Show me the data!

Week06: Loop

Big Data & Social Analysis R

Instructors: Chung-pei Pien

ZU1942001/266868001/Z23937001/ZM1941001

International College of INNOVATION
National Chengchi University
國立政治大學創新國際學院

CONTENTS

1 Homework 04

2 Loop

3 JSON and Loop

4 Read Files and Loop

5 Assignment

Homework 04

01

Homework 04

```
ntc_ref <- ntc_ref_2018 %>%
  left_join(ntc_ref_2021[, -1], by = "li_id")
```

```
ntc_ref <- ntc_ref_2018 %>%
  left_join(ntc_ref_2021, by = "li_id")
  #Two district names
```

```
ntc_ref <- ntc_ref_2018 %>%
  left_join(ntc_ref_2021, by = c("district", "li_id")
  #If two tables’ districts have different content?!
```

01 Homework 04

```
ntc_dist <- ntc_ref %>%
  group_by(district) %>%
  summarise(rf_16_yea = sum(rf_16_yea),
            rf_16_valid_vote = sum(rf_16_valid_vote),
            rf_17_yea = sum(rf_17_yea),
            rf_17_valid_vote = sum(rf_17_valid_vote),
            yearate_2018 = rf_16_yea /
       rf_16_valid_vote,
            yearate_2021 = rf_17_yea /
       rf_17_valid_vote,
            gap = yearate_2021 - yearate_2018,
            gap_rate = (rf_17_yea - rf_16_yea) /
       rf_17_valid_vote)
```

Loop

Loop

Introduction

Loop:
Repeat a set of actions N times

All skills you learned can put into a loop

02

Loop

Introduction

```
for (i in 1:10) {
    print(i)
}
```

02

Loop

Introduction

```
a <- 3

for (i in 1:10) {

    a <- a + i

}
```

02

Loop

Introduction

```
ici <- c("nccu", "taiwan", "best", "lovely")

for (i in 1:5) {

     print(ici[i])

}
```

02

Loop

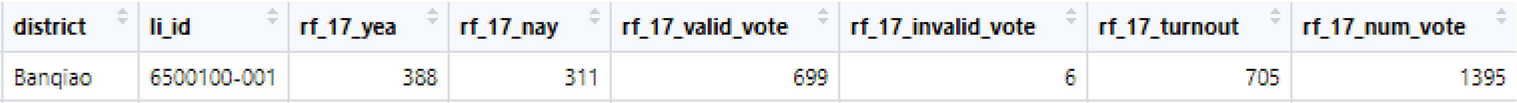
Introduction

```
ici <- c("nccu", "taiwan", "best", "lovely")

for (i in 1:length(ici)) {

     print(ici[i])

}
```

02

Loop

Introduction

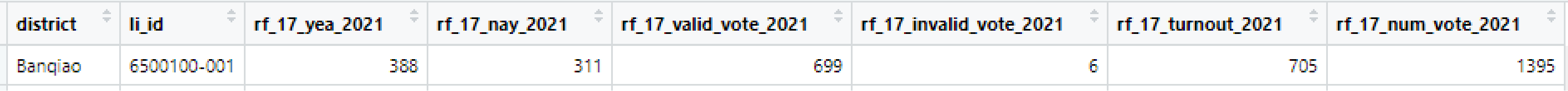


district	li_id	rf_17_yea	rf_17_nay	rf_17_valid_vote	rf_17_invalid_vote	rf_17_turnout	rf_17_num_vote
Banqiao	6500100-001	388	311	699	6	705	1395

Add “_2021” in the end of 3-8 column names of ntc_ref_2018:

rf_17_yea → rf_17_yea_2021
rf_17_nay → rf_17_nay_2021

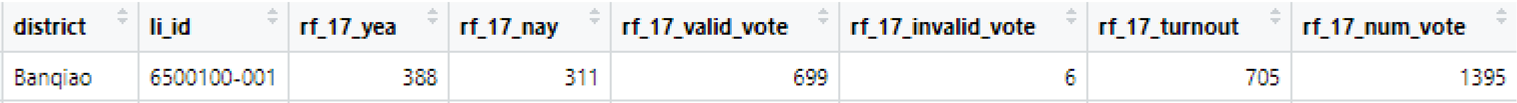
```
colnames(ntc_ref_2021)[3:8] <-
paste(colnames(ntc_ref_2021)[3:8], "2021", sep = "_")
```



district	li_id	rf_17_yea_2021	rf_17_nay_2021	rf_17_valid_vote_2021	rf_17_invalid_vote_2021	rf_17_turnout_2021	rf_17_num_vote_2021
Banqiao	6500100-001	388	311	699	6	705	1395

02 Loop

Introduction

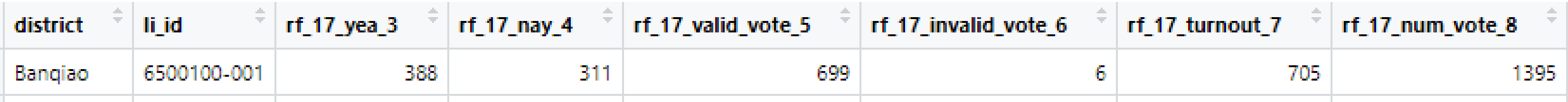


district	li_id	rf_17_yea	rf_17_nay	rf_17_valid_vote	rf_17_invalid_vote	rf_17_turnout	rf_17_num_vote
Banqiao	6500100-001	388	311	699	6	705	1395

```
ntc_ref_2021 <- read_xlsx("ntc_ref_2021.xlsx")
```

Add _ and column number in the end of 3-8 column names of ntc_ref_2018:

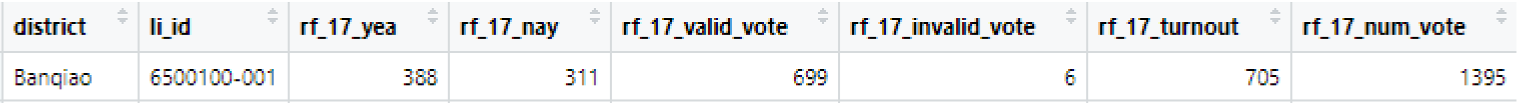
rf_17_yea → rf_17_yea_3
rf_17_nay → rf_17_nay_4
rf_17_valid_vote → rf_17_valid_vote_5



district	li_id	rf_17_yea_3	rf_17_nay_4	rf_17_valid_vote_5	rf_17_invalid_vote_6	rf_17_turnout_7	rf_17_num_vote_8
Banqiao	6500100-001	388	311	699	6	705	1395

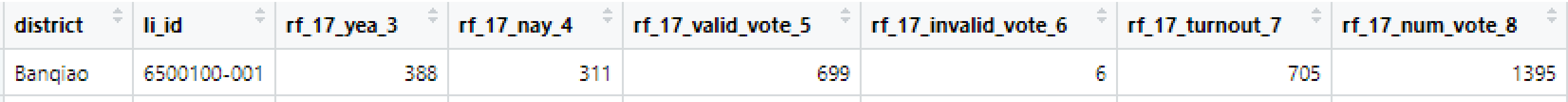
02 Loop

Introduction



district	li_id	rf_17_yea	rf_17_nay	rf_17_valid_vote	rf_17_invalid_vote	rf_17_turnout	rf_17_num_vote
Banqiao	6500100-001	388	311	699	6	705	1395

```
for (i in 3:8) {

  colnames(ntc_ref_2021)[i] <-
        paste(colnames(ntc_ref_2021)[i], i, sep = "_")
}
```



district	li_id	rf_17_yea_3	rf_17_nay_4	rf_17_valid_vote_5	rf_17_invalid_vote_6	rf_17_turnout_7	rf_17_num_vote_8
Banqiao	6500100-001	388	311	699	6	705	1395

JSON and Loop

03

JSON and Loop

Introduction

JSON format file:

Besides csv and xlsx, many websites provide JSON files.

What special JSON is?

JSON can stores complicated data: Nested data.

For example: A JSON have many lists, and every list involves a table.

03

JSON and Loop

New Taipei City Referendum Raw Data

We dealt with house_115 many weeks.

Do you know where the original file came from?

ProPublica Congress API: congress-115.json

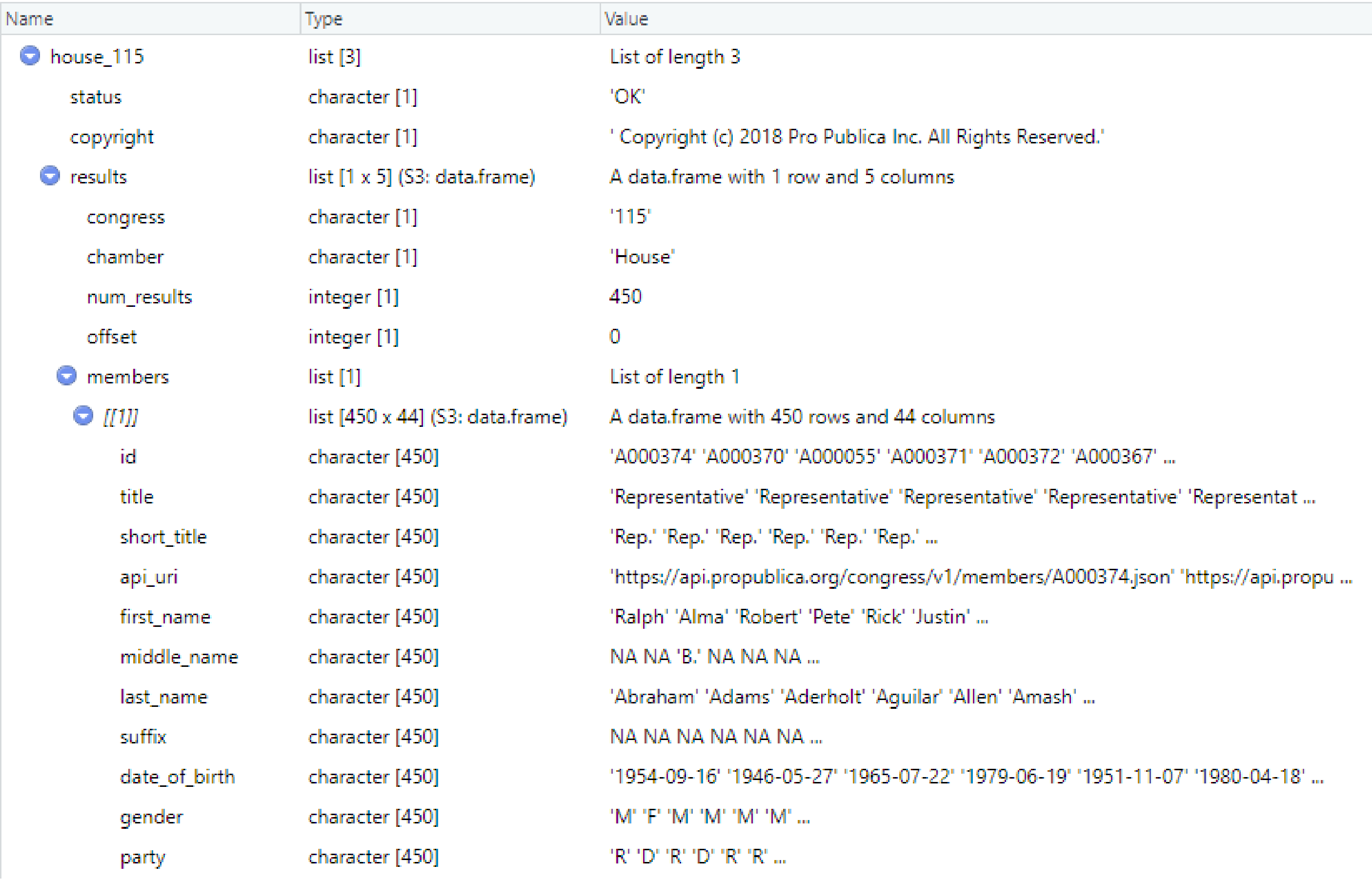
```
library(jsonlite)

house_115_json <- fromJSON("congress-115.json")
```

03

JSON and Loop

New Taipei City Referendum Raw Data

House_115 list → results sublist → members sublist → member sublist → [[1]] sublist



Name	Type	Value
house_115	list [3]	List of length 3
status	character [1]	'OK'
copyright	character [1]	' Copyright (c) 2018 Pro Publica Inc. All Rights Reserved.'
results	list [1 x 5] (S3: data.frame)	A data.frame with 1 row and 5 columns
congress	character [1]	'115'
chamber	character [1]	'House'
num_results	integer [1]	450
offset	integer [1]	0
members	list [1]	List of length 1
[[1]]	list [450 x 44] (S3: data.frame)	A data.frame with 450 rows and 44 columns
id	character [450]	'A000374' 'A000370' 'A000055' 'A000371' 'A000372' 'A000367' ...
title	character [450]	'Representative' 'Representative' 'Representative' 'Representative' 'Representat ...
short_title	character [450]	'Rep.' 'Rep.' 'Rep.' 'Rep.' 'Rep.' 'Rep.' ...
api_uri	character [450]	'https://api.propublica.org/congress/v1/members/A000374.json' 'https://api.propu ...
first_name	character [450]	'Ralph' 'Alma' 'Robert' 'Pete' 'Rick' 'Justin' ...
middle_name	character [450]	NA NA 'B.' NA NA NA ...
last_name	character [450]	'Abraham' 'Adams' 'Aderholt' 'Aguilar' 'Allen' 'Amash' ...
suffix	character [450]	NA NA NA NA NA NA ...
date_of_birth	character [450]	'1954-09-16' '1946-05-27' '1965-07-22' '1979-06-19' '1951-11-07' '1980-04-18' ...
gender	character [450]	'M' 'F' 'M' 'M' 'M' 'M' ...
party	character [450]	'R' 'D' 'R' 'D' 'R' 'R' ...

03

JSON and Loop

New Taipei City Referendum Raw Data

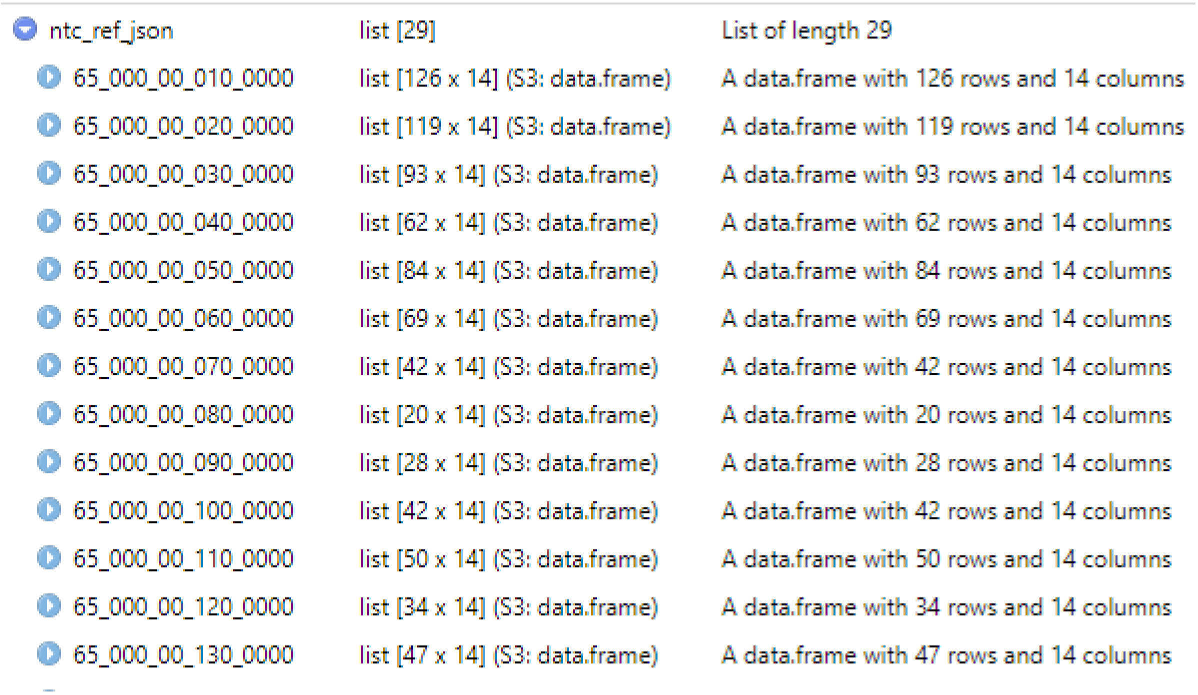
Let’s read a simple json file: 17_65000.json

```
ntc_ref_json <- fromJSON("17_65000.json")
```

03 JSON and Loop

New Taipei City Referendum Raw Data

ntc_ref_json <- fromJSON("17_65000.json")



```
ntc_ref_json          list [29]                          List of length 29
  65_000_00_010_0000  list [126 x 14] (S3: data.frame)   A data.frame with 126 rows and 14 columns
  65_000_00_020_0000  list [119 x 14] (S3: data.frame)   A data.frame with 119 rows and 14 columns
  65_000_00_030_0000  list [93 x 14] (S3: data.frame)    A data.frame with 93 rows and 14 columns
  65_000_00_040_0000  list [62 x 14] (S3: data.frame)    A data.frame with 62 rows and 14 columns
  65_000_00_050_0000  list [84 x 14] (S3: data.frame)    A data.frame with 84 rows and 14 columns
  65_000_00_060_0000  list [69 x 14] (S3: data.frame)    A data.frame with 69 rows and 14 columns
  65_000_00_070_0000  list [42 x 14] (S3: data.frame)    A data.frame with 42 rows and 14 columns
  65_000_00_080_0000  list [20 x 14] (S3: data.frame)    A data.frame with 20 rows and 14 columns
  65_000_00_090_0000  list [28 x 14] (S3: data.frame)    A data.frame with 28 rows and 14 columns
  65_000_00_100_0000  list [42 x 14] (S3: data.frame)    A data.frame with 42 rows and 14 columns
  65_000_00_110_0000  list [50 x 14] (S3: data.frame)    A data.frame with 50 rows and 14 columns
  65_000_00_120_0000  list [34 x 14] (S3: data.frame)    A data.frame with 34 rows and 14 columns
  65_000_00_130_0000  list [47 x 14] (S3: data.frame)    A data.frame with 47 rows and 14 columns
```

There are 29 lists: every list has a dataframe. Every dataframe records every district’s Case 17 reults at li level.

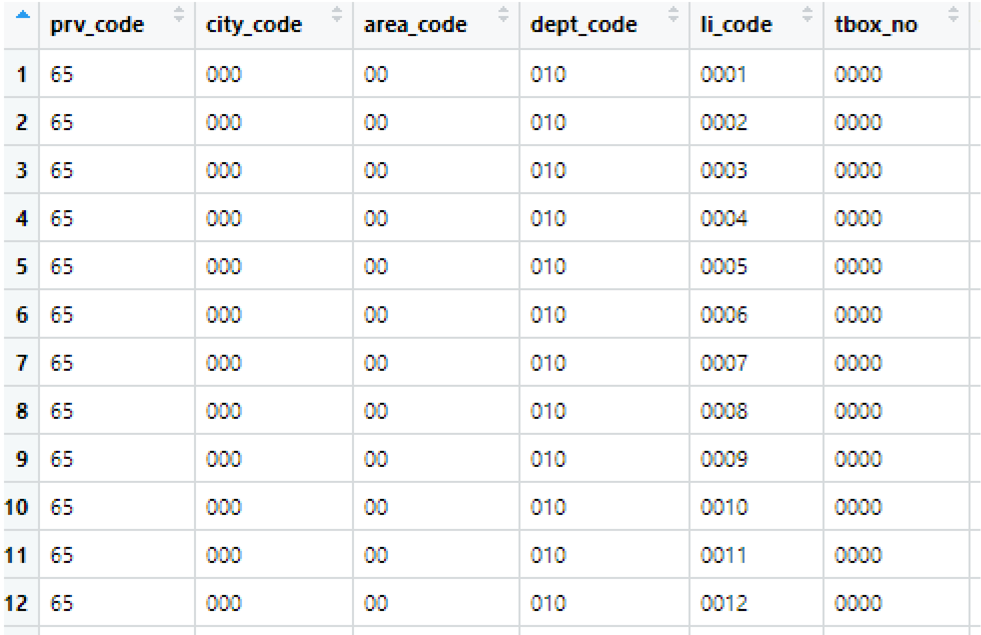
03 JSON and Loop

New Taipei City Referendum Raw Data

ntc_ref_json[[1]]

We can read the first list’s content.

ntc_ref_1 <- ntc_ref_json[[1]]



	prv_code	city_code	area_code	dept_code	li_code	tbox_no
1	65	000	00	010	0001	0000
2	65	000	00	010	0002	0000
3	65	000	00	010	0003	0000
4	65	000	00	010	0004	0000
5	65	000	00	010	0005	0000
6	65	000	00	010	0006	0000
7	65	000	00	010	0007	0000
8	65	000	00	010	0008	0000
9	65	000	00	010	0009	0000
10	65	000	00	010	0010	0000
11	65	000	00	010	0011	0000
12	65	000	00	010	0012	0000

03

JSON and Loop

New Taipei City Referendum Raw Data

```
ntc_ref_1 <- ntc_ref_json[[1]]

ntc_ref_2 <- ntc_ref_json[[2]]

ntc_ref_2021 <- rbind(ntc_ref_1, ntc_ref_2)
```

Therefore, if we can create a loop, we can rbind() all dataframes in ntc_ref_json.

03 JSON and Loop

New Taipei City Referendum Raw Data

```
ntc_ref_2021 <- data.frame() #Erease ntc_ref_2021 or create a new empty dataframe

for (i in 1:29) {

~~ ntc_ref_1 <- ntc_ref_json[[1]]~~
  temp_df <- ntc_ref_json[[i]]

~~ ntc_ref_2021 <- rbind(ntc_ref_1, ntc_ref_2)~~
  ntc_ref_2021 <- rbind(ntc_ref_2021, temp_df)

}
```

JSON and Loop

New Taipei City Referendum Raw Data

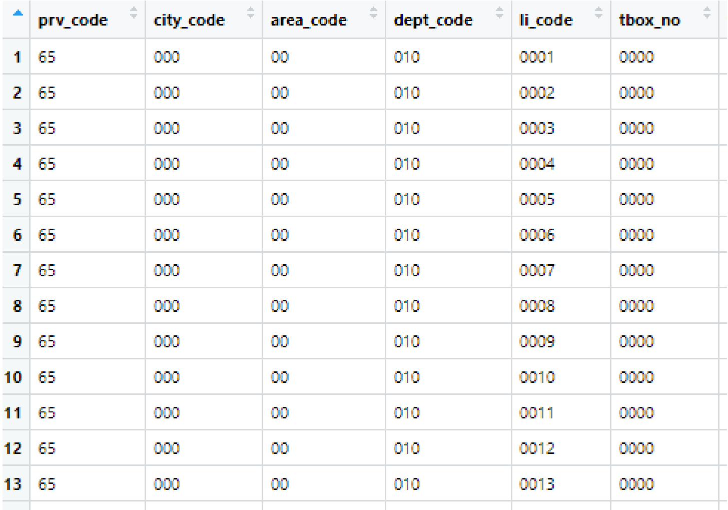
```
for (i in 1:29) {

  temp_df <- ntc_ref_json[[i]]
  ntc_ref_2021 <- rbind(ntc_ref_2021, temp_df)

}

i <- 1
```

03 JSON and Loop

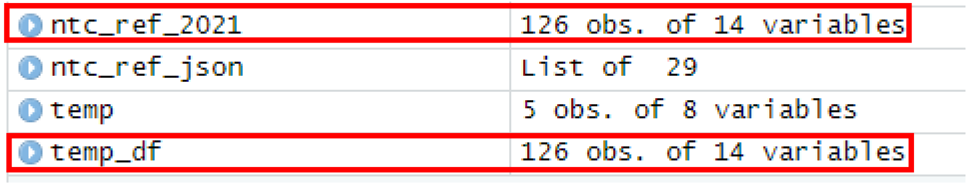
New Taipei City Referendum Raw Data

i <- 1

ntc_ref_2021 <- data.frame()
~~for (i in 1:29) {~~
temp_df <- ntc_ref_json[[i]]
ntc_ref_2021 <- rbind(ntc_ref_2021, temp_df)
}



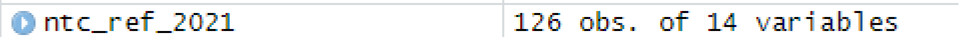
	prv_code	city_code	area_code	dept_code	li_code	tbox_no
1	65	000	00	010	0001	0000
2	65	000	00	010	0002	0000
3	65	000	00	010	0003	0000
4	65	000	00	010	0004	0000
5	65	000	00	010	0005	0000
6	65	000	00	010	0006	0000
7	65	000	00	010	0007	0000
8	65	000	00	010	0008	0000
9	65	000	00	010	0009	0000
10	65	000	00	010	0010	0000
11	65	000	00	010	0011	0000
12	65	000	00	010	0012	0000
13	65	000	00	010	0013	0000



ntc_ref_2021	126 obs. of 14 variables
ntc_ref_json	List of 29
temp	5 obs. of 8 variables
temp_df	126 obs. of 14 variables

03 JSON and Loop

New Taipei City Referendum Raw Data



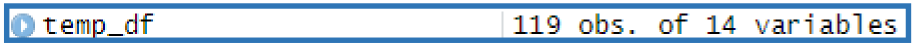
```
ntc_ref_2021        126 obs. of 14 variables
```

i <- 2

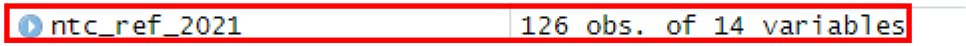
~~ntc_ref_2021 <- data.frame()~~

~~for (i in 1:29) {~~

temp_df <- ntc_ref_json[[i]]



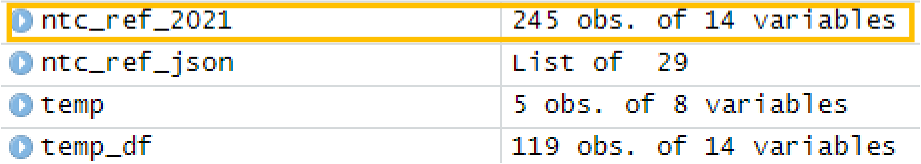
```
temp_df        119 obs. of 14 variables
```



```
ntc_ref_2021        126 obs. of 14 variables
```

ntc_ref_2021 <- rbind(ntc_ref_2021, temp_df)

}



```
ntc_ref_2021        245 obs. of 14 variables
ntc_ref_json        List of  29
temp                5 obs. of 8 variables
temp_df             119 obs. of 14 variables
```

03

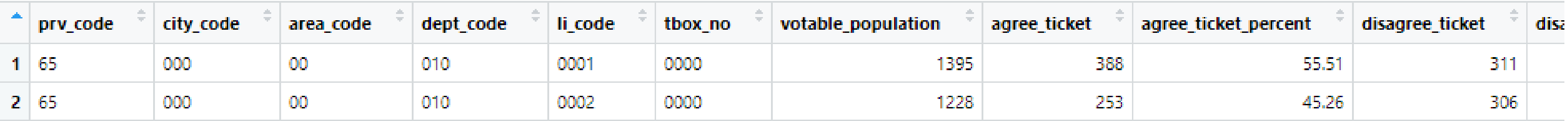
JSON and Loop

New Taipei City Referendum Raw Data

```
ntc_ref_2021 <- data.frame()

for (i in 1:29) {

  temp_df <- ntc_ref_json[[i]]

  ntc_ref_2021 <- rbind(ntc_ref_2021, temp_df)

}
```

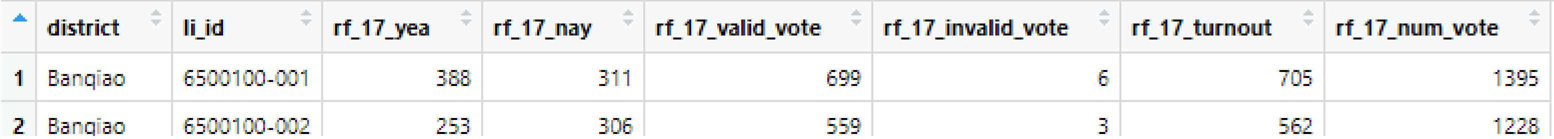
03 JSON and Loop

New Taipei City Referendum Raw Data



	prv_code	city_code	area_code	dept_code	li_code	tbox_no	votable_population	agree_ticket	agree_ticket_percent	disagree_ticket	disa
1	65	000	00	010	0001	0000	1395	388	55.51	311	
2	65	000	00	010	0002	0000	1228	253	45.26	306	

Comparing ntc_ref_2021 you create this week and ntc_ref_2021 in Homework 04, what’s different?

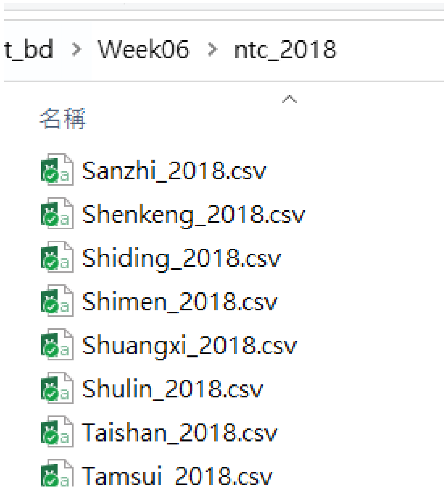


	district	li_id	rf_17_yea	rf_17_nay	rf_17_valid_vote	rf_17_invalid_vote	rf_17_turnout	rf_17_num_vote
1	Banqiao	6500100-001	388	311	699	6	705	1395
2	Banqiao	6500100-002	253	306	559	3	562	1228

Read Files and Loop

04 Read Files and Loop

Download ntc_2018.zip and unzip it into your working directory.



t_bd › Week06 › ntc_2018

名稱

- Sanzhi_2018.csv
- Shenkeng_2018.csv
- Shiding_2018.csv
- Shimen_2018.csv
- Shuangxi_2018.csv
- Shulin_2018.csv
- Taishan_2018.csv
- Tamsui_2018.csv

We have 29 files of New Taipei City districts’ 2018 election results at li level.

Use loop to read and rbind them!

04 Read Files and Loop

How to get files’ name in popu_edu_inc folder?

```
ntc_2018_list <- list.files("ntc_2018")
```

```
ntc_2018_1 <- read.csv(ntc_2018_list[1]) #error!!!!!
```

Why? Because your files are in **ntc_2018** folder.

04 Read Files and Loop

```
read.csv(ntc_2018_list[1])
```

You should provide correct path and file name in read.csv.

```
ntc_2018_1 <- read.csv("ntc_2018/ntc_2018_list[1]")
```

Error AGAIN!!!!

Why? R does not read any strings in "" as an object. There is no file named ntc_2018_list[1] in ntc_2018.

04 Read Files and Loop

```
ntc_2018_1 <- read.csv("ntc_2018/ntc_2018_list[1]")

paste("ntc_2018/", ntc_2018_list[1], sep = "")
paste0("ntc_2018/", ntc_2018_list[1])

ntc_2018_1 <- read.csv (paste0("ntc_2018/", ntc_2018_list[1]))
```

04 Read Files and Loop

```
ntc_2018_1 <- read.csv (paste0("ntc_2018/", ntc_2018_list[1]))

ntc_2018_2 <- read.csv (paste0("ntc_2018/", ntc_2018_list[2]))

ntc_2018 <- rbind(ntc_2018_1, ntc_2018_2)
```

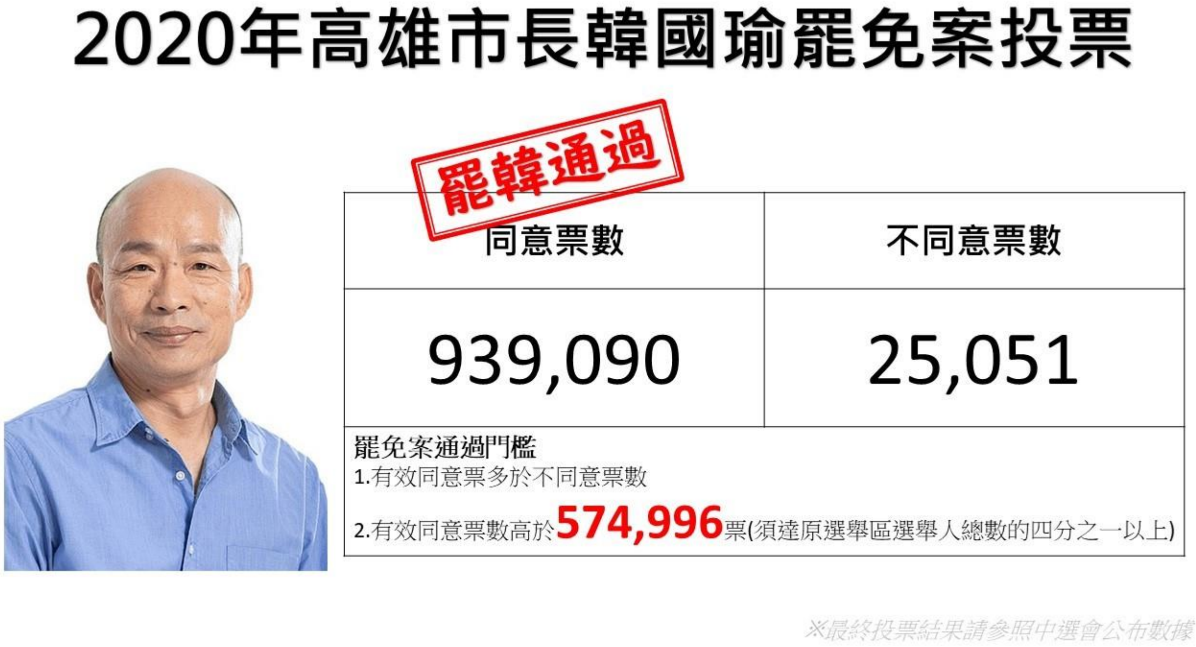
04 Read Files and Loop

Use loop function and codes of ntc_ref_2021 to rbind all districts’ 2018 election results into ntc_2018 (Practice 1)

Assignment

05 Assignment

In June 2020, Kaohsiung held a mayoral recall election. Finally, the recall was successful. Han Kuo-yu, who was elected as the mayor in 2018, was recalled.



2020年高雄市長韓國瑜罷免案投票

罷韓通過

同意票數	不同意票數
939,090	25,051
罷免案通過門檻 1.有效同意票多於不同意票數 2.有效同意票數高於**574,996**票(須達原選舉區選舉人總數的四分之一以上)	

※最終投票結果請參照中選會公布數據

Unzip kh_recall.zip, which involve all districts/towns' recall data.