

Show  
me the  
data!

Week04: Advanced Dataframe Manipulation

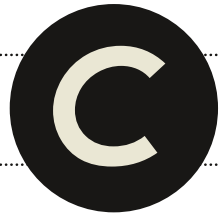
# Big Data & Social Analysis R

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 dplyr
- 2 Table Merge
- 3 Assignment

# dplyr

**dplyr is the most powerful package to help you to clean data.**

**dplyr is called**

**“grammar of data manipulation”**

**Pipeline: %>%**

**%>% is like **Relative Pronouns** in English.**

**You can modify an object and omit it.**

```
ici <- "best"
```

```
b <- length(ici)
```

```
a <- b
```

```
a <- b %>%  
  length()
```

**Last few weeks, we employed the following tasks by basic R functions:**

- 1. Get a dataframe's rows and columns**
- 2. Create new variables**
- 3. Calculate subsets' mean, sum, etc**

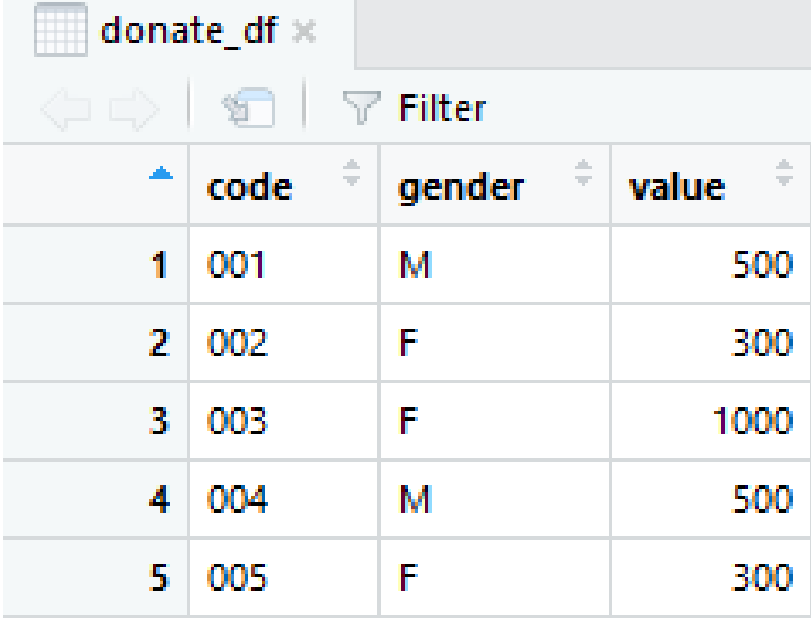
```
mean(donate_df$value[donate_df$gender == "M"])
```

**Too long, too complicated and hard to read.....**

## 01

# dplyr

## Key Functions in dplyr



	code	gender	value
1	001	M	500
2	002	F	300
3	003	F	1000
4	004	M	500
5	005	F	300

Create a table called male\_donate: Only male observations

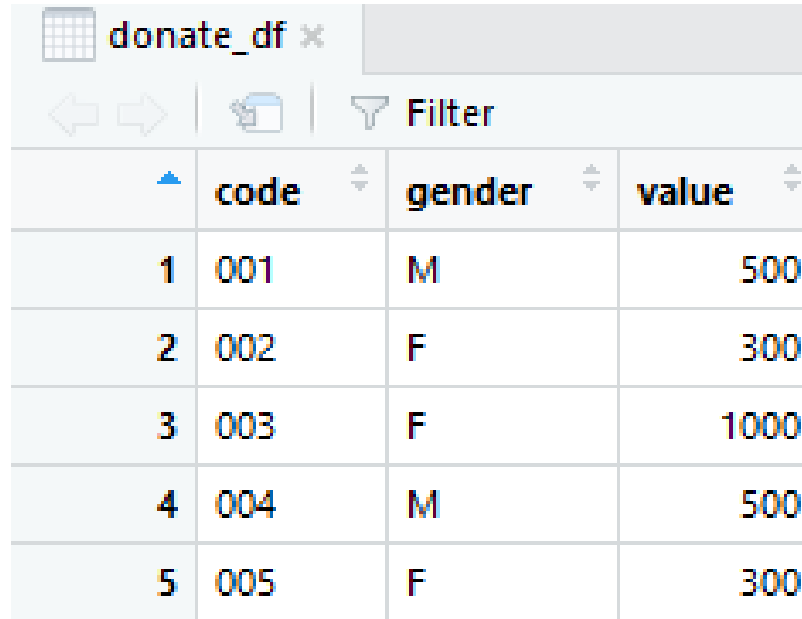
```
male_donate <- donate_df[donate_df$gender == "M",]
```

```
male_donate <- donate_df %>%  
  filter(donate_df, gender == "M")
```

## 01

# dplyr

## Key Functions in dplyr



	code	gender	value
1	001	M	500
2	002	F	300
3	003	F	1000
4	004	M	500
5	005	F	300

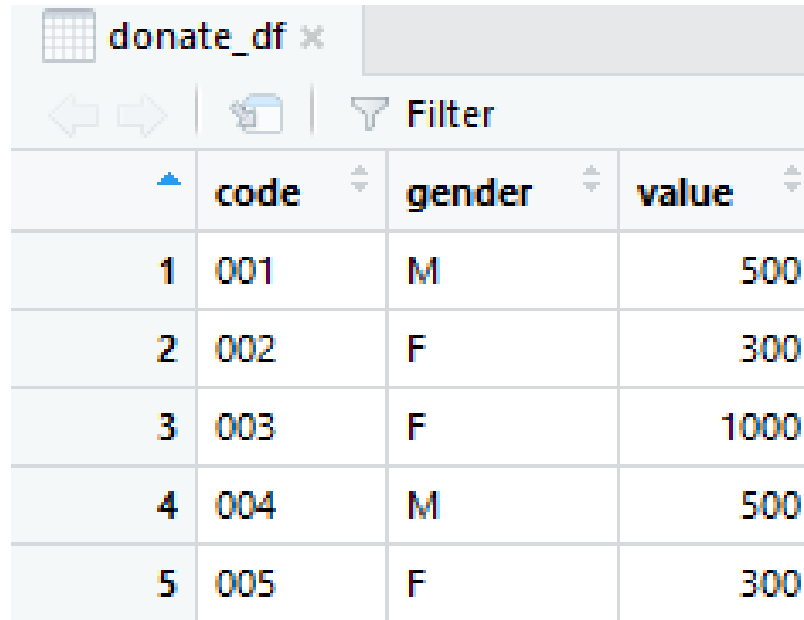
Create a table called value\_500: Observations' value is larger than 500



## 01

# dplyr

## Key Functions in dplyr



	code	gender	value
1	001	M	500
2	002	F	300
3	003	F	1000
4	004	M	500
5	005	F	300

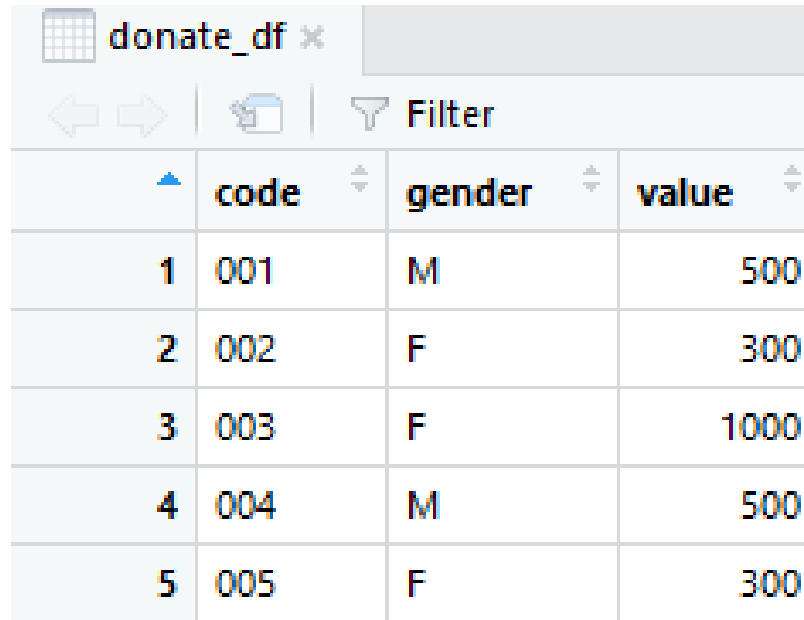
Create a table called value\_500: Observations' value is larger than 500

```
value_500 <- donate_df %>%  
  filter(donate_df, value >= 500)
```

## 01

# dplyr

## Key Functions in dplyr



	code	gender	value
1	001	M	500
2	002	F	300
3	003	F	1000
4	004	M	500
5	005	F	300

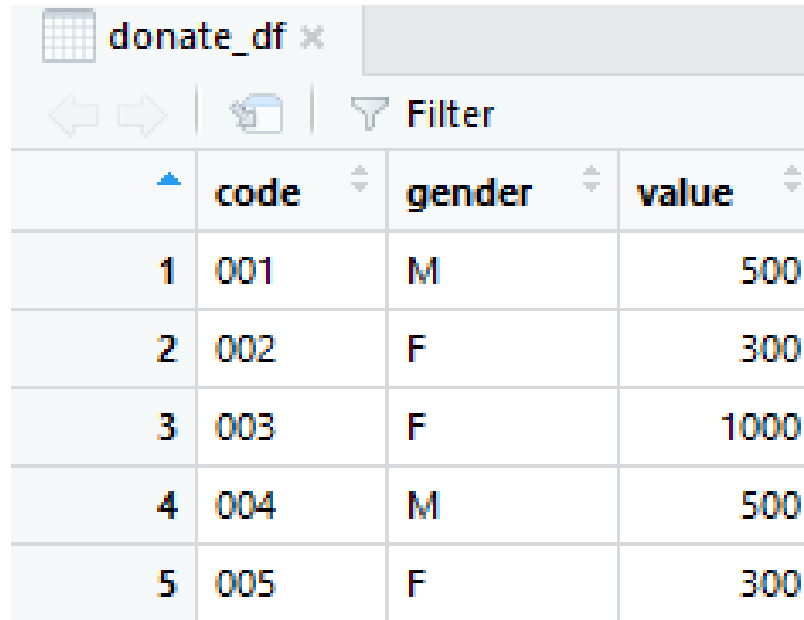
Create a table called donate\_1: delete code column

```
donate_1 <- donate_df %>%  
  select (donate_df, -code)
```

## 01

# dplyr

## Key Functions in dplyr



	code	gender	value
1	001	M	500
2	002	F	300
3	003	F	1000
4	004	M	500
5	005	F	300

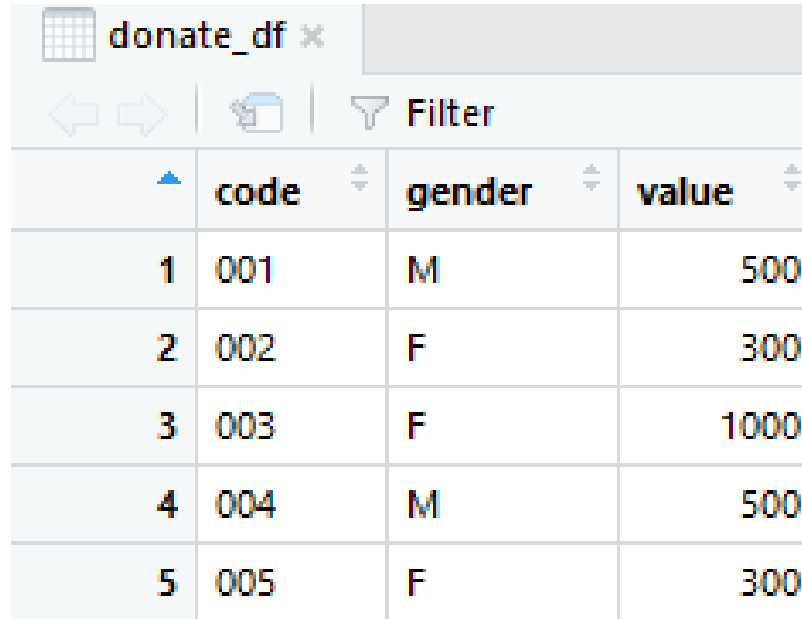
Create a table called `donate_2`: Only male observations and delete code column

```
donate_2 <- donate_df %>%  
  filter(value >= 500) %>%  
  select (-code)
```

## 01

# dplyr

## Key Functions in dplyr



	code	gender	value
1	001	M	500
2	002	F	300
3	003	F	1000
4	004	M	500
5	005	F	300

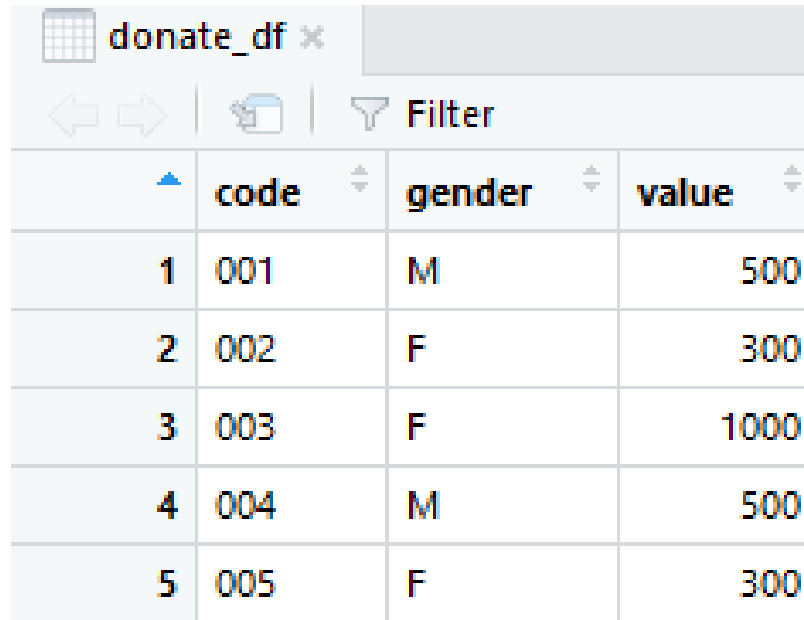
Create a variable called `usd`: Exchange value (TW) to US dollar

```
donate_df$usd <- donate_df$value / 28
```

## 01

# dplyr

## Key Functions in dplyr



	code	gender	value
1	001	M	500
2	002	F	300
3	003	F	1000
4	004	M	500
5	005	F	300

Create a variable called `usd`: Exchange value (TW) to US dollar

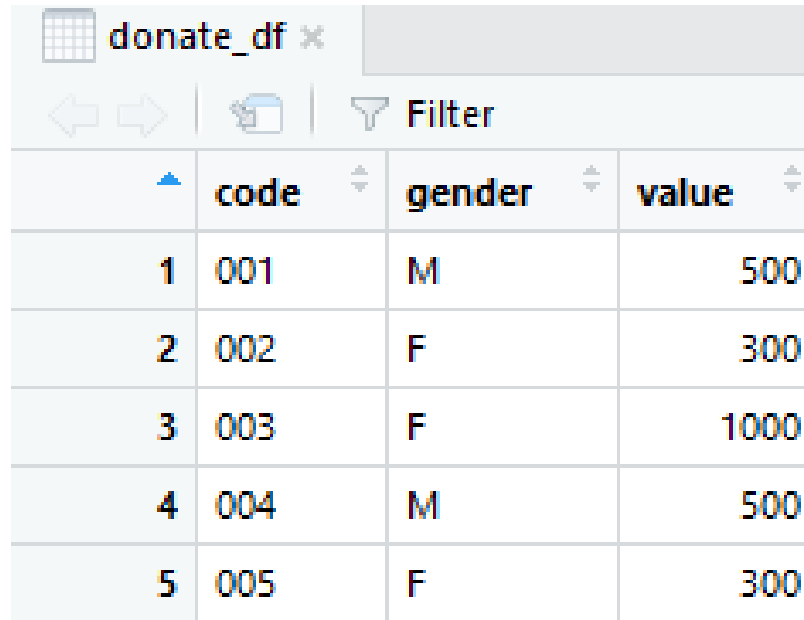
```
donate_df$usd <- donate_df$value / 28
```

```
donate_df <- donate_df %>%  
  mutate(donate_df, usd = donate_df$value / 28)
```

## 01

# dplyr

## Key Functions in dplyr



The screenshot shows a data table viewer for a dataset named 'donate\_df'. The table has three columns: 'code', 'gender', and 'value'. The rows are numbered 1 through 5. The 'code' column contains values 001, 002, 003, 004, and 005. The 'gender' column contains values M, F, F, M, and F. The 'value' column contains values 500, 300, 1000, 500, and 300.

	code	gender	value
1	001	M	500
2	002	F	300
3	003	F	1000
4	004	M	500
5	005	F	300

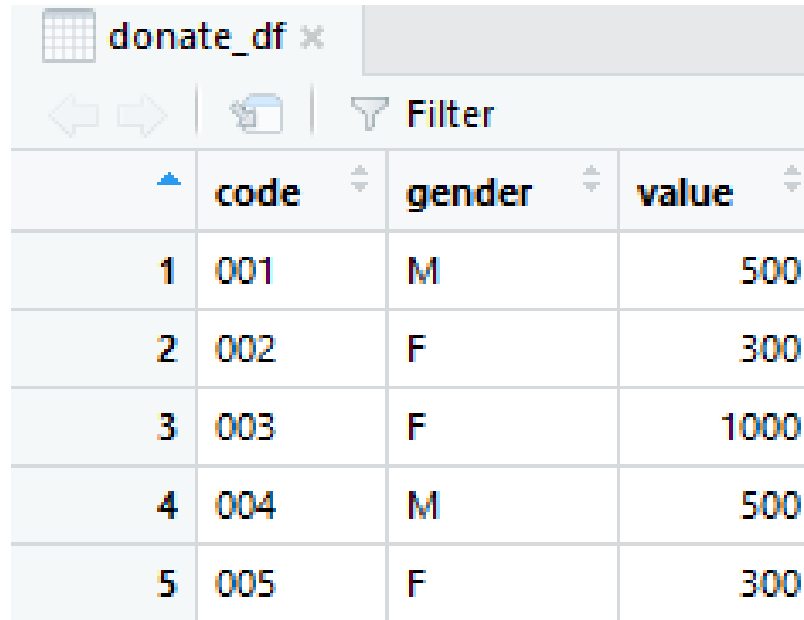
Create a variable called `usd`: Exchange value (TW) to US dollar

```
donate_df <- donate_df %>%  
  mutate(usd = value / 28,  
         jpn = value * 4.12)
```

## 01

# dplyr

## Key Functions in dplyr



	code	gender	value
1	001	M	500
2	002	F	300
3	003	F	1000
4	004	M	500
5	005	F	300

Create a variable called sex: Female is 1 and male is 0

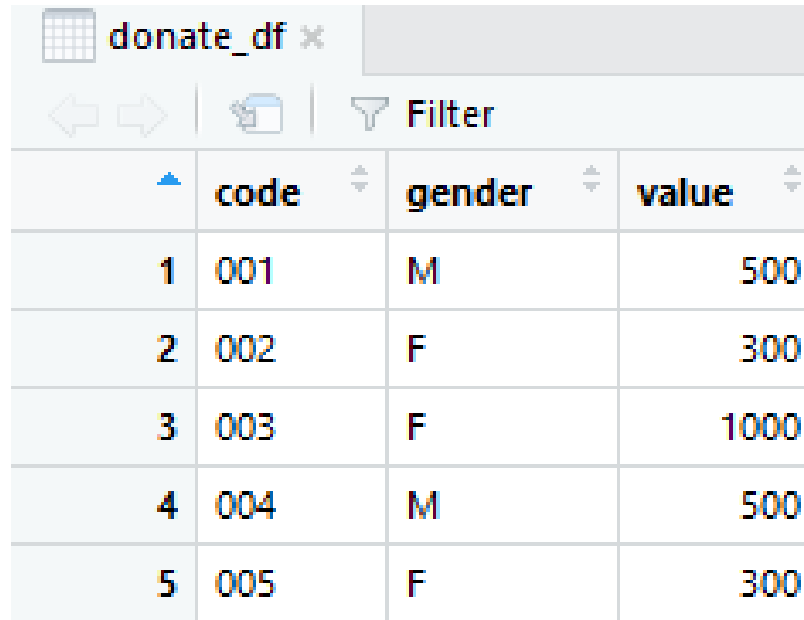
```
donate_df$sex <- 0
```

```
donate_df$sex[donate_df$gender == "F"] <- 1
```

## 01

# dplyr

## Key Functions in dplyr



	code	gender	value
1	001	M	500
2	002	F	300
3	003	F	1000
4	004	M	500
5	005	F	300

Create a variable called sex: Female is 1 and male is 0

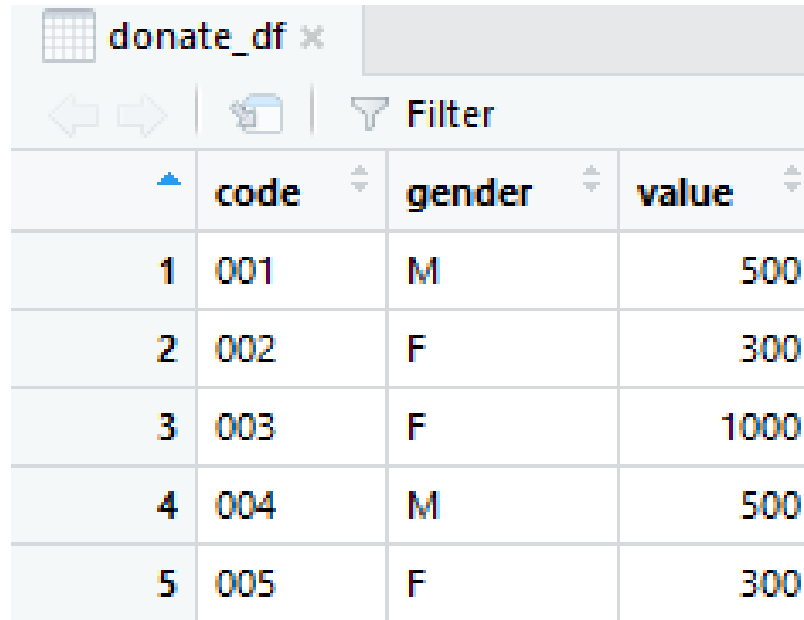
```
donate_df <- donate_df %>%  
  mutate(sex = case_when(gender == "M" ~ 0  
                          True ~ 1))
```



## 01

# dplyr

## Key Functions in dplyr



	code	gender	value
1	001	M	500
2	002	F	300
3	003	F	1000
4	004	M	500
5	005	F	300

Create a variable called donate: value  $\geq 500$  is 2, and value small than 500 is 1

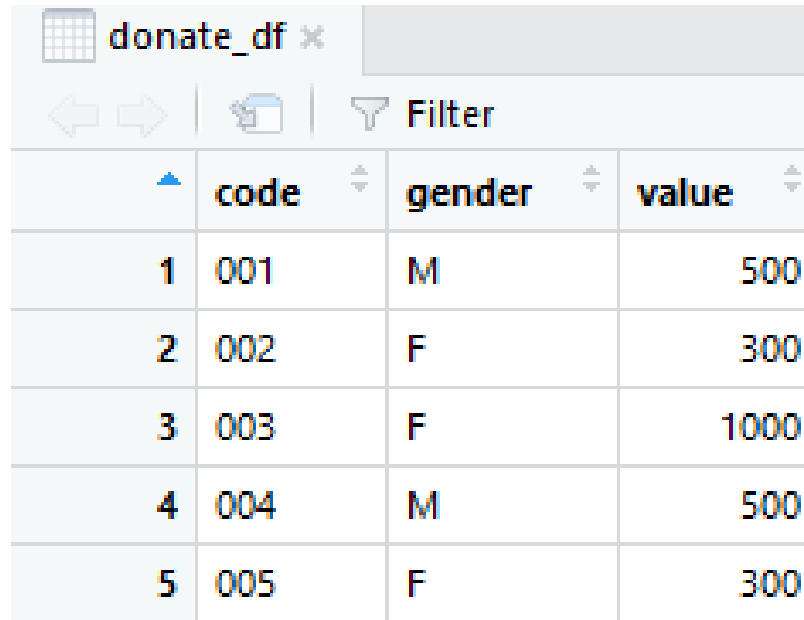
```
donate_df$donate <- 1
```

```
donate_df$donate[donate_df$value  $\geq$  500] <- 2
```

## 01

# dplyr

## Key Functions in dplyr



	code	gender	value
1	001	M	500
2	002	F	300
3	003	F	1000
4	004	M	500
5	005	F	300

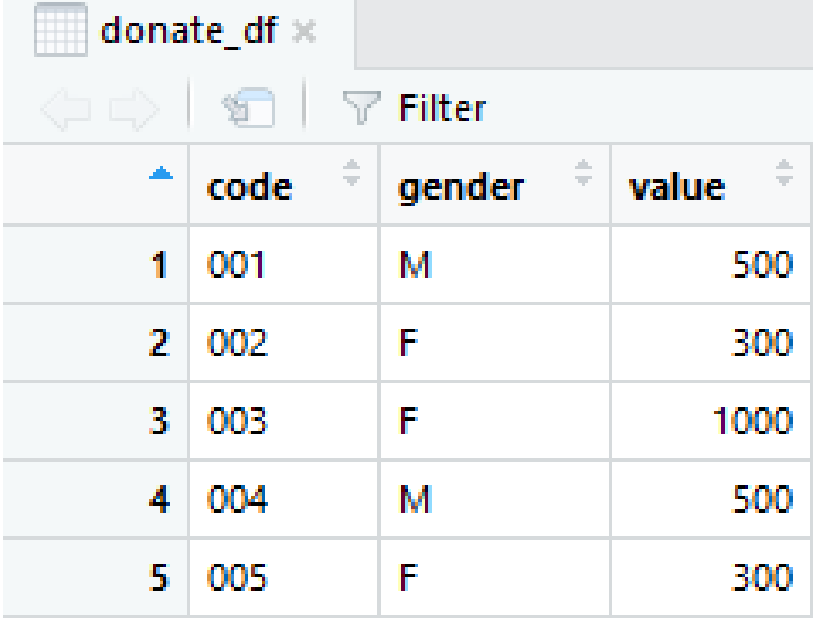
Create a variable called donate: value  $\geq 500$  is 2, and value small than 500 is 1

```
donate_df <- donate_df %>%  
  mutate(donate = case_when(value >= 500 ~ 2,  
                             True ~ 1))
```

## 01

# dplyr

## Key Functions in dplyr



The screenshot shows a data frame viewer for a table named 'donate\_df'. The table has 5 rows and 4 columns. The columns are 'code', 'gender', and 'value'. The rows are indexed 1 to 5. The 'value' column contains the values 500, 300, 1000, 500, and 300 respectively.

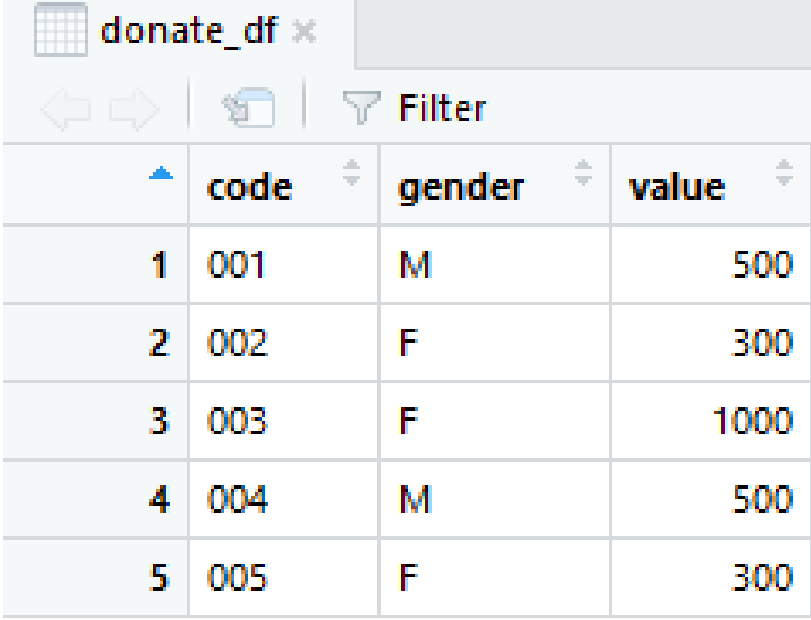
	code	gender	value
1	001	M	500
2	002	F	300
3	003	F	1000
4	004	M	500
5	005	F	300

Create a variable called donate: value  $\geq 1000$  is 3,  $\leq 500$  and  $< 1000$  is 2, and value  $< 500$  is 1

## 01

# dplyr

## Key Functions in dplyr



	code	gender	value
1	001	M	500
2	002	F	300
3	003	F	1000
4	004	M	500
5	005	F	300

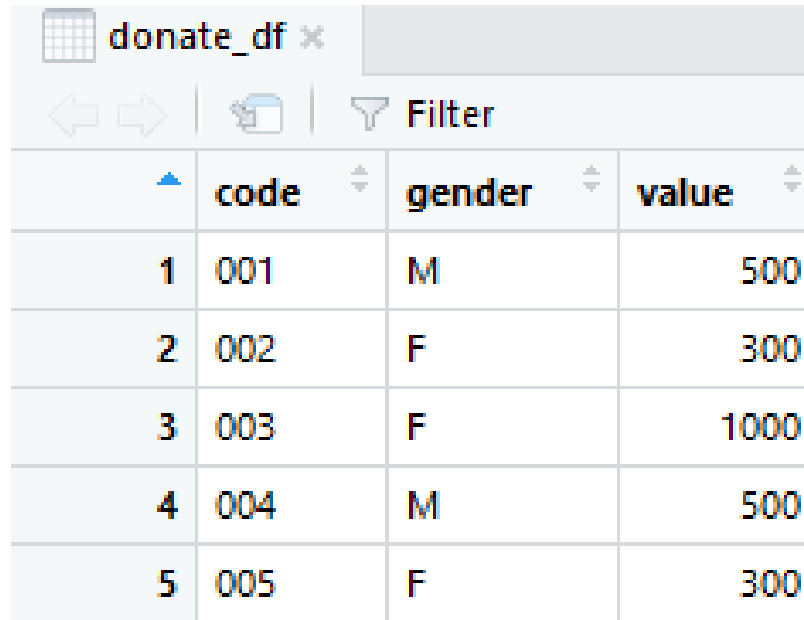
Create a variable called donate: value  $\geq 1000$  is 3, value  $\geq 500$  and  $< 1000$  is 2, and value  $< 500$  is 1

```
donate_df <- donate_df %>%  
  mutate(donate = case_when(value >= 1000 ~ 3,  
                             value >= 500 & < 1000 ~ 2,  
                             True ~ 1))
```

## 01

# dplyr

group\_by and summarise



	code	gender	value
1	001	M	500
2	002	F	300
3	003	F	1000
4	004	M	500
5	005	F	300

Compare male and female's donation

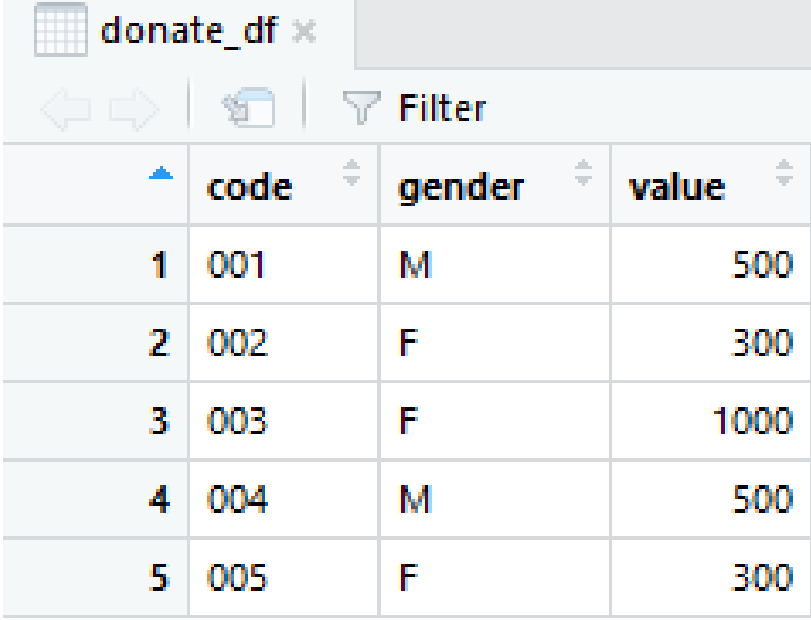
```
mean(donate_df$value[donate_df$gender == "M"])
```

```
mean(donate_df$value[donate_df$gender == "F"])
```

## 01

## dplyr

group\_by and summarise



	code	gender	value
1	001	M	500
2	002	F	300
3	003	F	1000
4	004	M	500
5	005	F	300

Compare male and female's donation

```
gender_df <- donate_df %>%  
  group_by(gender) %>%  
  summarise(donate = mean(value))
```

## Let's go back to house\_115

id	title	short_title	api_uri	first_name	last_name	date_of_birth	gender	
1	A000374	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000374.j...">https://api.propublica.org/congress/v1/members/A000374.j...</a>	Ralph	Abraham	1954-09-16	M
2	A000370	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000370.j...">https://api.propublica.org/congress/v1/members/A000370.j...</a>	Alma	Adams	1946-05-27	F
3	A000055	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000055.j...">https://api.propublica.org/congress/v1/members/A000055.j...</a>	Robert	Aderholt	1965-07-22	M
4	A000371	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000371.j...">https://api.propublica.org/congress/v1/members/A000371.j...</a>	Pete	Aguilar	1979-06-19	M
5	A000372	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000372.j...">https://api.propublica.org/congress/v1/members/A000372.j...</a>	Rick	Allen	1951-11-07	M
6	A000367	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000367.j...">https://api.propublica.org/congress/v1/members/A000367.j...</a>	Justin	Amash	1980-04-18	M
7	A000369	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000369.j...">https://api.propublica.org/congress/v1/members/A000369.j...</a>	Mark	Amodei	1958-06-12	M
8	A000375	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000375.j...">https://api.propublica.org/congress/v1/members/A000375.j...</a>	Jodey	Arrington	1972-03-09	M
9	B001291	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001291.j...">https://api.propublica.org/congress/v1/members/B001291.j...</a>	Brian	Babin	1948-03-23	M
10	B001298	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001298.j...">https://api.propublica.org/congress/v1/members/B001298.j...</a>	Don	Bacon	1963-08-16	M
11	B001306	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001306.j...">https://api.propublica.org/congress/v1/members/B001306.j...</a>	Troy	Balderson	1962-01-16	M
12	B001299	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001299.j...">https://api.propublica.org/congress/v1/members/B001299.j...</a>	Jim	Banks	1979-07-16	M
13	B001269	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001269.j...">https://api.propublica.org/congress/v1/members/B001269.j...</a>	Lou	Barietta	1956-01-28	M
14	B001282	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001282.j...">https://api.propublica.org/congress/v1/members/B001282.j...</a>	Andy	Barr	1973-07-24	M
15	B001300	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001300.j...">https://api.propublica.org/congress/v1/members/B001300.j...</a>	Nanette	Barragán	1976-09-15	F
16	B000213	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B000213.j...">https://api.propublica.org/congress/v1/members/B000213.j...</a>	Joe	Barton	1949-09-15	M
17	B001270	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001270.j...">https://api.propublica.org/congress/v1/members/B001270.j...</a>	Karen	Bass	1953-10-03	F
18	B001281	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001281.j...">https://api.propublica.org/congress/v1/members/B001281.j...</a>	Joyce	Beatty	1950-03-12	F
19	B000287	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B000287.j...">https://api.propublica.org/congress/v1/members/B000287.j...</a>	Xavier	Becerra	1958-01-26	M
20	B001287	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001287.j...">https://api.propublica.org/congress/v1/members/B001287.j...</a>	Ami	Bera	1965-03-02	M

## 01

## dplyr

## Practice

	id	title	short_title	api_uri	first_name	last_name	date_of_birth	gender
1	A000374	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000374.j...">https://api.propublica.org/congress/v1/members/A000374.j...</a>	Ralph	Abraham	1954-09-16	M
2	A000370	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000370.j...">https://api.propublica.org/congress/v1/members/A000370.j...</a>	Alma	Adams	1946-05-27	F
3	A000055	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000055.j...">https://api.propublica.org/congress/v1/members/A000055.j...</a>	Robert	Aderholt	1965-07-22	M
4	A000371	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000371.j...">https://api.propublica.org/congress/v1/members/A000371.j...</a>	Pete	Aguilar	1979-06-19	M
5	A000372	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000372.j...">https://api.propublica.org/congress/v1/members/A000372.j...</a>	Rick	Allen	1951-11-07	M
6	A000367	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000367.j...">https://api.propublica.org/congress/v1/members/A000367.j...</a>	Justin	Amash	1980-04-18	M
7	A000369	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000369.j...">https://api.propublica.org/congress/v1/members/A000369.j...</a>	Mark	Amodei	1958-06-12	M
8	A000375	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000375.j...">https://api.propublica.org/congress/v1/members/A000375.j...</a>	Jodey	Arrington	1972-03-09	M
9	B001291	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001291.j...">https://api.propublica.org/congress/v1/members/B001291.j...</a>	Brian	Babin	1948-03-23	M
10	B001298	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001298.j...">https://api.propublica.org/congress/v1/members/B001298.j...</a>	Don	Bacon	1963-08-16	M
11	B001306	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001306.j...">https://api.propublica.org/congress/v1/members/B001306.j...</a>	Troy	Balderson	1962-01-16	M
12	B001299	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001299.j...">https://api.propublica.org/congress/v1/members/B001299.j...</a>	Jim	Banks	1979-07-16	M
13	B001269	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001269.j...">https://api.propublica.org/congress/v1/members/B001269.j...</a>	Lou	Barletta	1956-01-28	M
14	B001282	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001282.j...">https://api.propublica.org/congress/v1/members/B001282.j...</a>	Andy	Barr	1973-07-24	M
15	B001300	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001300.j...">https://api.propublica.org/congress/v1/members/B001300.j...</a>	Nanette	Barragán	1976-09-15	F
16	B000213	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B000213.j...">https://api.propublica.org/congress/v1/members/B000213.j...</a>	Joe	Barton	1949-09-15	M
17	B001270	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001270.j...">https://api.propublica.org/congress/v1/members/B001270.j...</a>	Karen	Bass	1953-10-03	F
18	B001281	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001281.j...">https://api.propublica.org/congress/v1/members/B001281.j...</a>	Joyce	Beatty	1950-03-12	F
19	B000287	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B000287.j...">https://api.propublica.org/congress/v1/members/B000287.j...</a>	Xavier	Becerra	1958-01-26	M
20	B001287	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001287.j...">https://api.propublica.org/congress/v1/members/B001287.j...</a>	Ami	Bera	1965-03-02	M

```
house_115_2016 <- house_115[!(house_115$successor == 1)
& !(house_115$non_voting == 1),]
```

```
house_115_2018 <- house_115[!(house_115$vacate == 1)
& !(house_115$non_voting == 1),]
```



## 01

## dplyr

## Practice

	id	title	short_title	api_uri	first_name	last_name	date_of_birth	gender
1	A000374	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000374.j...">https://api.propublica.org/congress/v1/members/A000374.j...</a>	Ralph	Abraham	1954-09-16	M
2	A000370	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000370.j...">https://api.propublica.org/congress/v1/members/A000370.j...</a>	Alma	Adams	1946-05-27	F
3	A000055	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000055.j...">https://api.propublica.org/congress/v1/members/A000055.j...</a>	Robert	Aderholt	1965-07-22	M
4	A000371	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000371.j...">https://api.propublica.org/congress/v1/members/A000371.j...</a>	Pete	Aguilar	1979-06-19	M
5	A000372	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000372.j...">https://api.propublica.org/congress/v1/members/A000372.j...</a>	Rick	Allen	1951-11-07	M
6	A000367	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000367.j...">https://api.propublica.org/congress/v1/members/A000367.j...</a>	Justin	Amash	1980-04-18	M
7	A000369	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000369.j...">https://api.propublica.org/congress/v1/members/A000369.j...</a>	Mark	Amodei	1958-06-12	M
8	A000375	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/A000375.j...">https://api.propublica.org/congress/v1/members/A000375.j...</a>	Jodey	Arrington	1972-03-09	M
9	B001291	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001291.j...">https://api.propublica.org/congress/v1/members/B001291.j...</a>	Brian	Babin	1948-03-23	M
10	B001298	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001298.j...">https://api.propublica.org/congress/v1/members/B001298.j...</a>	Don	Bacon	1963-08-16	M
11	B001306	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001306.j...">https://api.propublica.org/congress/v1/members/B001306.j...</a>	Troy	Balderson	1962-01-16	M
12	B001299	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001299.j...">https://api.propublica.org/congress/v1/members/B001299.j...</a>	Jim	Banks	1979-07-16	M
13	B001269	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001269.j...">https://api.propublica.org/congress/v1/members/B001269.j...</a>	Lou	Barletta	1956-01-28	M
14	B001282	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001282.j...">https://api.propublica.org/congress/v1/members/B001282.j...</a>	Andy	Barr	1973-07-24	M
15	B001300	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001300.j...">https://api.propublica.org/congress/v1/members/B001300.j...</a>	Nanette	Barragán	1976-09-15	F
16	B000213	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B000213.j...">https://api.propublica.org/congress/v1/members/B000213.j...</a>	Joe	Barton	1949-09-15	M
17	B001270	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001270.j...">https://api.propublica.org/congress/v1/members/B001270.j...</a>	Karen	Bass	1953-10-03	F
18	B001281	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001281.j...">https://api.propublica.org/congress/v1/members/B001281.j...</a>	Joyce	Beatty	1950-03-12	F
19	B000287	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B000287.j...">https://api.propublica.org/congress/v1/members/B000287.j...</a>	Xavier	Becerra	1958-01-26	M
20	B001287	Representative	Rep.	<a href="https://api.propublica.org/congress/v1/members/B001287.j...">https://api.propublica.org/congress/v1/members/B001287.j...</a>	Ami	Bera	1965-03-02	M

```
house_115_2016 <- house_115 %>%
  filter(successor != 1) %>%
  filter(non_voting != 1)
```

```
house_115_2018 <- house_115 %>%
  filter(vacate != 1) %>%
  filter(non_voting != 1)
```

Using dplyr's `filter()` to create table objects  
`house_116_2019`, `house_116_2020`,  
`house_117_2021`, `house_117_2022`, and copy your  
codes in Moodle (Practice 1)

# Table Merge

In big data era, merging data is a common and important task:

1. Tables come from different data sources
2. Merging them into a new table to do analysis

```
set.seed(123)
```

```
df1 <- data.frame(id = 1:10, math = rpois(10,50))
```

```
set.seed(123)
```

```
df2 <- data.frame(id = 10:1, art = rpois(10,70))
```

**Don't use cbind!!!!**

**cbind works only when two tables' observations are arranged in same order.**

```
set.seed(123)
```

```
df1 <- data.frame(id = 1:10, math = rpois(10,50))
```

```
set.seed(123)
```

```
df2 <- data.frame(id = 10:1, art = rpois(10,70))
```

```
df <- cbind(df1, df2)
```

```
df <- merge(df1, df2, by = "id")
```

**by:** combine columns that have matching values in a column called "id"

## Different size of observations

```
set.seed(123)
```

```
df3 <- data.frame(id = 1:9, reading = rpois(9,65))
```

```
df_a <- merge(df, df3, by = "id")
```

```
df_b <- merge(df, df3, by = "id", all = TRUE)
```

**left\_join, right\_join, full\_join, inner\_join**

**Basic R's merge:**

```
df <- merge(df1, df2, by = "id")
```

**dplyr's merge:**

```
df <- df1 %>%  
  left_join(df2, by = "id")
```

```
df <- df %>%  
  full_join(df3, by = "id")
```



The names of by column are different

```
set.seed(123)
```

```
df4 <- data.frame(sid = 1:11, writing =  
rpois(11, 75))
```

```
df_e <- df %>%  
  left_join(df4, by = "id")
```

```
df_f <- df %>%  
  full_join(df4, by = "id")
```

## Merge tables by two columns

df5

	id	year	math
1	1	2020	65
2	2	2020	79
3	3	2020	55
4	4	2020	71
5	5	2020	84
6	6	2020	73
7	7	2020	59
8	8	2020	55
9	9	2020	80
10	10	2020	73
11	1	2021	73
12	2	2021	70
13	3	2021	65
14	4	2021	80

df6

	id	year	reading
1	1	2020	72
2	1	2021	69
3	2	2020	69
4	2	2021	81
5	3	2020	68
6	3	2021	82
7	4	2020	57
8	4	2021	61
9	5	2020	71
10	5	2021	73
11	6	2020	65
12	6	2021	87
13	7	2020	58
14	7	2021	72

### Merge tables by two columns

```
df_new <- df5 %>%  
  left_join(df6, by = c("id", "year"))
```

# Assignment

Using ntc\_ref\_2018.xlsx and ntc\_ref\_2021.xlsx to analyze the results of New Taipei City's 2018 and 2021 pro-nuclear power referendums (Case no. 16 and 17) at li level.