Getting data from tabular sources

Getting and Cleaning Data

CSVs

Each column separated by a comma

sample_data.csv — Edited

```
name, height, blood_type
Natasha, 5'2" ,A-
Hassan, 6' , B-
Chun, 5'8", O
```

Has the extension ".csv"

Each row is separated by a new line

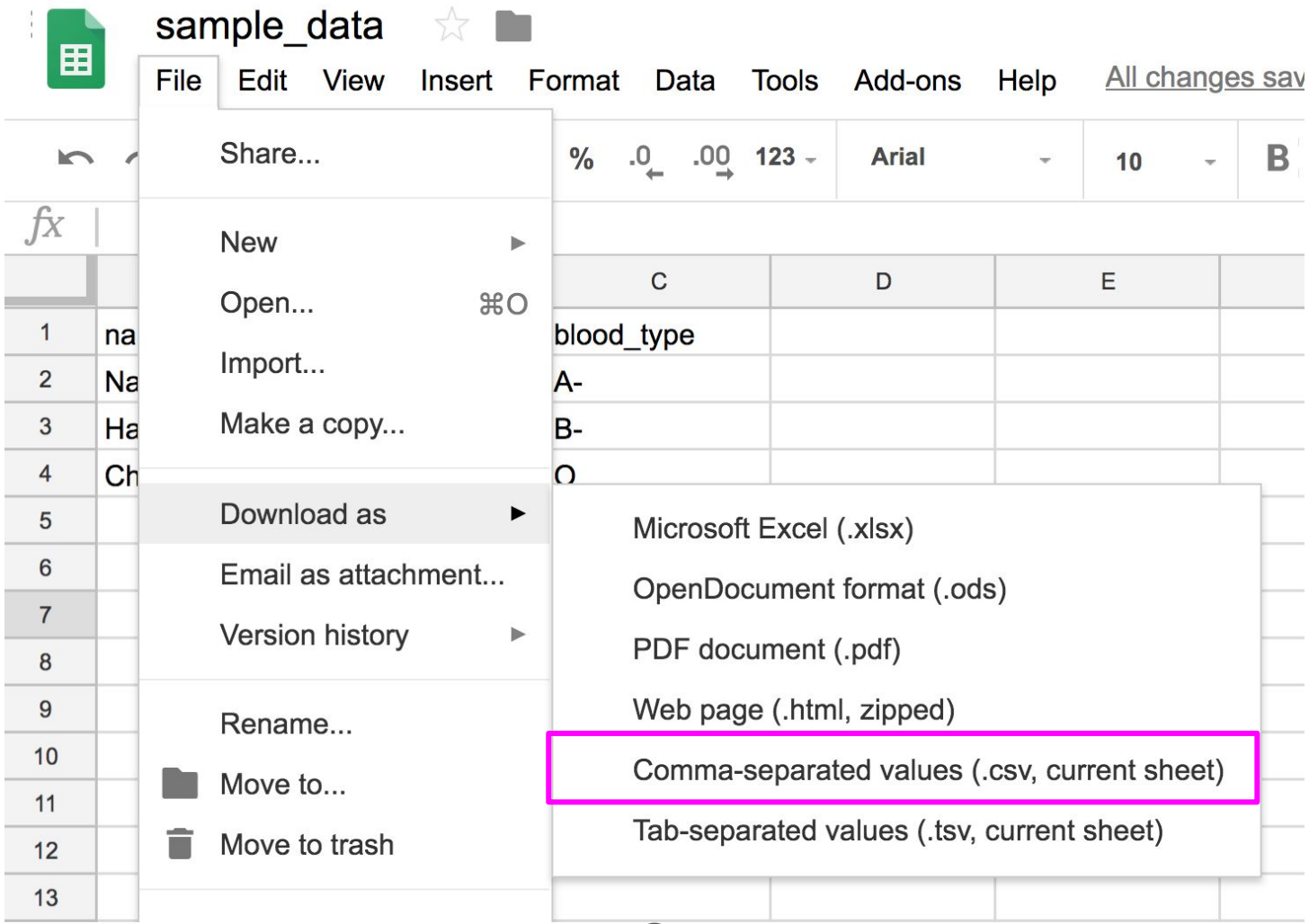
sample_data

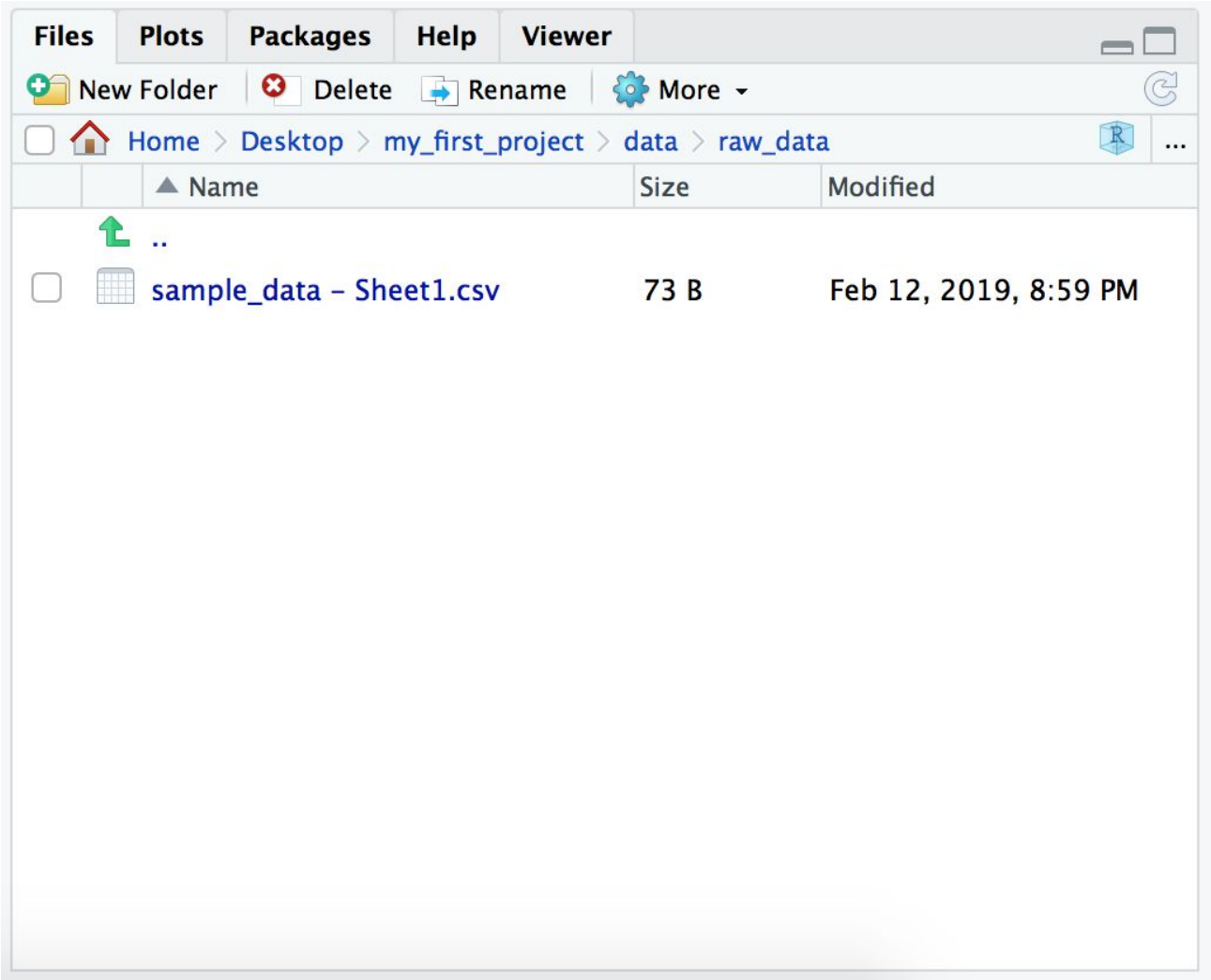
File Edit View Insert Format Data T

100% $ % .0 .00 1

fx

	A	B	C
1	name	height	blood_type
2	Natasha	5'2"	A-
3	Hassan	6'	B-
4	Chun	5'8"	O


sample_data
File
Edit
View
Insert
Format
Data
Tools
Add-ons
Help
All changes sav
Share...
New
Open... ⌘O
Import...
Make a copy...
Download as
Email as attachment...
Version history
Rename...
Move to...
Move to trash
Microsoft Excel (.xlsx)
OpenDocument format (.ods)
PDF document (.pdf)
Web page (.html, zipped)
Comma-separated values (.csv, current sheet)
Tab-separated values (.tsv, current sheet)
Arial
10
C
D
E
blood_type
A-
B-
O


Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home > Desktop > my_first_project > data > raw_data
Name
Size
Modified
..
sample_data – Sheet1.csv
73 B
Feb 12, 2019, 8:59 PM

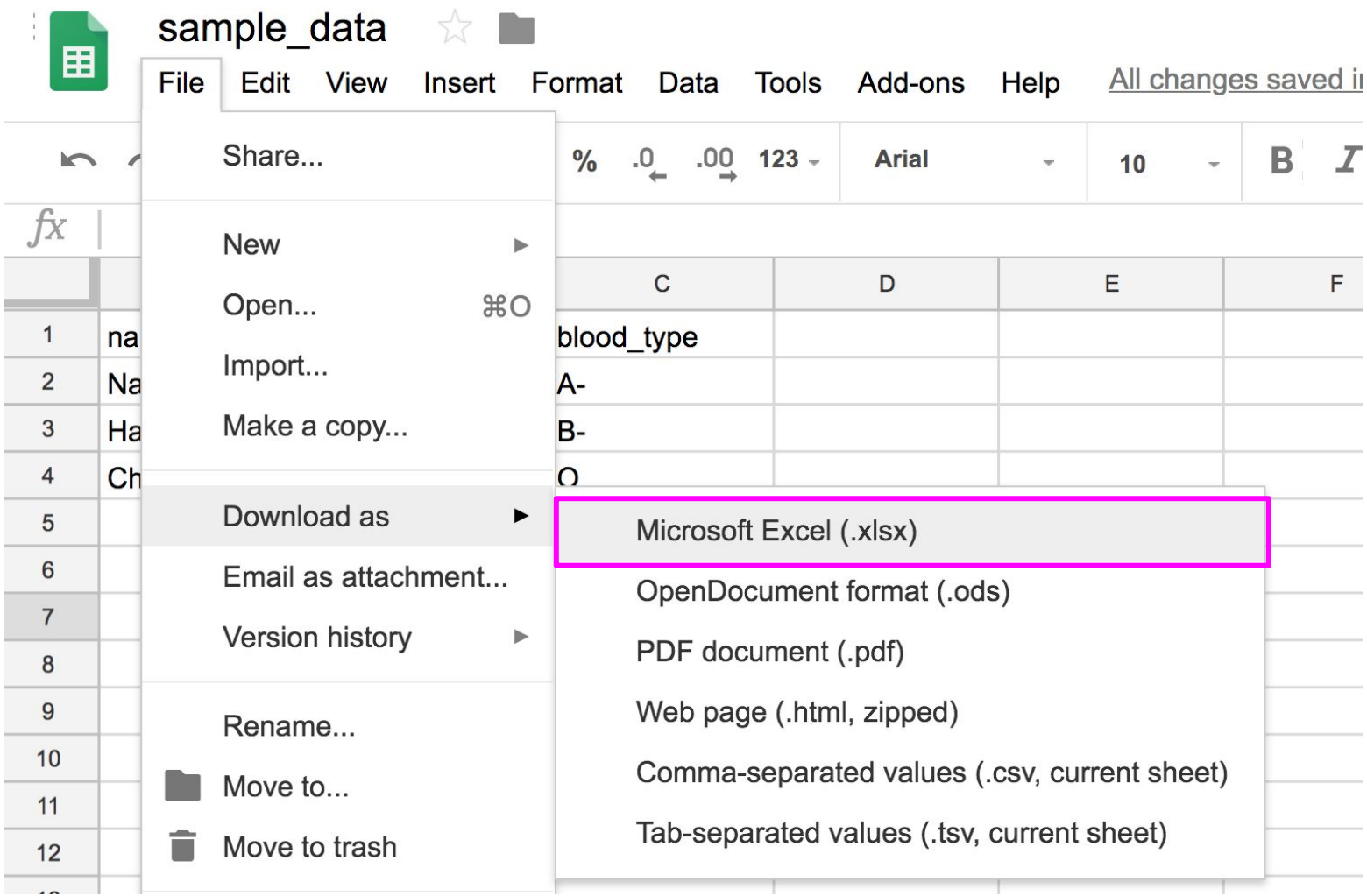
Untitled1*

Source on Save | Run | Source

```r
# Reading in the sample_data CSV file using readr

## install and load package
install.packages("readr")
library(readr)

## read CSV file into R
df_csv <- read_csv("data/raw_data/sample_data - Sheet1.csv")

## check the contents of your new data frame
head(df_csv)
```

12:1 (Top Level) R Script

- `- col_names = FALSE` to specify that the first row does NOT contain column names
- `- skip = 2` will skip the first 2 rows.
- `- n_max = 100` will only read in the first 100 rows.

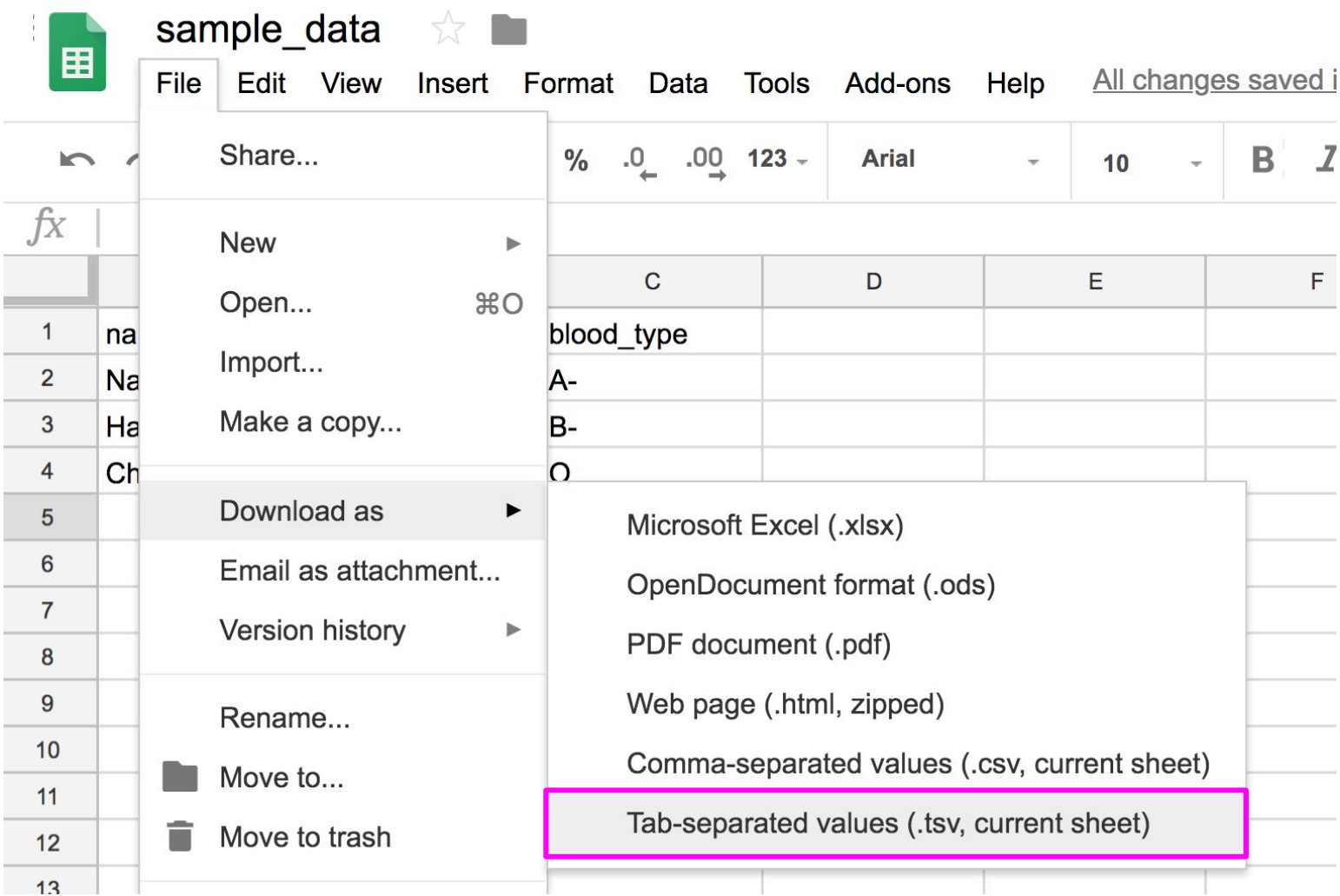

sample_data
File
Edit
View
Insert
Format
Data
Tools
Add-ons
Help
All changes saved i
Share...
New
Open... ⌘O
Import...
Make a copy...
Download as
Email as attachment...
Version history
Rename...
Move to...
Move to trash
Microsoft Excel (.xlsx)
OpenDocument format (.ods)
PDF document (.pdf)
Web page (.html, zipped)
Comma-separated values (.csv, current sheet)
Tab-separated values (.tsv, current sheet)
Arial
10
123
blood_type
A-
B-

Untitled1*

Source on Save | Run | Source

```r
# Reading in the sample_data XLSX file using readxl

## install and load package
install.packages("readxl")
library(readxl)

## read XLSX file into R
df_excel <- read_excel("data/raw_data/sample_data.xlsx")

## check the contents of your new data frame
head(df_excel)

```

12:1 (Top Level) R Script


sample_data
File
Edit
View
Insert
Format
Data
Tools
Add-ons
Help
All changes saved i
Share...
New
Open... ⌘O
Import...
Make a copy...
Download as
Email as attachment...
Version history
Rename...
Move to...
Move to trash
Microsoft Excel (.xlsx)
OpenDocument format (.ods)
PDF document (.pdf)
Web page (.html, zipped)
Comma-separated values (.csv, current sheet)
Tab-separated values (.tsv, current sheet)
Arial
10
123
blood_type
A-
B-

Untitled1*

Source on Save | Run | Source

```r
# Reading in the sample_data TSV file using readr

## install and load package
install.packages("readr")
library(readr)

## read TSV file into R
df_tsv <- read_tsv("data/raw_data/sample_data - Sheet1.tsv")

## check the contents of your new data frame
head(df_tsv)

```

12:1 (Top Level) R Script

Untitled1*

Source on Save | Run | Source

```r
# Reading in the sample_data TXT file using readr

## install and load package
install.packages("readr")
library(readr)

## read TXT into R
df_txt <- read_delim("data/raw_data/sample_data.txt", delim = "\t")

## check the contents of your new data frame
head(df_txt)

```

12:1 (Top Level) R Script

Untitled1*

Source on Save | Run | Source

```r
# Writing out the sample_data CSV file using readr

## install and load package
install.packages("readr")
library(readr)

## write CSV file
write_csv(df_csv, path = "my_csv_file.csv")
```

9:1 (Top Level) R Script

Console

Environment | History | Connections

Import Dataset | List

Global Environment

Data

df_csv	3 obs. of 3 variables
df_excel	3 obs. of 3 variables
df_tsv	3 obs. of 3 variables
df_txt	3 obs. of 3 variables

Files | Plots | Packages | Help | Viewer

New Folder | Delete | Rename | More

Home > Desktop > my_first_project

Name	Size	Modified
..		
.Rhistory	0 B	Feb 10, 20
code		
data		
figures		
my_csv_file.csv	70 B	Feb 12, 20
my_first_project.Rproj	205 B	Feb 12, 20
products		
README.md	539 B	Feb 10, 20

Untitled1*

Source

```r
# Previewing our data using head and tail

head(df_csv)

tail(df_csv)

head(df_csv, n = 2)

```

```
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 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 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
> str(mtcars)
'data.frame':   32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

glimpse()

skim()

Untitled1*

Source on Save | Run | Source

```r
# Preview and summarize data using glimpse and skim

## install dplyr, if you have not already (you probably have)
install.packages("dplyr")

## install skimr
install.packages("skimr")

## load these libraries
library(dplyr)
library(skimr)

## call the functions glimpse() and skim() on the mtcars dataset
glimpse(mtcars)
skim(mtcars)
```

16:1 (Top Level) R Script

```
> glimpse(mtcars)
Observations: 32
Variables: 11
$ mpg  <dbl> 21.0, 21.0, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19.2, 17.8, 16.4, 17.3, 15.2, 10.4, 10.4, 14.7, 32.4, 30....
$ cyl  <dbl> 6, 6, 4, 6, 8, 6, 8, 4, 4, 6, 6, 8, 8, 8, 8, 8, 8, 4, 4, 4, 4, 8, 8, 8, 8, 4, 4, 4, 8, 6, 8, 4
$ disp <dbl> 160.0, 160.0, 108.0, 258.0, 360.0, 225.0, 360.0, 146.7, 140.8, 167.6, 167.6, 275.8, 275.8, 275.8, 472.0, 460.0,...
$ hp   <dbl> 110, 110, 93, 110, 175, 105, 245, 62, 95, 123, 123, 180, 180, 180, 205, 215, 230, 66, 52, 65, 97, 150, 150, 245...
$ drat <dbl> 3.90, 3.90, 3.85, 3.08, 3.15, 2.76, 3.21, 3.69, 3.92, 3.92, 3.92, 3.07, 3.07, 3.07, 2.93, 3.00, 3.23, 4.08, 4.9...
$ wt   <dbl> 2.620, 2.875, 2.320, 3.215, 3.440, 3.460, 3.570, 3.190, 3.150, 3.440, 3.440, 4.070, 3.730, 3.780, 5.250, 5.424,...
$ qsec <dbl> 16.46, 17.02, 18.61, 19.44, 17.02, 20.22, 15.84, 20.00, 22.90, 18.30, 18.90, 17.40, 17.60, 18.00, 17.98, 17.82,...
$ vs   <dbl> 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1
$ am   <dbl> 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1
$ gear <dbl> 4, 4, 4, 3, 3, 3, 3, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 4, 4, 4, 3, 3, 3, 3, 3, 4, 5, 5, 5, 5, 5, 4
$ carb <dbl> 4, 4, 1, 1, 2, 1, 4, 2, 2, 4, 4, 3, 3, 3, 4, 4, 4, 1, 2, 1, 1, 2, 2, 4, 2, 1, 2, 2, 4, 6, 8, 2
> skim(mtcars)
Skim summary statistics
 n obs: 32
 n variables: 11

── Variable type:numeric ─────────────────────────────────────────────────────────────────────────────────────────────────────────
 variable missing complete  n   mean     sd    p0    p25    p50    p75   p100     hist
       am       0       32 32   0.41   0.5   0      0      0      1      1
     carb       0       32 32   2.81   1.62  1      2      2      4      8
      cyl       0       32 32   6.19   1.79  4      4      6      8      8
     disp       0       32 32 230.72 123.94 71.1  120.83 196.3  326    472
     drat       0       32 32   3.6    0.53  2.76   3.08   3.7    3.92   4.93
     gear       0       32 32   3.69   0.74  3      3      4      4      5
       hp       0       32 32 146.69  68.56 52     96.5  123    180    335
      mpg       0       32 32  20.09   6.03 10.4   15.43  19.2   22.8   33.9
     qsec       0       32 32  17.85   1.79 14.5   16.89  17.71  18.9   22.9
       vs       0       32 32   0.44   0.5   0      0      0      1      1
       wt       0       32 32   3.22   0.98  1.51   2.58   3.33   3.61   5.42
```

Summarizing: Getting data from tabular sources

Getting and Cleaning Data