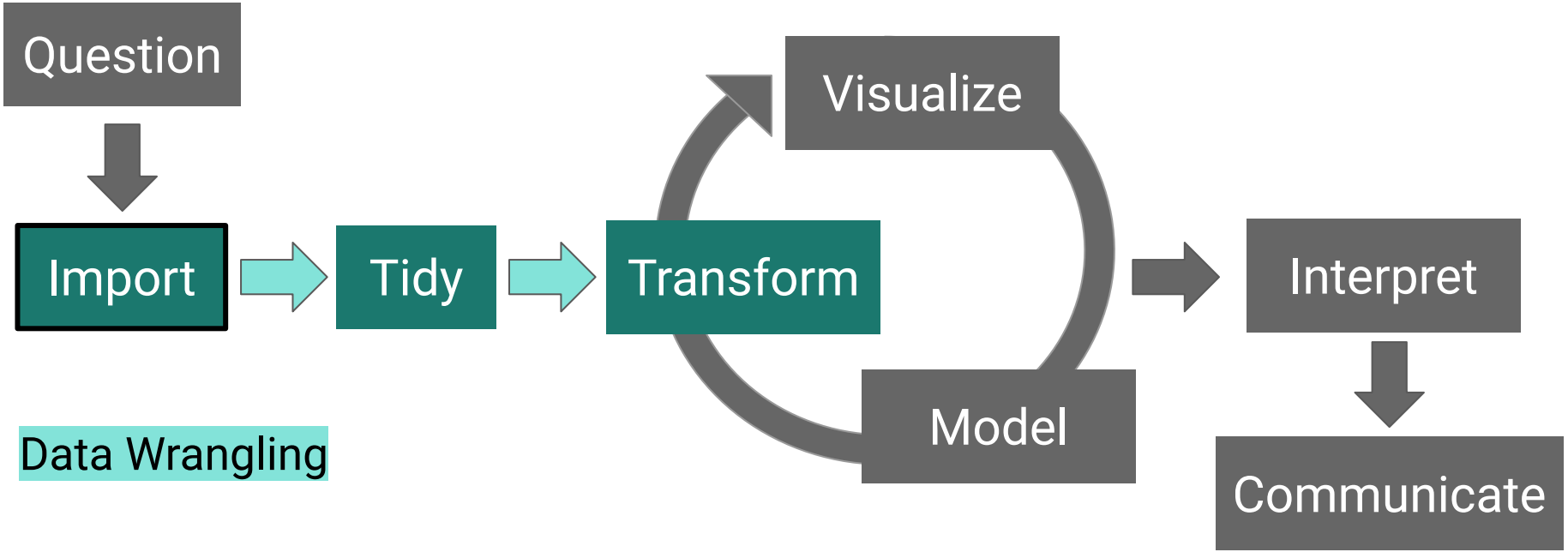
Importing Data

Lecture 4

Objectives

- To import files
- To download files
- To write to a file

Motivation



Modified from https://r4ds.had.co.nz/explore-intro.html#explore-intro

base R vs package

- R packages are collections of functions developed by the community
- R packages improve existing base R functionalities, or by add new ones

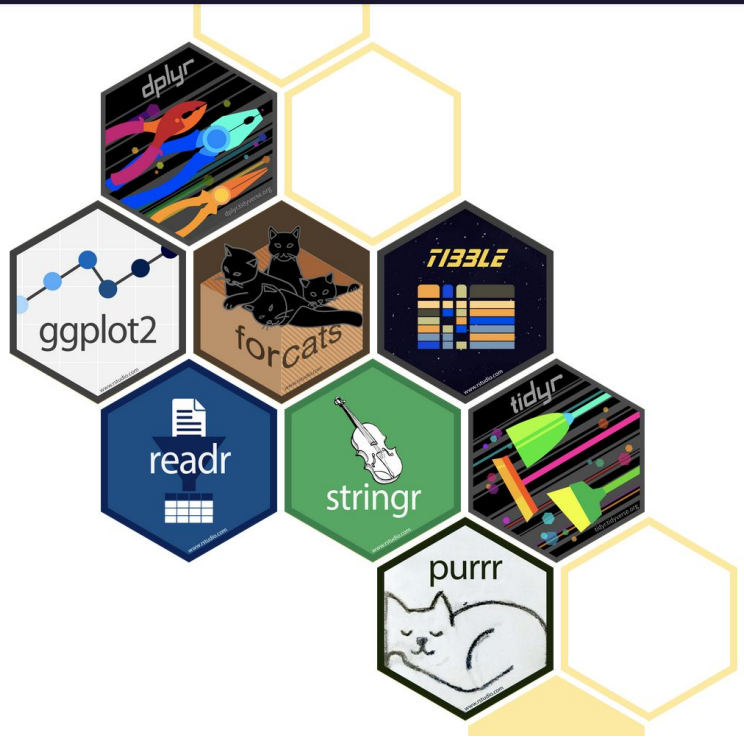
```
# Read documentation
> packageDescription("tidyverse")
> help(package = "tidyverse")
```

```
Package: tidyverse
Title: Easily Install and Load the 'Tidyverse'
Version: 1.3.0
Authors@R: c(person(given = "Hadley", family = "Wickham", role = c("aut", "cre"), email = "hadley@rstudio.com"), person(given =
           "RStudio", role = c("cph", "fnd")))
Description: The 'tidyverse' is a set of packages that work in harmony because they share common data representations and 'API'
design. This package is designed to make it easy to install and load multiple 'tidyverse' packages in a single step. Learn more about the
'tidyverse' at <https://tidyverse.org>. ...
```

Tidyverse

Tidyverse

Packages Blog Learn Help Contribute



R packages for data science

The tidyverse is an opinionated **collection of R packages** designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

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

https://www.tidyverse.org/

Tibble vs Dataframe

Tidyverse

Packages

tibble

tibble is a modern re-imagining of the data frame, keeping what time has proven to be effective, and throwing out what it has not. Tibbles are data.frames that are lazy and surly: they do less and complain more forcing you to confront problems earlier, typically leading to cleaner, more expressive code. Go to docs...

https://www.tidyverse.org/packages/

readr package

Tidyverse

Packages

readr

readr provides a fast and friendly way to read rectangular data (like csv, tsv, and fwf). It is designed to flexibly parse many types of data found in the wild, while still cleanly failing when data unexpectedly changes. Go to docs...

https://www.tidyverse.org/packages/

Load tidyverse

```
> library(tidyverse)

── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──
✓ ggplot2 3.3.2  ✓ purrr   0.3.4
✓ tibble  3.0.4  ✓ dplyr   1.0.2
✓ tidyr   1.1.2  ✓ stringr 1.4.0
✓ readr   1.4.0  ✓ forcats 0.5.0
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
x dplyr::filter()     masks stats::filter()
x dplyr::lag()        masks stats::lag()
```

Importing text files

- open plain text files and convert to data frames
- text files
  - comma-separated value (csv) files
  - tab-delimited files
- other delimiters
  - semi-colon (;)
  - forward slash (/)

Importing text files

patient.csv

```
patient_ID,sex,age_year,weight_kg,height_cm
P001,female,1,9.1,73
P002,female,4,16.4,96
P003,female,2,10.5,85
P004,male,3,13.2,95
P005,male,4,15.9,104
```

```
> patient
```

	patient_ID	sex	age_year	weight_kg	height_cm
1	P001	female	1	9.1	73
2	P002	female	4	16.4	96
3	P003	female	2	10.5	85
4	P004	male	3	13.2	95
5	P005	male	4	15.9	104

read_csv( )

```
> patient <- read_csv("patient.csv")
```

```
── Column specification ────────────────────────────────────────────────────────
cols(
  patient_ID = col_character(),
  sex = col_character(),
  age_year = col_double(),
  weight_kg = col_double(),
  height_cm = col_double()
)
```

```
> patient
```

```
# A tibble: 5 x 5
  patient_ID sex    age_year weight_kg height_cm
  <chr>      <chr>     <dbl>     <dbl>     <dbl>
1 P001       female        1       9.1        73
2 P002       female        4      16.4        96
3 P003       female        2      10.5        85
4 P004       male          3      13.2        95
5 P005       male          4      15.9       104
```

Parsing a file

- **readr** automatically guesses the type of vector in each column
- uses the **parse_*( )** function
  - *** = logical, integer, double, character, date, factor**

```
> parse_double(c("123", "456", "abc", "123.456"))
```

```
[1] 123.000 456.000     NA 123.456
attr(,"problems")
# A tibble: 1 x 4
        row   col expected actual
        <int> <int> <chr> <chr>
1       3       NA a double abc
```

read_tsv( )

patient.txt

patient_ID	sex	age_year	weight_kg	height_cm
P001	female	1	9.1	73
P002	female	4	16.4	96
P003	female	2	10.5	85
P004	male	3	13.2	95
P005	male	4	15.9	104

```
> patient <- read_tsv("patient.txt")
> patient
```

```
# A tibble: 5 x 5
  patient_ID  sex     age_year  weight_kg  height_cm
  <chr>       <chr>      <dbl>      <dbl>      <dbl>
1 P001        female         1        9.1         73
2 P002        female         4       16.4         96
3 P003        female         2       10.5         85
4 P004        male           3       13.2         95
5 P005        male           4       15.9        104
```

No headers

patient_headless.csv

```
P001,female,1,9.1,73
P002,female,4,16.4,96
P003,female,2,10.5,85
P004,male,3,13.2,95
P005,male,4,15.9,104
```

```
> patient <- read_csv("patient_headless.csv",
                      col_names = FALSE)
> patient
```

```
# A tibble: 5 x 5
  X1    X2     X3     X4    X5
  <chr> <chr>  <dbl> <dbl> <dbl>
1 P001  female     1   9.1    73
2 P002  female     4  16.4    96
3 P003  female     2  10.5    85
4 P004  male       3  13.2    95
5 P005  male       4  15.9   104
```

Provide headers

```
> headers <- c("patient", "sex", "age", "weight", "height")
> patient <- read_csv("patient_headless.csv",
                      col_names = headers)
> patient
```

```
# A tibble: 5 x 5
  patient sex        age weight height
  <chr>   <chr>  <dbl>  <dbl>  <dbl>
1 P001    female     1    9.1     73
2 P002    female     4   16.4     96
3 P003    female     2   10.5     85
4 P004    male       3   13.2     95
5 P005    male       4   15.9    104
```

Skip lines

```
> patient <- read_csv("patient_headless.csv",
                      col_names = FALSE,
                      skip = 2)
> patient
```

```
# A tibble: 3 x 5
  X1   X2     X3     X4     X5
  <chr> <chr>  <dbl> <dbl> <dbl>
1 P003  female   2  10.5    85
2 P004  male     3  13.2    95
3 P005  male     4  15.9   104
```

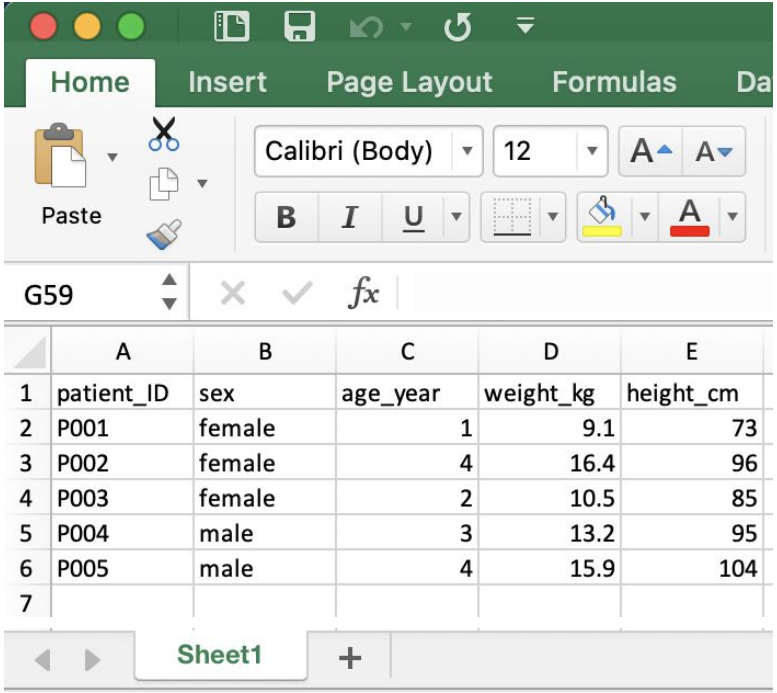
Importing Excel files

- different sheets with tabular data

```
> install.packages("readxl")
> library(readxl)

# list different sheets
> excel_sheets("patient.xlsx")
```

```
1] "Sheet1"
```



	A	B	C	D	E
1	patient_ID	sex	age_year	weight_kg	height_cm
2	P001	female	1	9.1	73
3	P002	female	4	16.4	96
4	P003	female	2	10.5	85
5	P004	male	3	13.2	95
6	P005	male	4	15.9	104
7					

read_excel( )

```
> patient <- read_excel("patient.xlsx", sheet = "Sheet1")    # name of sheet
> patient <- read_excel("patient.xlsx", sheet = 1)           # sheet number
```

```
> patient
```

```
# A tibble: 5 x 5
  patient_ID  sex     age_year  weight_kg  height_cm
  <chr>       <chr>      <dbl>     <dbl>      <dbl>
1 P001        female         1       9.1         73
2 P002        female         4      16.4         96
3 P003        female         2      10.5         85
4 P004        male           3      13.2         95
5 P005        male           4      15.9        104
```

Importing files from the internet

https://github.com/CSSEGISandData/COVID-19

Most Visited Getting Started

Why GitHub? Team Enterprise Explore Marketplace Pricing Search / Sign in Sign up

CSSEGISandData / COVID-19

Watch 945 Star 25.2k Fork 16.1k

Code Issues 1.3k Pull requests 281 Actions Projects Security Insights

master 19 branches 0 tags Go to file Code

CSSEGISandData Automated update 269936d 6 hours ago 2,354 commits

archived_data	archived_0325	9 months ago
csse_covid_19_data	Automated update	6 hours ago
who_covid_19_situation_reports	update who readme	4 months ago
.gitignore	update	11 months ago
README.md	Update README.md	10 days ago

About

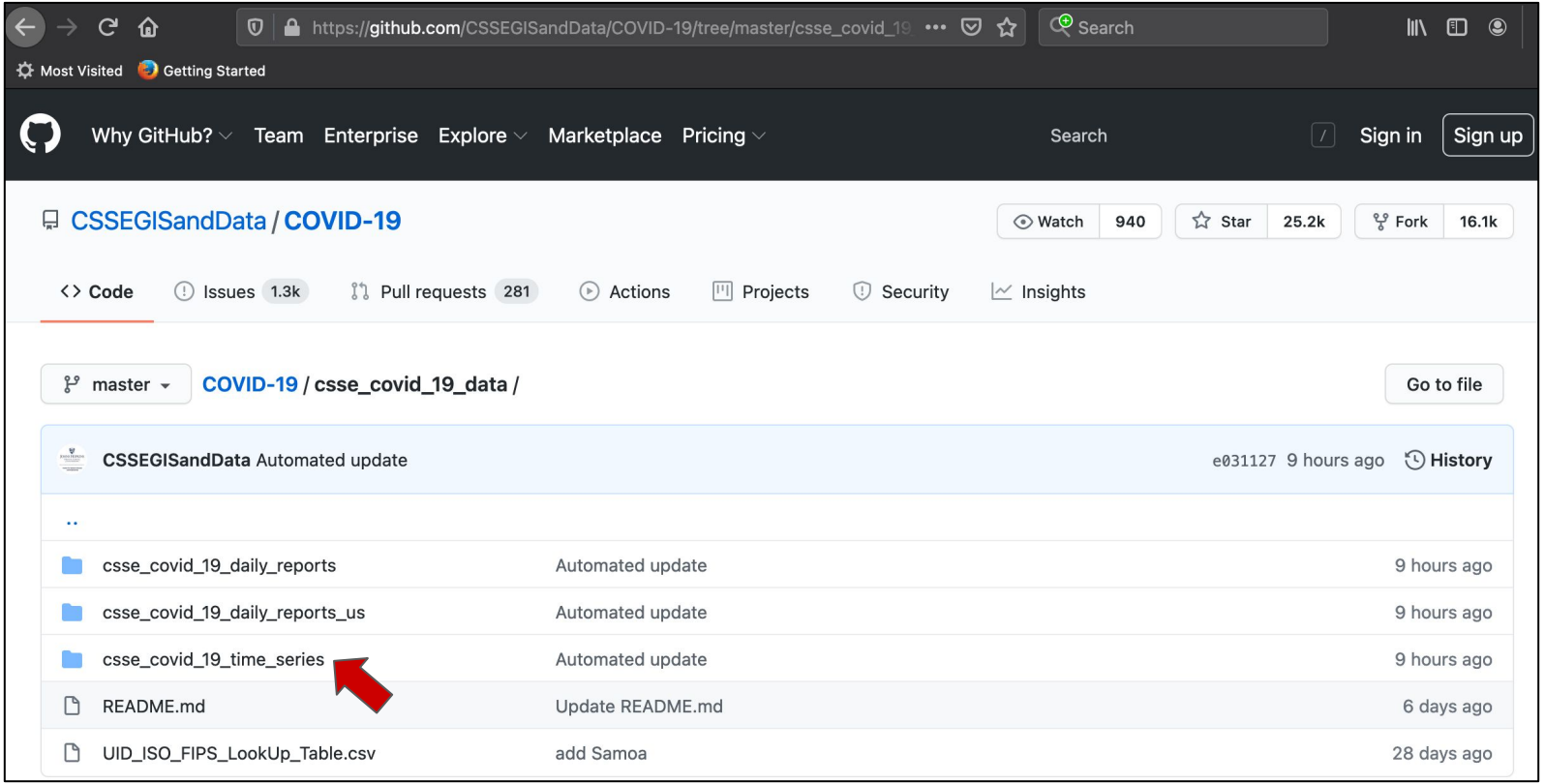
Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE

systems.jhu.edu/research/public-...

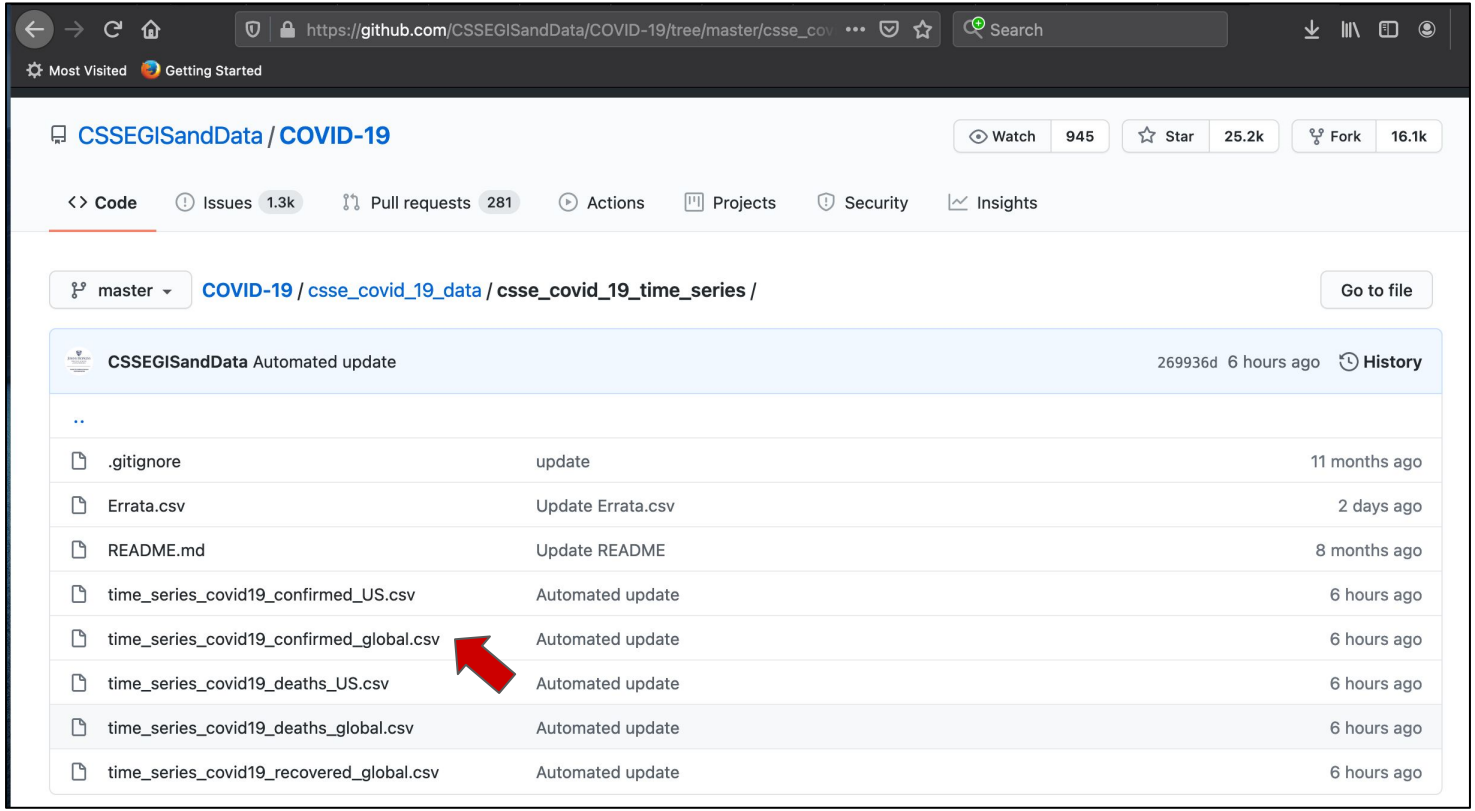
johns-hopkins-university
systems-science engineering
covid-19 2019-ncov coronavirus
csse jhu

Readme

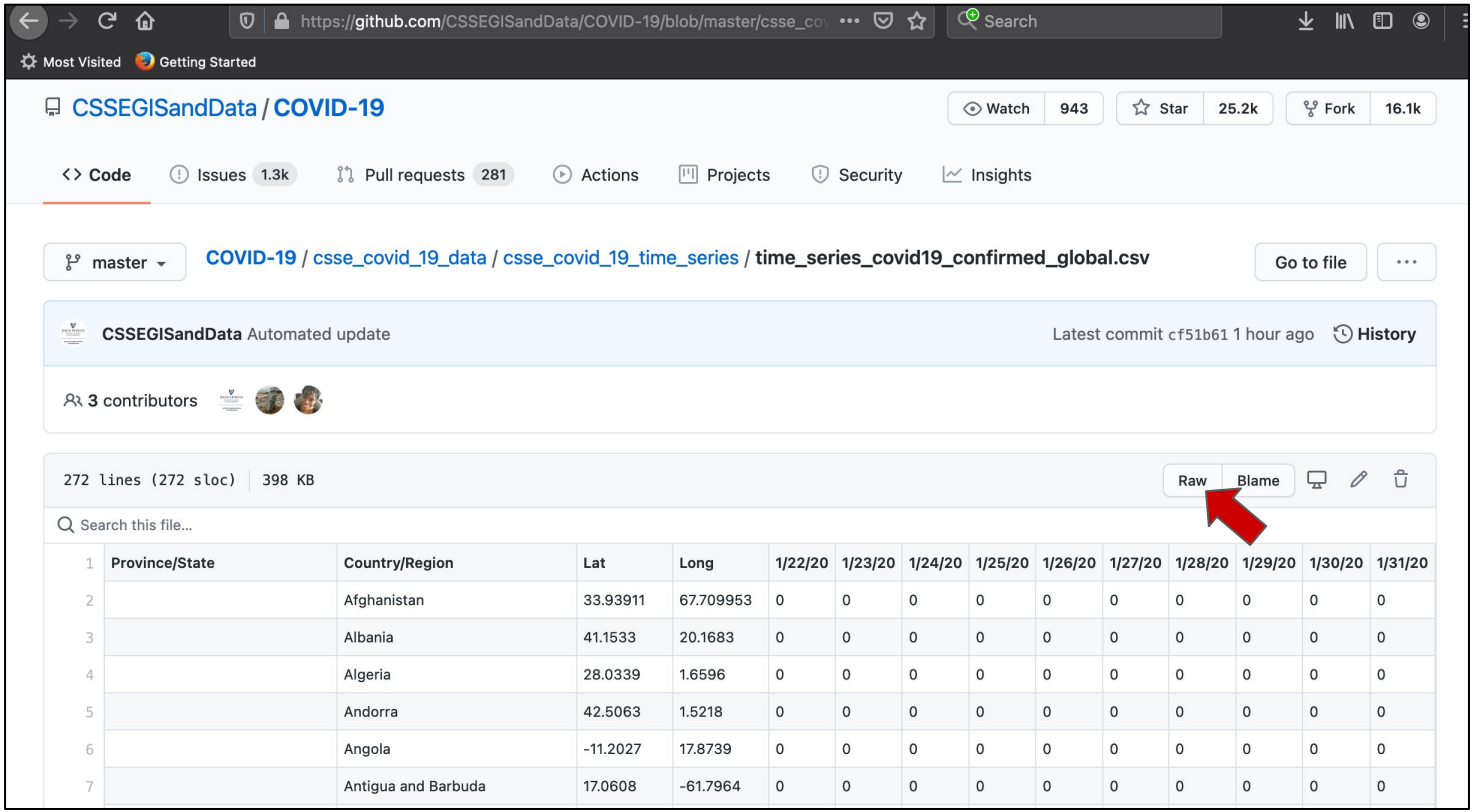
Importing files from the internet



Importing files from the internet



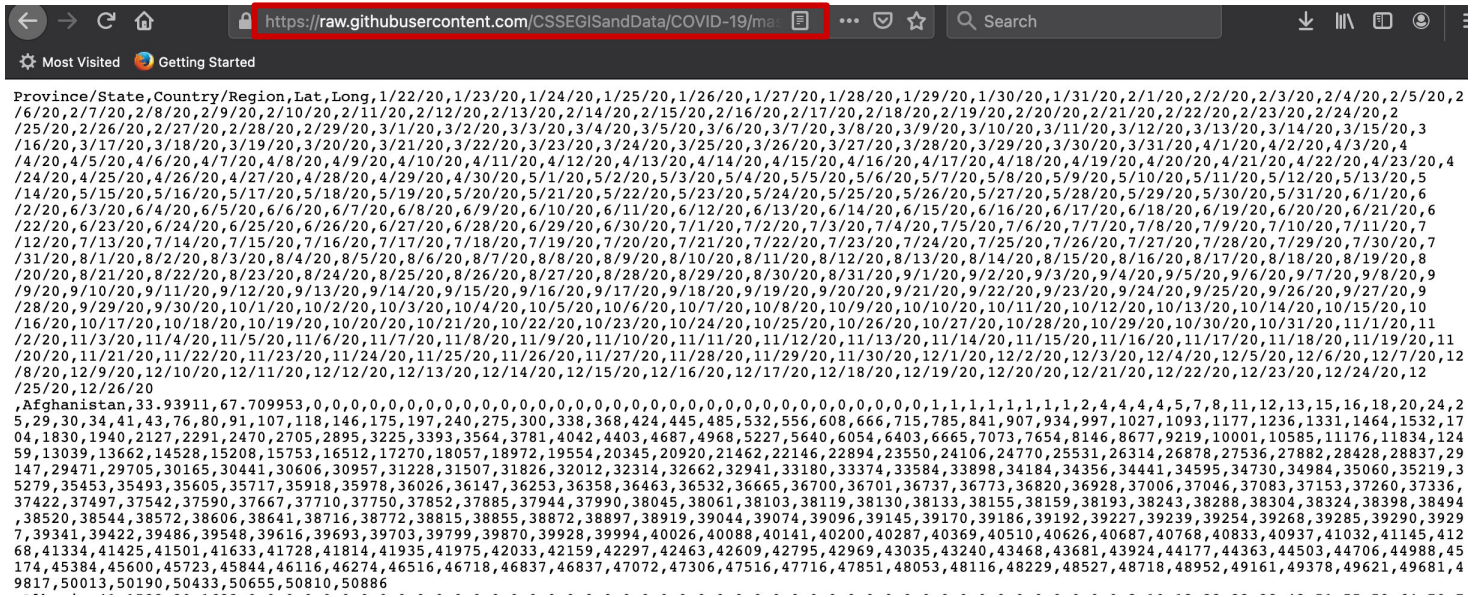
Importing files from the internet



	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	1/28/20	1/29/20	1/30/20	1/31/20
2		Afghanistan	33.93911	67.709953	0	0	0	0	0	0	0	0	0	0
3		Albania	41.1533	20.1683	0	0	0	0	0	0	0	0	0	0
4		Algeria	28.0339	1.6596	0	0	0	0	0	0	0	0	0	0
5		Andorra	42.5063	1.5218	0	0	0	0	0	0	0	0	0	0
6		Angola	-11.2027	17.8739	0	0	0	0	0	0	0	0	0	0
7		Antigua and Barbuda	17.0608	-61.7964	0	0	0	0	0	0	0	0	0	0

Importing files from the internet

- Save the raw file as plain text file and download to your local machine
- Copy the link the of the raw file and download in R




```
Province/State,Country/Region,Lat,Long,1/22/20,1/23/20,1/24/20,1/25/20,1/26/20,1/27/20,1/28/20,1/29/20,1/30/20,1/31/20,2/1/20,2/2/20,2/3/20,2/4/20,2/5/20,2
/6/20,2/7/20,2/8/20,2/9/20,2/10/20,2/11/20,2/12/20,2/13/20,2/14/20,2/15/20,2/16/20,2/17/20,2/18/20,2/19/20,2/20/20,2/21/20,2/22/20,2/23/20,2/24/20,2
/25/20,2/26/20,2/27/20,2/28/20,2/29/20,3/1/20,3/2/20,3/3/20,3/4/20,3/5/20,3/6/20,3/7/20,3/8/20,3/9/20,3/10/20,3/11/20,3/12/20,3/13/20,3/14/20,3/15/20,3
/16/20,3/17/20,3/18/20,3/19/20,3/20/20,3/21/20,3/22/20,3/23/20,3/24/20,3/25/20,3/26/20,3/27/20,3/28/20,3/29/20,3/30/20,3/31/20,4/1/20,4/2/20,4/3/20,4
/4/20,4/5/20,4/6/20,4/7/20,4/8/20,4/9/20,4/10/20,4/11/20,4/12/20,4/13/20,4/14/20,4/15/20,4/16/20,4/17/20,4/18/20,4/19/20,4/20/20,4/21/20,4/22/20,4/23/20,4
/24/20,4/25/20,4/26/20,4/27/20,4/28/20,4/29/20,4/30/20,5/1/20,5/2/20,5/3/20,5/4/20,5/5/20,5/6/20,5/7/20,5/8/20,5/9/20,5/10/20,5/11/20,5/12/20,5/13/20,5
/14/20,5/15/20,5/16/20,5/17/20,5/18/20,5/19/20,5/20/20,5/21/20,5/22/20,5/23/20,5/24/20,5/25/20,5/26/20,5/27/20,5/28/20,5/29/20,5/30/20,5/31/20,6/1/20,6
/2/20,6/3/20,6/4/20,6/5/20,6/6/20,6/7/20,6/8/20,6/9/20,6/10/20,6/11/20,6/12/20,6/13/20,6/14/20,6/15/20,6/16/20,6/17/20,6/18/20,6/19/20,6/20/20,6/21/20,6
/22/20,6/23/20,6/24/20,6/25/20,6/26/20,6/27/20,6/28/20,6/29/20,6/30/20,7/1/20,7/2/20,7/3/20,7/4/20,7/5/20,7/6/20,7/7/20,7/8/20,7/9/20,7/10/20,7/11/20,7
/12/20,7/13/20,7/14/20,7/15/20,7/16/20,7/17/20,7/18/20,7/19/20,7/20/20,7/21/20,7/22/20,7/23/20,7/24/20,7/25/20,7/26/20,7/27/20,7/28/20,7/29/20,7/30/20,7
/31/20,8/1/20,8/2/20,8/3/20,8/4/20,8/5/20,8/6/20,8/7/20,8/8/20,8/9/20,8/10/20,8/11/20,8/12/20,8/13/20,8/14/20,8/15/20,8/16/20,8/17/20,8/18/20,8/19/20,8
/20/20,8/21/20,8/22/20,8/23/20,8/24/20,8/25/20,8/26/20,8/27/20,8/28/20,8/29/20,8/30/20,8/31/20,9/1/20,9/2/20,9/3/20,9/4/20,9/5/20,9/6/20,9/7/20,9/8/20,9
/9/20,9/10/20,9/11/20,9/12/20,9/13/20,9/14/20,9/15/20,9/16/20,9/17/20,9/18/20,9/19/20,9/20/20,9/21/20,9/22/20,9/23/20,9/24/20,9/25/20,9/26/20,9/27/20,9
/28/20,9/29/20,9/30/20,10/1/20,10/2/20,10/3/20,10/4/20,10/5/20,10/6/20,10/7/20,10/8/20,10/9/20,10/10/20,10/11/20,10/12/20,10/13/20,10/14/20,10/15/20,10
/16/20,10/17/20,10/18/20,10/19/20,10/20/20,10/21/20,10/22/20,10/23/20,10/24/20,10/25/20,10/26/20,10/27/20,10/28/20,10/29/20,10/30/20,10/31/20,11/1/20,11
/2/20,11/3/20,11/4/20,11/5/20,11/6/20,11/7/20,11/8/20,11/9/20,11/10/20,11/11/20,11/12/20,11/13/20,11/14/20,11/15/20,11/16/20,11/17/20,11/18/20,11/19/20,11
/20/20,11/21/20,11/22/20,11/23/20,11/24/20,11/25/20,11/26/20,11/27/20,11/28/20,11/29/20,11/30/20,12/1/20,12/2/20,12/3/20,12/4/20,12/5/20,12/6/20,12/7/20,12
/8/20,12/9/20,12/10/20,12/11/20,12/12/20,12/13/20,12/14/20,12/15/20,12/16/20,12/17/20,12/18/20,12/19/20,12/20/20,12/21/20,12/22/20,12/23/20,12/24/20,12
/25/20,12/26/20
,Afghanistan,33.93911,67.709953,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,2,4,4,4,4,5,7,8,11,12,13,15,16,18,20,24,2
5,29,30,34,41,43,76,80,91,107,118,146,175,197,240,275,300,338,368,424,445,485,532,556,608,666,715,785,841,907,934,997,1027,1093,1177,1236,1331,1464,1532,17
04,1830,1940,2127,2291,2470,2705,2895,3225,3393,3564,3781,4042,4403,4687,4968,5227,5640,6054,6403,6665,7073,7654,8146,8677,9219,10001,10585,11176,11834,124
59,13039,13662,14528,15208,15753,16512,17270,18057,18972,19554,20345,20920,21462,22146,22894,23550,24106,24770,25531,26314,26878,27536,27882,28428,28837,29
147,29471,29705,30165,30441,30606,30957,31228,31507,31826,32012,32314,32662,32941,33180,33374,33584,33898,34184,34356,34441,34595,34730,34984,35060,35219,3
5279,35453,35493,35605,35717,35918,35978,36026,36147,36253,36358,36463,36532,36665,36700,36701,36737,36773,36820,36928,37006,37046,37083,37153,37260,37336,
37422,37497,37542,37590,37667,37710,37750,37852,37885,37944,37990,38045,38061,38103,38119,38130,38133,38155,38159,38193,38243,38288,38304,38324,38398,38494
,38520,38544,38572,38606,38641,38716,38772,38815,38855,38872,38897,38919,39044,39074,39096,39145,39170,39186,39192,39227,39239,39254,39268,39285,39290,3929
7,39341,39422,39486,39548,39616,39693,39703,39799,39870,39928,39994,40026,40088,40141,40200,40287,40369,40510,40626,40687,40768,40833,40937,41032,41145,412
68,41334,41425,41501,41633,41728,41814,41935,41975,42033,42159,42297,42463,42609,42795,42969,43035,43240,43468,43681,43924,44177,44363,44503,44706,44988,45
174,45384,45600,45723,45844,46116,46274,46516,46718,46837,46837,47072,47306,47516,47716,47851,48053,48116,48229,48527,48718,48952,49161,49378,49621,49681,4
9817,50013,50190,50433,50655,50810,50886
```

Importing files from the internet

```
# Require curl package
> install.packages("curl")
> library(curl)

# Create an object for the link
> url <-
"https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/cs
se_covid_19_data/csse_covid_19_time_series/time_series_covid19_confir
med_global.csv"

> covid_confirmed <- read_csv(url)
> dim(covid_confirmed)
```

```
[1] 271 343
```

Importing files from the internet

```
> head(covid_confirmed)
```

```
# A tibble: 6 x 343
  `Province/State` `Country/Region`   Lat   Long `1/22/20` `1/23/20`
  <chr>            <chr>            <dbl>  <dbl>   <dbl>   <dbl>
1 NA               Afghanistan       33.9  67.7        0       0
2 NA               Albania           41.2  20.2        0       0
3 NA               Algeria           28.0   1.66       0       0
4 NA               Andorra           42.5   1.52       0       0
5 NA               Angola           -11.2  17.9        0       0
6 NA               Antigua and Bar…  17.1 -61.8        0       0 ...
```

```
> tail(covid_confirmed)
```

```
# A tibble: 6 x 343
  `Province/State` `Country/Region`      Lat  Long `1/22/20` `1/23/20`
  <chr>            <chr>               <dbl> <dbl>   <dbl>   <dbl>
1 NA               Venezuela           6.42 -66.6       0       0
2 NA               Vietnam             14.1  108.       0       2
3 NA               West Bank and G…    32.0   35.2      0       0
4 NA               Yemen               15.6   48.5      0       0
5 NA               Zambia             -13.1   27.8      0       0
6 NA               Zimbabwe           -19.0   29.2      0       0 ...
```

Writing to a file

write_csv(x, file)

x = data frame or tibble

file = file to write into

```
> write_csv(covid_confirmed, file = "covid_confirmed.csv")
```

Writing to a file

- **RDS** is R's custom binary format
- save column type specification

write_rds(x, file)

x = data frame or tibble

file = file to write into

```
> write_rds(covid_confirmed, file = "covid_confirmed.rds")
```

```
# open RDS file
> read_rds("covid_confirmed.rds")
```

Writing to a file

- **RDS** is R's custom binary format
- save column type specification

```
# open RDS file
> read_rds("covid_confirmed.rds")
```

```
# A tibble: 271 x 343
   `Province/State` `Country/Region`   Lat   Long `1/22/20` `1/23/20`
   <chr>            <chr>            <dbl>  <dbl>    <dbl>    <dbl>
 1 NA               Afghanistan       33.9  67.7         0        0
 2 NA               Albania           41.2  20.2         0        0
 3 NA               Algeria           28.0   1.66        0        0
 4 NA               Andorra           42.5   1.52        0        0
 5 NA               Angola           -11.2  17.9         0        0
 ...
```

Take-away message

- readr package of tidyverse is useful for reading and writing textfiles
- readxl for importing Excel files in R
- use R to fetch files from the internet