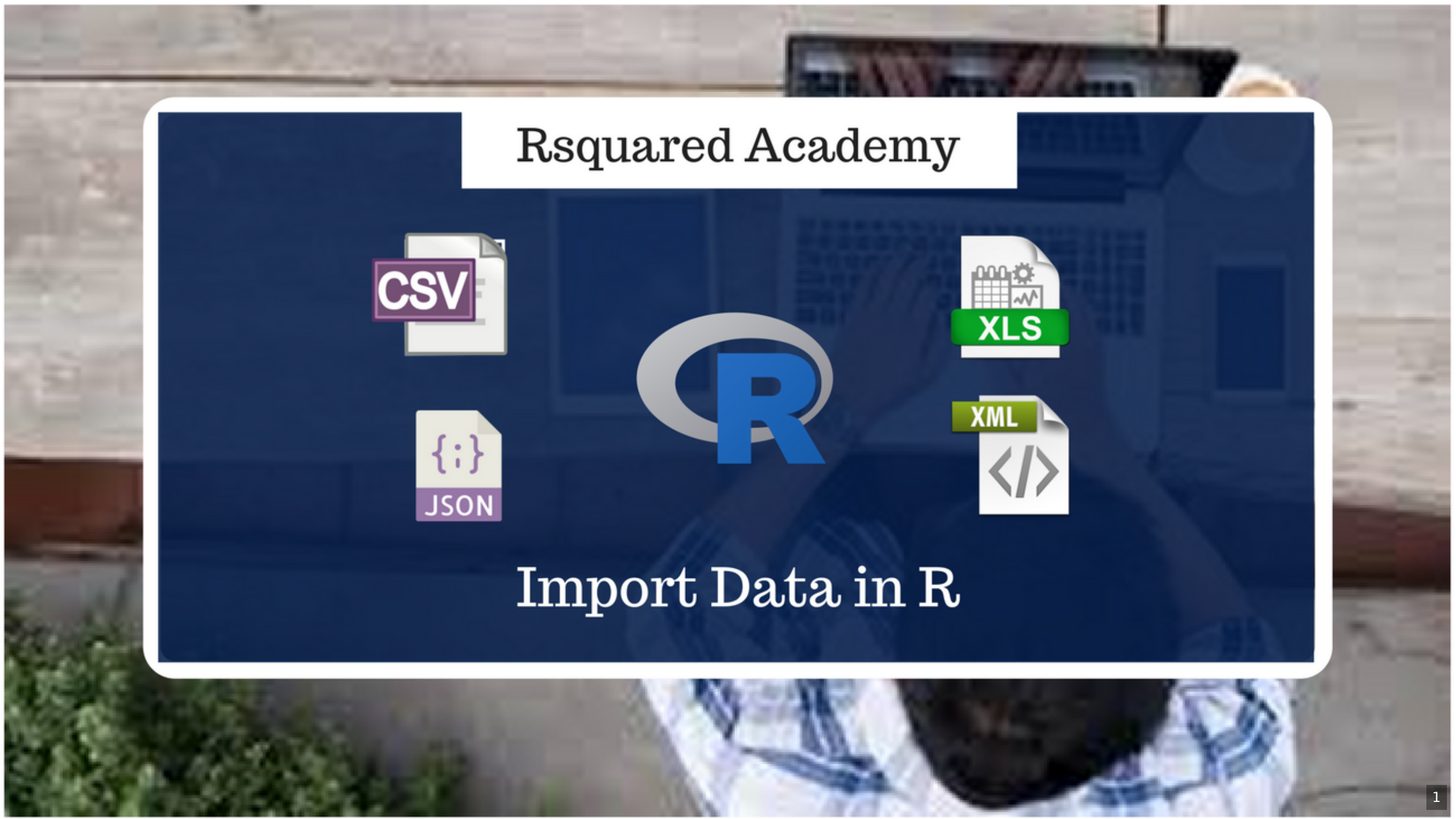

Rsquared Academy
CSV
XLS
JSON
XML
Import Data in R

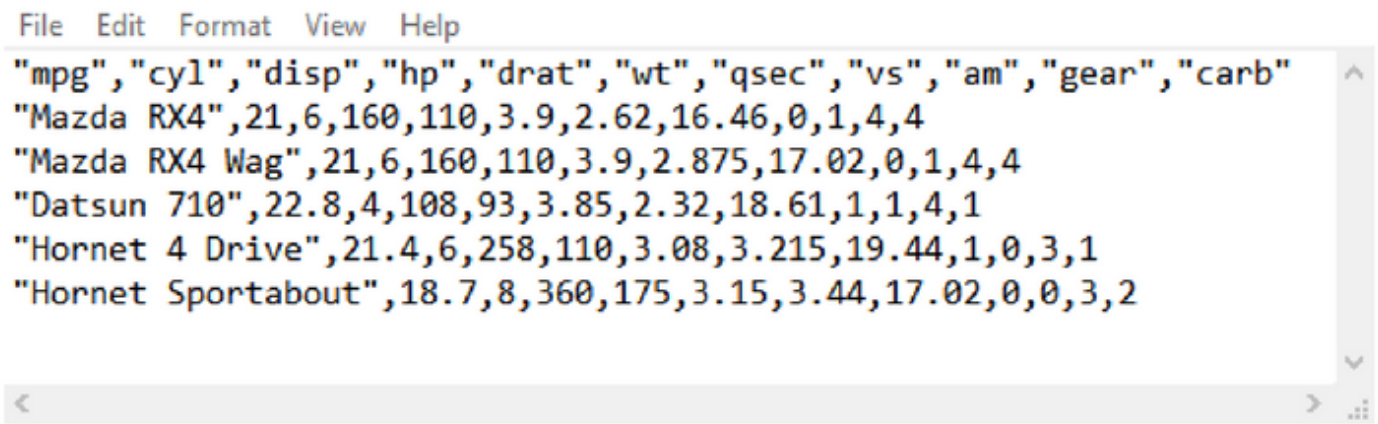
Agenda

- read data from flat or delimited files
- read data from excel spreadsheets
- read data from other statistical softwares
- specify column/variable types
- read specific columns/variables

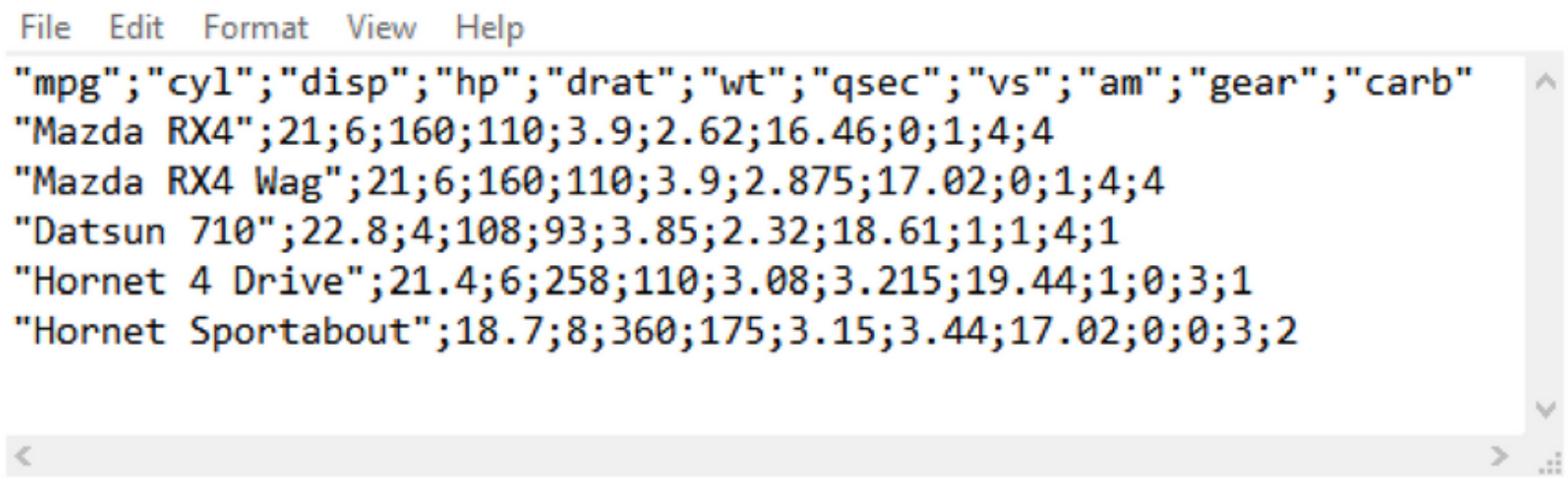
Libraries

```
library(readr)
library(readxl)
library(haven)
```

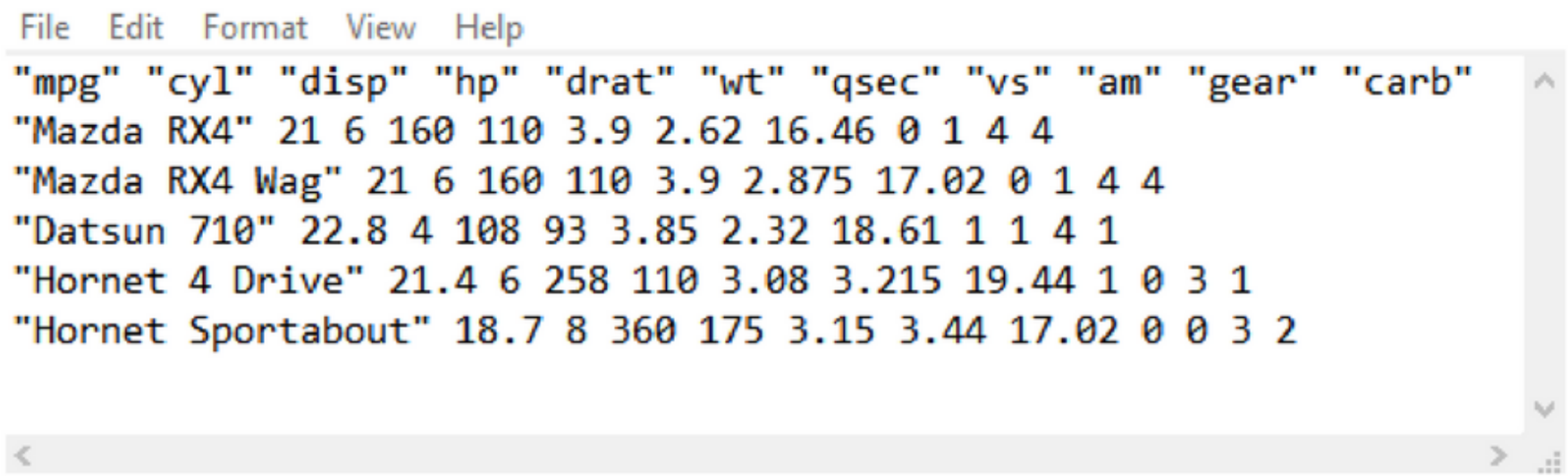
Comma Separated Values



Semi Colon Separated Values



Space Separated Values



Tab Separated Values

File Edit Format View Help

"mpg"	"cyl"	"disp"	"hp"	"drat"	"wt"	"qsec"	"vs"	"am"	"gear"	"carb"
"Mazda RX4"	21	6	160	110	3.9	2.62	16.46	0	1	4	4
"Mazda RX4 Wag"	21	6	160	110	3.9	2.875	17.02	0	1	4	4
"Datsun 710"	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
"Hornet 4 Drive"	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
"Hornet Sportabout"	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2

```
read_csv('data/mtcars.csv')
```

```
## # A tibble: 32 x 11
##      mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
##    <dbl> <int> <dbl> <int> <dbl> <dbl> <dbl> <int> <int> <int> <int>
##  1  21.0     6   160   110  3.90  2.62  16.5     0     1     4     4
##  2  21.0     6   160   110  3.90  2.88  17.0     0     1     4     4
##  3  22.8     4   108    93  3.85  2.32  18.6     1     1     4     1
##  4  21.4     6   258   110  3.08  3.22  19.4     1     0     3     1
##  5  18.7     8   360   175  3.15  3.44  17.0     0     0     3     2
##  6  18.1     6   225   105  2.76  3.46  20.2     1     0     3     1
##  7  14.3     8   360   245  3.21  3.57  15.8     0     0     3     4
##  8  24.4     4   147    62  3.69  3.19  20.0     1     0     4     2
##  9  22.8     4   141    95  3.92  3.15  22.9     1     0     4     2
## 10  19.2     6   168   123  3.92  3.44  18.3     1     0     4     4
## # ... with 22 more rows
```

```
read_delim('data/mtcars.csv', delim = ",")
```

```
## # A tibble: 32 x 11
##      mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
##    <dbl> <int> <dbl> <int> <dbl> <dbl> <dbl> <int> <int> <int> <int>
##  1  21.0     6   160   110  3.90  2.62  16.5     0     1     4     4
##  2  21.0     6   160   110  3.90  2.88  17.0     0     1     4     4
##  3  22.8     4   108    93  3.85  2.32  18.6     1     1     4     1
##  4  21.4     6   258   110  3.08  3.22  19.4     1     0     3     1
##  5  18.7     8   360   175  3.15  3.44  17.0     0     0     3     2
##  6  18.1     6   225   105  2.76  3.46  20.2     1     0     3     1
##  7  14.3     8   360   245  3.21  3.57  15.8     0     0     3     4
##  8  24.4     4   147    62  3.69  3.19  20.0     1     0     4     2
##  9  22.8     4   141    95  3.92  3.15  22.9     1     0     4     2
## 10  19.2     6   168   123  3.92  3.44  18.3     1     0     4     4
## # ... with 22 more rows
```

Column Names

File Edit Format View Help

```
"mpg","cyl","disp","hp","drat","wt","qsec","vs","am","gear","carb"
"Mazda RX4",21,6,160,110,3.9,2.62,16.46,0,1,4,4
"Mazda RX4 Wag",21,6,160,110,3.9,2.875,17.02,0,1,4,4
"Datsun 710",22.8,4,108,93,3.85,2.32,18.61,1,1,4,1
"Hornet 4 Drive",21.4,6,258,110,3.08,3.215,19.44,1,0,3,1
"Hornet Sportabout",18.7,8,360,175,3.15,3.44,17.02,0,0,3,2
```

File Edit Format View Help

```
"Mazda RX4",21,6,160,110,3.9,2.62,16.46,0,1,4,4
"Mazda RX4 Wag",21,6,160,110,3.9,2.875,17.02,0,1,4,4
"Datsun 710",22.8,4,108,93,3.85,2.32,18.61,1,1,4,1
"Hornet 4 Drive",21.4,6,258,110,3.08,3.215,19.44,1,0,3,1
"Hornet Sportabout",18.7,8,360,175,3.15,3.44,17.02,0,0,3,2
```

Column Names

```
read_csv('data/mtcars1.csv')
```

```
## Warning: Duplicated column names deduplicated: '4' => '4_1' [11]
```

```
## # A tibble: 31 x 11
##     `21`   `6` `160` `110` `3.9` `2.62` `16.46`   `0`   `1`   `4` `4_
##    <dbl> <int> <dbl> <int> <dbl>  <dbl>   <dbl> <int> <int> <int> <ir
##  1  21.0     6   160   110  3.90   2.88    17.0     0     1     4
##  2  22.8     4   108    93  3.85   2.32    18.6     1     1     4
##  3  21.4     6   258   110  3.08   3.22    19.4     1     0     3
##  4  18.7     8   360   175  3.15   3.44    17.0     0     0     3
##  5  18.1     6   225   105  2.76   3.46    20.2     1     0     3
##  6  14.3     8   360   245  3.21   3.57    15.8     0     0     3
##  7  24.4     4   147    62  3.69   3.19    20.0     1     0     4
##  8  22.8     4   141    95  3.92   3.15    22.9     1     0     4
##  9  19.2     6   168   123  3.92   3.44    18.3     1     0     4
## 10  17.8     6   168   123  3.92   3.44    18.9     1     0     4
## # ... with 21 more rows
```

Column Names

```
read_csv('data/mtcars1.csv', col_names = FALSE)
```

```
## # A tibble: 32 x 11
##       X1    X2    X3    X4    X5    X6    X7    X8    X9   X10   X11
##    <dbl> <int> <dbl> <int> <dbl> <dbl> <dbl> <int> <int> <int> <int>
##  1  21.0     6   160   110  3.90  2.62  16.5     0     1     4     4
##  2  21.0     6   160   110  3.90  2.88  17.0     0     1     4     4
##  3  22.8     4   108    93  3.85  2.32  18.6     1     1     4     1
##  4  21.4     6   258   110  3.08  3.22  19.4     1     0     3     1
##  5  18.7     8   360   175  3.15  3.44  17.0     0     0     3     2
##  6  18.1     6   225   105  2.76  3.46  20.2     1     0     3     1
##  7  14.3     8   360   245  3.21  3.57  15.8     0     0     3     4
##  8  24.4     4   147    62  3.69  3.19  20.0     1     0     4     2
##  9  22.8     4   141    95  3.92  3.15  22.9     1     0     4     2
## 10  19.2     6   168   123  3.92  3.44  18.3     1     0     4     4
## # ... with 22 more rows
```

Skip Lines

File Edit Format View Help

```
"The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design
,,,,,,,,,,
A data frame with 32 observations on 11 variables.,,,,,,,,,,
,,,,,,,,,,
"[, 1]", mpg, Miles/(US) gallon,,,,,,,,
"[, 2]", cyl, Number of cylinders,,,,,,,,
"[, 3]", disp, Displacement (cu.in.),,,,,,,,
"[, 4]", hp, Gross horsepower,,,,,,,,
"[, 5]", drat, Rear axle ratio,,,,,,,,
"[, 6]", wt, Weight (1000 lbs),,,,,,,,
"[, 7]", qsec, 1/4 mile time,,,,,,,,
"[, 8]", vs, V/S,,,,,,,,
"[, 9]", am," Transmission (0 = automatic, 1 = manual)",,,,,,,,
"[,10]", gear, Number of forward gears,,,,,,,,
"[,11]", carb, Number of carburetors,,,,,,,,
,,,,,,,,,,
,,,,,,,,,,
"Henderson and Velleman (1981), Building multiple regression models interactively. Biometrics, 37, 391–411.",,,,,,,,,,
,,,,,,,,,,
mpg,cyl,disp,hp,drat,wt,qsec,vs,am,gear,carb
21,6,160,110,3.9,2.62,16.46,0,1,4,4
21,6,160,110,3.9,2.875,17.02,0,1,4,4
22.8,4,108,93,3.85,2.32,18.61,1,1,4,1
21.4,6,258,110,3.08,3.215,19.44,1,0,3,1
18.7,8,360,175,3.15,3.44,17.02,0,0,3,2
18.1,6,225,105,2.76,3.46,20.22,1,0,3,1
14.3,8,360,245,3.21,3.57,15.84,0,0,3,4
```

Skip Lines

```
read_csv('data/mtcars2.csv')
```

```
## Warning: Missing column names filled in: 'X2' [2], 'X3' [3], 'X4' [4]
## 'X5' [5], 'X6' [6], 'X7' [7], 'X8' [8], 'X9' [9], 'X10' [10], 'X11' [
```

```
## # A tibble: 51 x 11
##    `The data was ex~ X2    X3    X4    X5    X6    X7    X8    X9
##    <chr>             <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
##  1 <NA>              <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
##  2 A data frame wit~ <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
##  3 <NA>              <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
##  4 [, 1]             mpg   Mile~ <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
##  5 [, 2]             cyl   Numb~ <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
##  6 [, 3]             disp  Disp~ <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
##  7 [, 4]             hp    Gros~ <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
##  8 [, 5]             drat  Rear~ <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
##  9 [, 6]             wt    Weig~ <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
## 10 [, 7]             qsec  1/4 ~ <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
## # ... with 41 more rows, and 1 more variable: X11 <chr>
```

Skip Lines

```
read_csv('data/mtcars2.csv', skip = 19)
```

```
## # A tibble: 32 x 11
##      mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
##    <dbl> <int> <dbl> <int> <dbl> <dbl> <dbl> <int> <int> <int> <int>
##  1  21.0     6   160   110  3.90  2.62  16.5     0     1     4     4
##  2  21.0     6   160   110  3.90  2.88  17.0     0     1     4     4
##  3  22.8     4   108    93  3.85  2.32  18.6     1     1     4     1
##  4  21.4     6   258   110  3.08  3.22  19.4     1     0     3     1
##  5  18.7     8   360   175  3.15  3.44  17.0     0     0     3     2
##  6  18.1     6   225   105  2.76  3.46  20.2     1     0     3     1
##  7  14.3     8   360   245  3.21  3.57  15.8     0     0     3     4
##  8  24.4     4   147    62  3.69  3.19  20.0     1     0     4     2
##  9  22.8     4   141    95  3.92  3.15  22.9     1     0     4     2
## 10  19.2     6   168   123  3.92  3.44  18.3     1     0     4     4
## # ... with 22 more rows
```

Maximum Lines

```
read_csv('data/mtcars.csv', n_max = 20)
```

```
## # A tibble: 20 x 11
##      mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
##    <dbl> <int> <dbl> <int> <dbl> <dbl> <dbl> <int> <int> <int> <int>
##  1  21.0     6 160     110  3.90  2.62  16.5     0     1     4     4
##  2  21.0     6 160     110  3.90  2.88  17.0     0     1     4     4
##  3  22.8     4 108      93  3.85  2.32  18.6     1     1     4     1
##  4  21.4     6 258     110  3.08  3.22  19.4     1     0     3     1
##  5  18.7     8 360     175  3.15  3.44  17.0     0     0     3     2
##  6  18.1     6 225     105  2.76  3.46  20.2     1     0     3     1
##  7  14.3     8 360     245  3.21  3.57  15.8     0     0     3     4
##  8  24.4     4 147      62  3.69  3.19  20.0     1     0     4     2
##  9  22.8     4 141      95  3.92  3.15  22.9     1     0     4     2
## 10  19.2     6 168     123  3.92  3.44  18.3     1     0     4     4
## 11  17.8     6 168     123  3.92  3.44  18.9     1     0     4     4
## 12  16.4     8 276     180  3.07  4.07  17.4     0     0     3     3
## 13  17.3     8 276     180  3.07  3.73  17.6     0     0     3     3
## 14  15.2     8 276     180  3.07  3.78  18.0     0     0     3     3
```

Column Types

Data Type	Function
Integer	`col_integer()`
Double	col_double()
Logical	col_logical()
Categorical	col_factor()
Character	col_character()
Date/Time	col_datetime(), col_date(), col_time()
Skip	col_skip()

Column Types

```
spec_csv('data/mtcars5.csv')
```

```
## cols(
##   mpg = col_double(),
##   cyl = col_integer(),
##   disp = col_double(),
##   hp = col_integer()
## )
```

Column Types

Objective	Function
Specify column data types	`col_types()`
Skip column	col_skip()
Read spcecific columns	cols_only()

```
read_csv('data/mtcars5.csv',
         col_types = list(col_double(), col_factor(levels = c(4, 6, 8)),
                          col_double(), col_integer()))
```

```
## # A tibble: 32 x 4
##      mpg cyl    disp    hp
##    <dbl> <fct> <dbl> <int>
##  1  21.0 6       160   110
##  2  21.0 6       160   110
##  3  22.8 4       108    93
##  4  21.4 6       258   110
##  5  18.7 8       360   175
##  6  18.1 6       225   105
##  7  14.3 8       360   245
##  8  24.4 4       147    62
##  9  22.8 4       141    95
## 10  19.2 6       168   123
## # ... with 22 more rows
```

Skip Columns

```
read_csv('data/mtcars5.csv',
         col_types = list(col_double(), col_factor(levels = c(4, 6, 8)),
                          col_skip(), col_integer()))
```

```
## # A tibble: 32 x 3
##      mpg cyl      hp
##    <dbl> <fct> <int>
##  1  21.0 6       110
##  2  21.0 6       110
##  3  22.8 4        93
##  4  21.4 6       110
##  5  18.7 8       175
##  6  18.1 6       105
##  7  14.3 8       245
##  8  24.4 4        62
##  9  22.8 4        95
## 10  19.2 6       123
## # ... with 22 more rows
```

Read Specific Columns

```
read_csv('data/mtcars5.csv',
         col_types = cols_only(mpg = col_double(),
                               cyl = col_factor(levels = c(4, 6, 8))))
```

```
## # A tibble: 32 x 2
##      mpg cyl
##    <dbl> <fct>
##  1  21.0 6
##  2  21.0 6
##  3  22.8 4
##  4  21.4 6
##  5  18.7 8
##  6  18.1 6
##  7  14.3 8
##  8  24.4 4
##  9  22.8 4
## 10  19.2 6
## # ... with 22 more rows
```

readr & Base R

Type	readr	Base R
comma	`read_csv()`	`read.csv()`
semicolon	read_csv2()	read.csv2()
tab	read_tsv()	read.delim() / read.table()
space	read_table()	read.table()
multiple spaces	read_table2()	read.table()
any delimiter	read_delim()	read.delim()

Your turn…

Open the below files, examine how the values are separated and read them into R using the appropriate function listed in the previous slide:

- mtcars.txt
- mtcars.tsv
- mtcars3.txt
- mtcars4.txt

- list sheets in an excel file
- read data from an excel sheet
- read specific cells
- read specific rows
- read specific columns

List Sheets

```
excel_sheets('data/sample.xls')
```

```
## [1] "ecom"
```

Read Sheet

```
read_excel('data/sample.xls', sheet = 1)
```

```
## # A tibble: 7 x 5
##   channel        users new_users sessions bounce_rate
##   <chr>          <dbl>     <dbl>    <dbl> <chr>
## 1 Organic Search 43296     40238    50810 48.72%
## 2 Direct         12916     12311    16419 49.27%
## 3 Referral       10983      7636    18105 22.26%
## 4 Social         10346     10029    11101 61.92%
## 5 Display         5564      4790     7220 83.30%
## 6 Paid Search     2687      2205     3438 38.02%
## 7 Affiliates      1773      1585     2167 55.75%
```

Read Specific Cells

	A	B	C	D	E
1	channel	users	new_users	sessions	bounce_rate
2	Organic Search	43296	40238	50810	48.72%
3	Direct	12916	12311	16419	49.27%
4	Referral	10983	7636	18105	22.26%
5	Social	10346	10029	11101	61.92%
6	Display	5564	4790	7220	83.30%
7	Paid Search	2687	2205	3438	38.02%
8	Affiliates	1773	1585	2167	55.75%
9					

B	C
users	new_users
43296	40238
12916	12311
10983	7636

Read Specific Cells

```
read_excel('data/sample.xls', sheet = 1, range = "B1:C4")
```

```
## # A tibble: 3 x 2
##   users new_users
##   <dbl>     <dbl>
## 1 43296     40238
## 2 12916     12311
## 3 10983      7636
```

Read Single Column

```
readxl::read_excel('data/sample.xls', sheet = 1, range = cell_cols(2))
```

```
## # A tibble: 7 x 1
##   users
##   <dbl>
## 1 43296
## 2 12916
## 3 10983
## 4 10346
## 5  5564
## 6  2687
## 7  1773
```

Specific Rows

```
readxl::read_excel('data/sample.xls', sheet = 1, range = cell_rows(1:4))
```

```
## # A tibble: 3 x 5
##   channel         users new_users sessions bounce_rate
##   <chr>           <dbl>     <dbl>    <dbl> <chr>
## 1 Organic Search 43296     40238    50810 48.72%
## 2 Direct         12916     12311    16419 49.27%
## 3 Referral       10983      7636    18105 22.26%
```

Specific Columns

```
readxl::read_excel('data/sample.xls', sheet = 1, range = cell_cols(2:3))
```

```
## # A tibble: 7 x 2
##   users new_users
##   <dbl>     <dbl>
## 1 43296     40238
## 2 12916     12311
## 3 10983      7636
## 4 10346     10029
## 5  5564      4790
## 6  2687      2205
## 7  1773      1585
```

Statistical Softwares

- SAS
- SPSS
- STATA

```
haven::read_stata('data/airline.dta')
```

```
## # A tibble: 32 x 6
##     year     y     w     r     l     k
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1  1948  1.21 0.243 0.145  1.41 0.612
##  2  1949  1.35 0.260 0.218  1.38 0.559
##  3  1950  1.57 0.278 0.316  1.39 0.573
##  4  1951  1.95 0.297 0.394  1.55 0.564
##  5  1952  2.27 0.310 0.356  1.80 0.574
##  6  1953  2.73 0.322 0.359  1.93 0.711
##  7  1954  3.03 0.335 0.403  1.96 0.776
##  8  1955  3.56 0.350 0.396  2.12 0.827
##  9  1956  3.98 0.361 0.382  2.43 0.800
## 10  1957  4.42 0.379 0.305  2.71 0.921
## # ... with 22 more rows
```

```
read_spss('data/employee.sav')
```

```
## # A tibble: 474 x 9
##       id gender    educ   jobcat  salary salbegin jobtime prevexp mir
##    <dbl> <chr+lbl> <dbl+> <dbl+l> <dbl+> <dbl+lb> <dbl+l> <dbl+l> <db
##  1  1.00 m         15     3       57000  27000    98       144     0
##  2  2.00 m         16     1       40200  18750    98       36      0
##  3  3.00 f         12     1       21450  12000    98       381     0
##  4  4.00 f         8      1       21900  13200    98       190     0
##  5  5.00 m         15     1       45000  21000    98       138     0
##  6  6.00 m         15     1       32100  13500    98       67      0
##  7  7.00 m         15     1       36000  18750    98       114     0
##  8  8.00 f         12     1       21900  9750     98       0       0
##  9  9.00 f         15     1       27900  12750    98       115     0
## 10 10.0  f         12     1       24000  13500    98       244     0
## # ... with 464 more rows
```

```
read_sas('data/airline.sas7bdat')
```

```
## # A tibble: 32 x 6
##     YEAR     Y     W     R     L     K
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1  1948  1.21 0.243 0.145  1.41 0.612
##  2  1949  1.35 0.260 0.218  1.38 0.559
##  3  1950  1.57 0.278 0.316  1.39 0.573
##  4  1951  1.95 0.297 0.394  1.55 0.564
##  5  1952  2.27 0.310 0.356  1.80 0.574
##  6  1953  2.73 0.322 0.359  1.93 0.711
##  7  1954  3.03 0.335 0.403  1.96 0.776
##  8  1955  3.56 0.350 0.396  2.12 0.827
##  9  1956  3.98 0.361 0.382  2.43 0.800
## 10  1957  4.42 0.379 0.305  2.71 0.921
## # ... with 22 more rows
```

Summary

File Type	readr	foreign/sas7bdat
excel	`read_excel()`	
sas	read_sas()	read.sas7bdat()
spss	read_sav() / read_spss()	read.spss()
stata	read_dta() / read_stata()	read.dta()

Thank You

For more information please visit our website
www.rsquaredacademy.com