

Show
me the
data!

Week03: Dataframe and Import Data

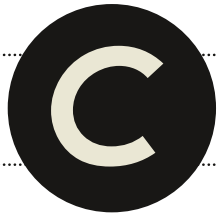
Big Data & Social Analysis R

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 Dataframe
- 2 Import Data
- 3 Basic Cleaning of Dataframe
- 4 Assignment

Dataframe

Dataframes/Tables in R just like Excel's spreadsheet consisted by rows and columns.

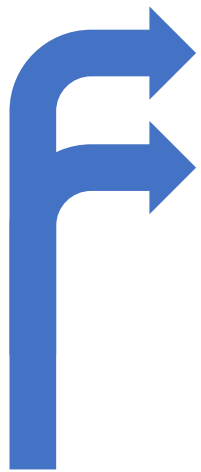
01

Dataframe

columns : observations' variables



	last_name	first_name	birthday	gender	type	state	reelection	TC2018	district	party
1	Young	Don	1933/6/9	M	rep	AK	1	0	0	Republican
2	Byrne	Bradley	1955/2/16	M	rep	AL	1	0	1	Republican
3	Roby	Martha	1976/7/27	F	rep	AL	1	0	2	Republican
4	Rogers	Mike	1958/7/16	M	rep	AL	1	1	3	Republican
5	Aderholt	Robert	1965/7/22	M	rep	AL	1	0	4	Republican
6	Brooks	Mo	1954/4/29	M	rep	AL	1	1	5	Republican
7	Palmer	Gary	1954/5/14	M	rep	AL	1	0	6	Republican
8	Sewell	Terri	1965/1/1	F	rep	AL	1	0	7	Democrat
9	Crawford	Eric	1966/1/22	M	rep	AR	1	1	1	Republican
10	Hill	French	1956/12/5	M	rep	AR	1	0	2	Republican
11	Womack	Steve	1957/2/18	M	rep	AR	1	0	3	Republican
12	Westerman	Bruce	1967/11/18	M	rep	AR	1	0	4	Republican
13	Amata	Aumua	1947/12/29	F	rep	AS	1	0	0	Republican
14	O?alleran	Tom	1946/1/24	M	rep	AZ	1	0	1	Democrat
15	McSally	Martha	1966/3/22	F	rep	AZ	1	0	2	Republican
16	Grijalva	Raul	1948/2/19	M	rep	AZ	1	0	3	Democrat
17	Gosar	Paul	1958/11/22	M	rep	AZ	1	0	4	Republican
18	Biggs	Andy	1958/11/7	M	rep	AZ	1	1	5	Republican
19	Schweikert	David	1962/3/3	M	rep	AZ	1	0	6	Republican
20	Collins	Robert	1970/11/20	M	rep	AZ	1	0	7	Democrat



rows : observations

How to create a table object in R?

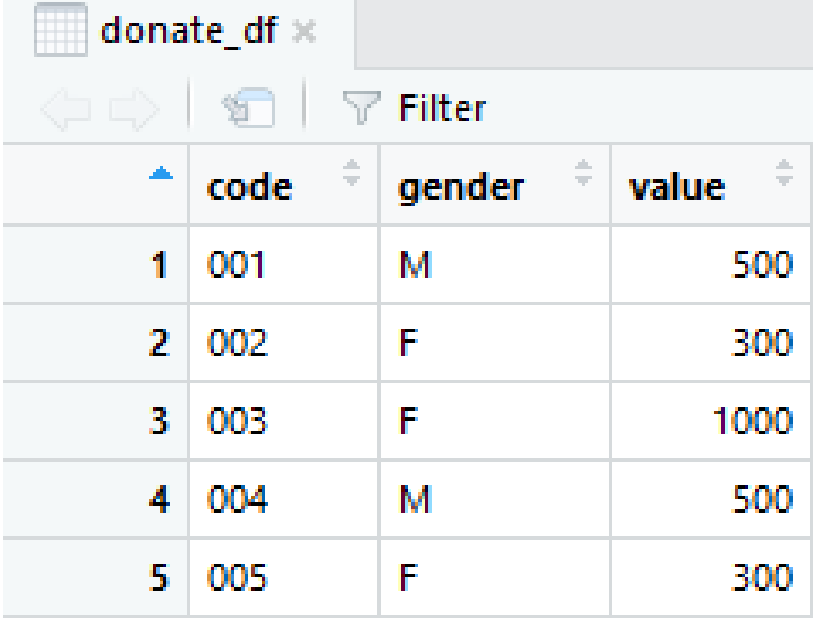
1. Created by `data.frame()` function
2. Imported from an existed file

Using `data.frame()` :

```
donate_df <- data.frame(code = c("001", "002", "003", "004", "005"),  
                        gender = c("M", "F", "F", "M", "F"),  
                        value = c(500, 300, 1000, 500, 300))
```

01

Dataframe



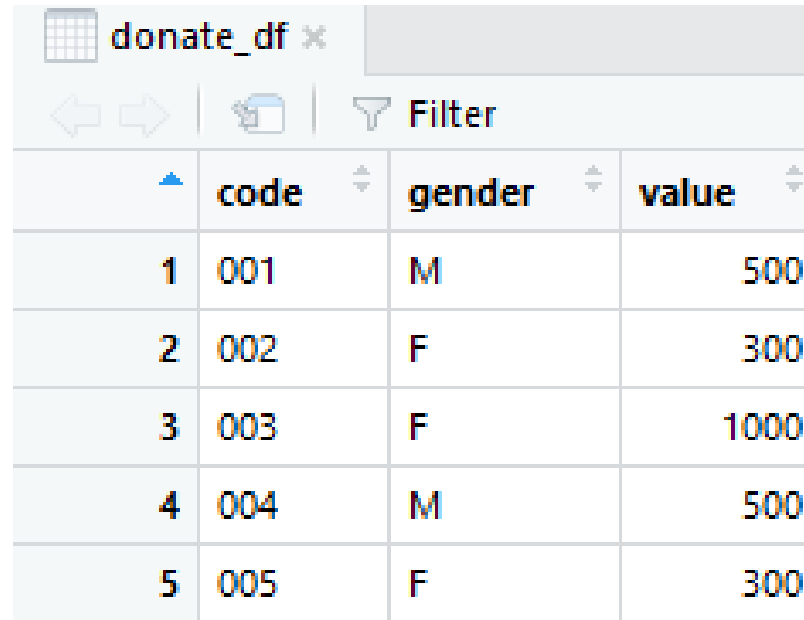
The screenshot shows a Jupyter Notebook interface with a dataframe named 'donate_df'. The dataframe has 5 rows and 4 columns: index, code, gender, and value. The data is as follows:

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

Take all data in a column:
Using **\$** : `donate_df$value`

01

Dataframe



The screenshot shows a data frame viewer window titled 'donate_df x'. It contains a table with 5 rows and 4 columns. The columns are labeled 'code', 'gender', and 'value'. The first column is an index column with values 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

You can create a vector object and assign values from df

```
money <- donate_df$value
```

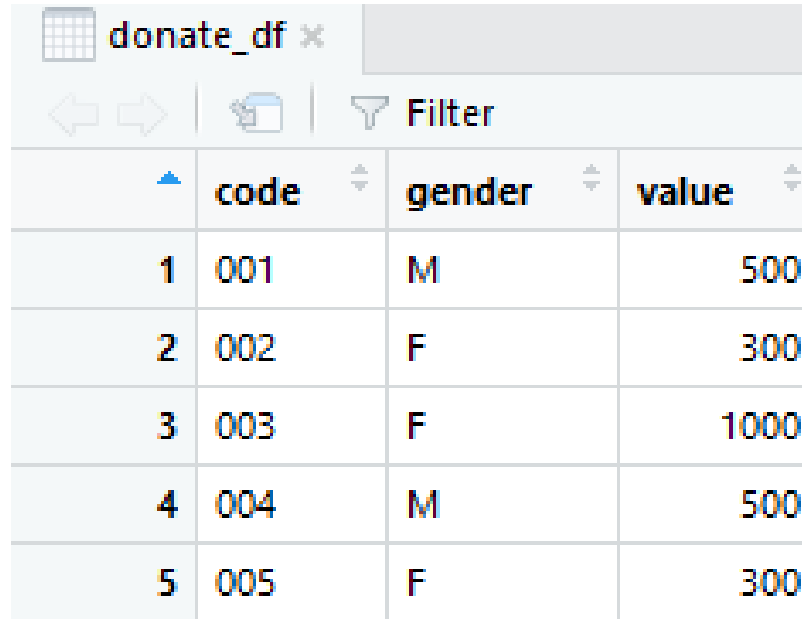
```
money <- c(500, 300, 1000, 500, 300)
```

You can use reverse way to create df

```
donate_df <- data.frame(code = c("001", "002", "003", "004", "005"),  
                        gender = c("M", "F", "F", "M", "F"),  
                        value = money)
```

01

Dataframe



donate_df x

← → | 📄 | 🏠 Filter

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

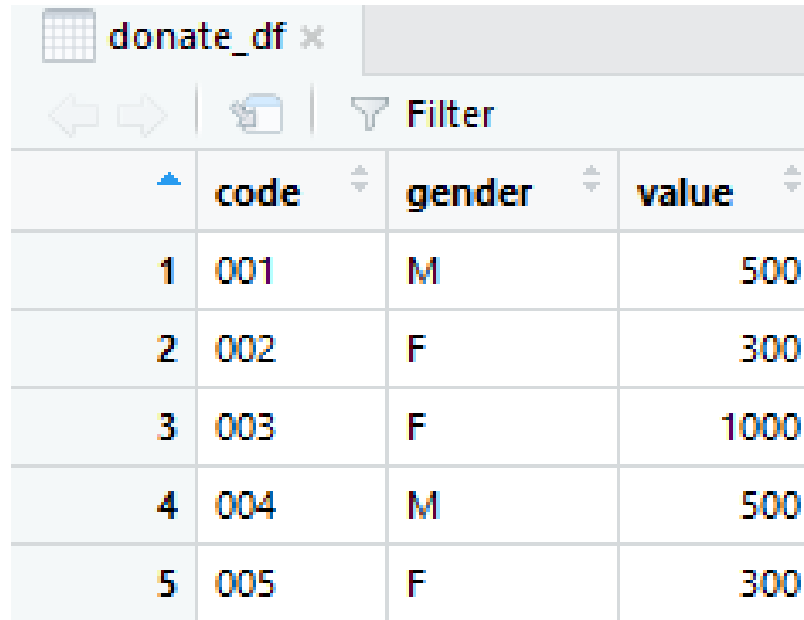
Calculate the sum and average of the donation

```
money <- donate_df$value  
sum(money)  
mean(money)
```

```
sum(donate_df$value)  
mean(donate_df$value)
```

01

Dataframe



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

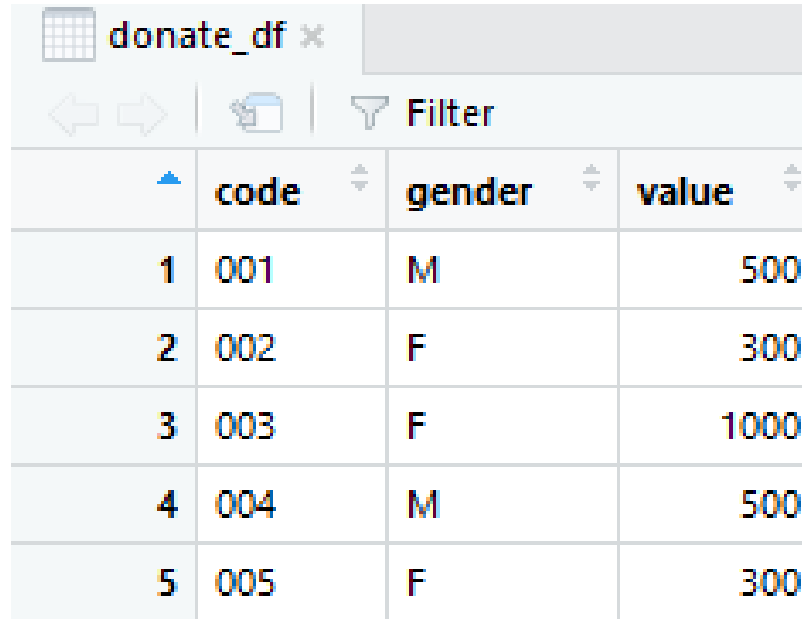
Compare male and female's donation

```
donate_df$value[donate_df$gender == "M"]
```

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

01

Dataframe



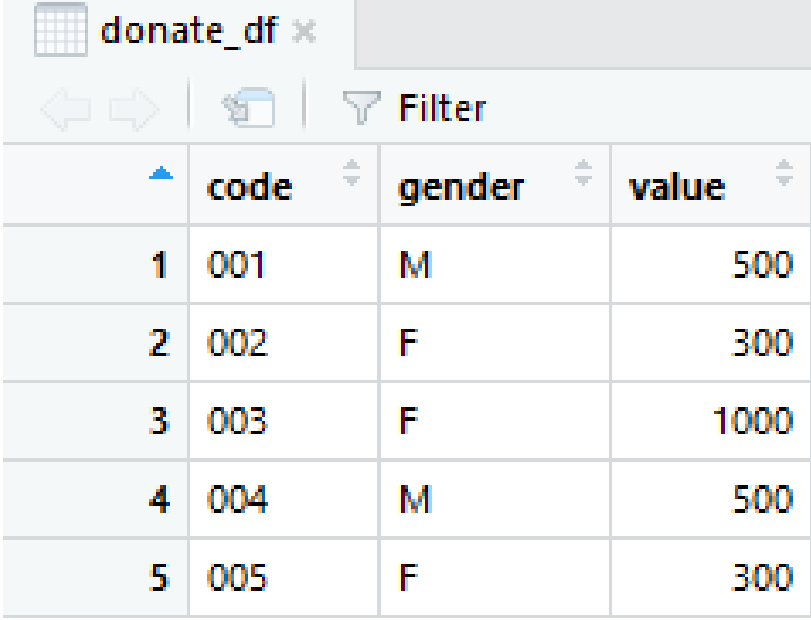
	code	gender	value
1	001	M	500
2	002	F	300
3	003	F	1000
4	004	M	500
5	005	F	300

Compare male and female's donation

Calculate sum and mean of female's donation in Moodle (Practice 1)

01

Dataframe



	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"])
```

Import Data

How to create a table object in R?

1. Created by `data.frame()` function
2. Imported from an existed file

Usually, we can download datasets from external sources:

Your friends

Your teammates

Websites

Standard Operating Procedure (SOP) for employing a data project before you coding:

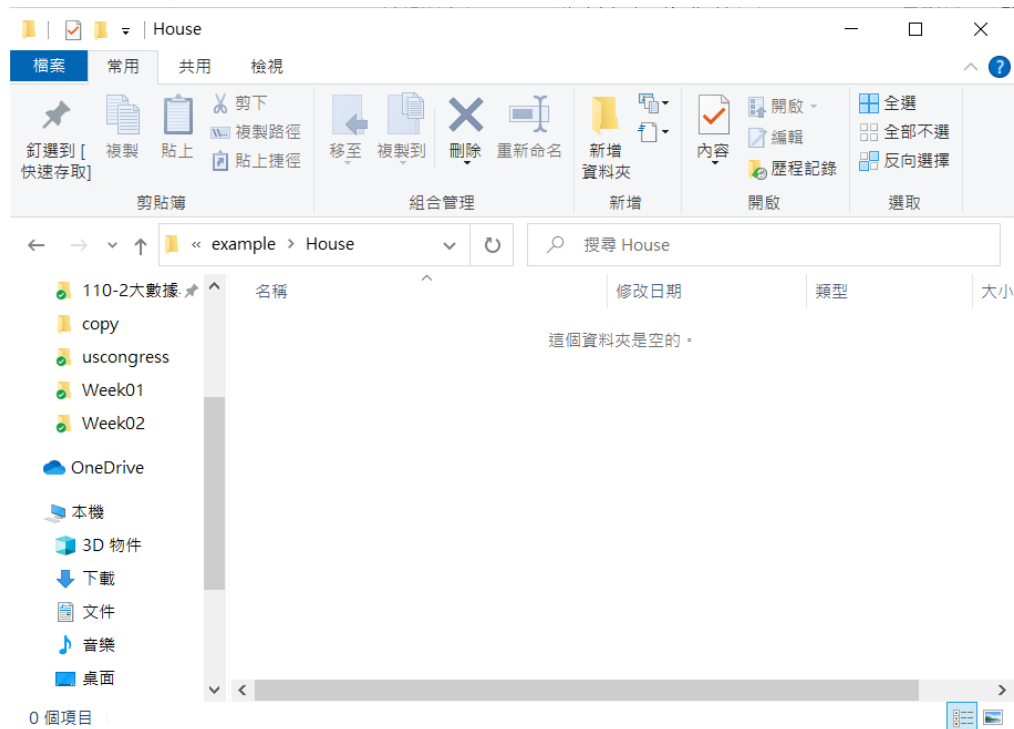
1. Create a folder for this data project
2. Create a R script and save it into the folder
3. Copy dataset files into this folder
4. Set the working directory in R
5. Load packages the project needs
6. Read the datasets

02

Import Data

We want to do US House data project

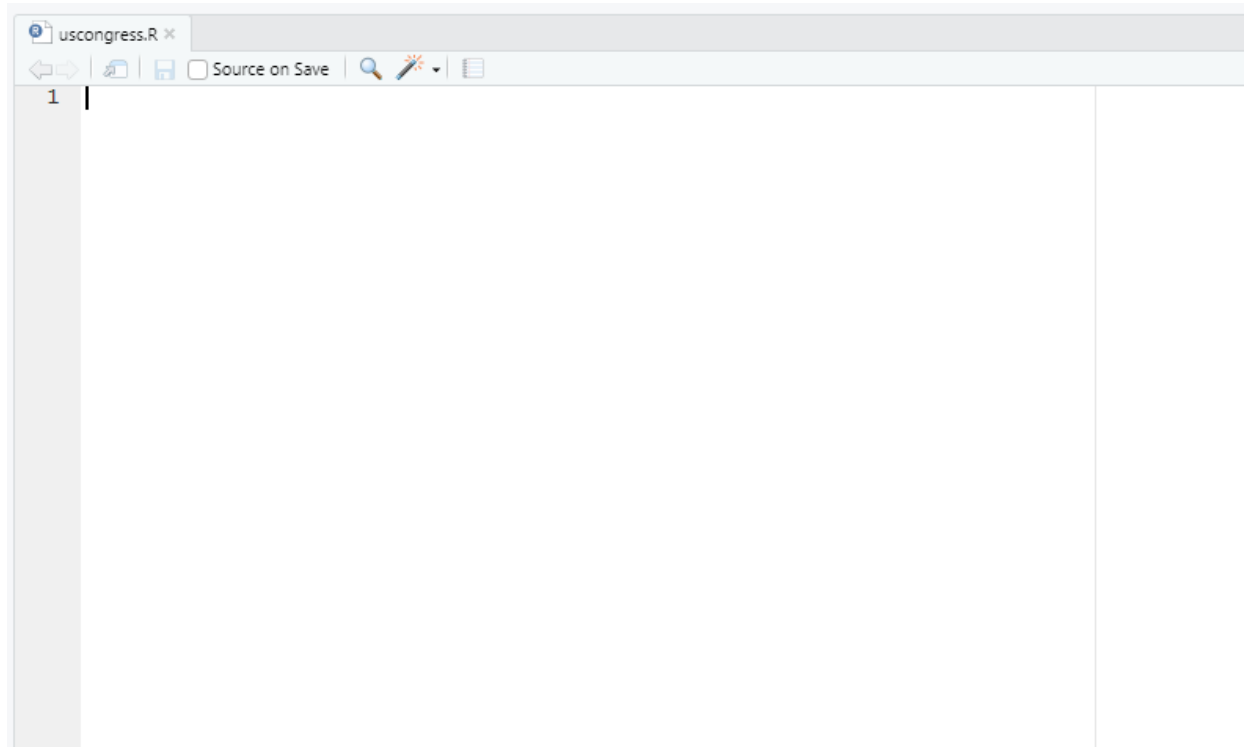
Create a folder called House



02

Import Data

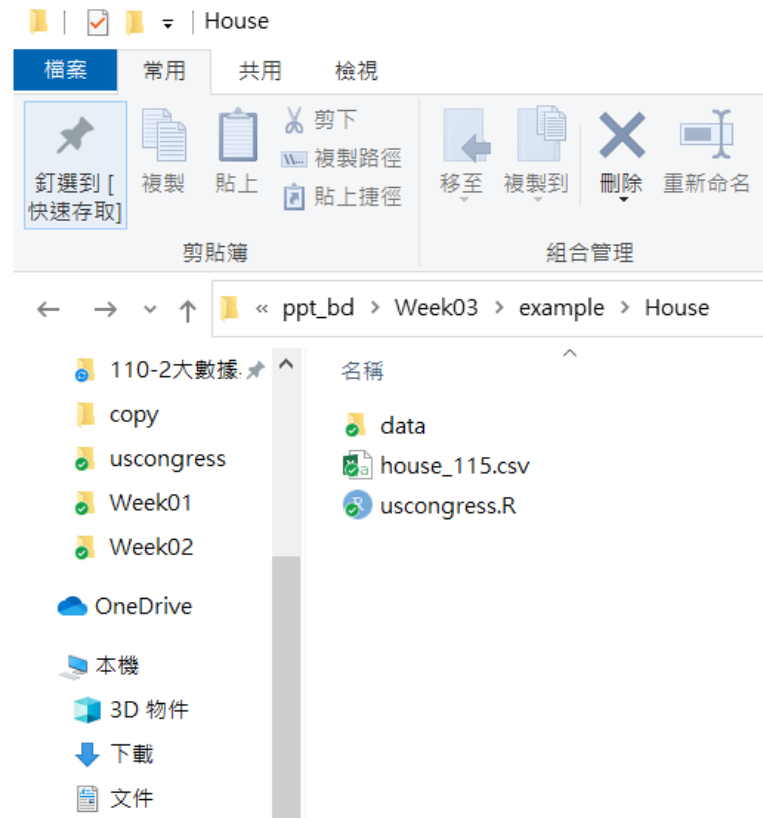
Create a R script file called
uscongress.R and save it into House



02

Import Data

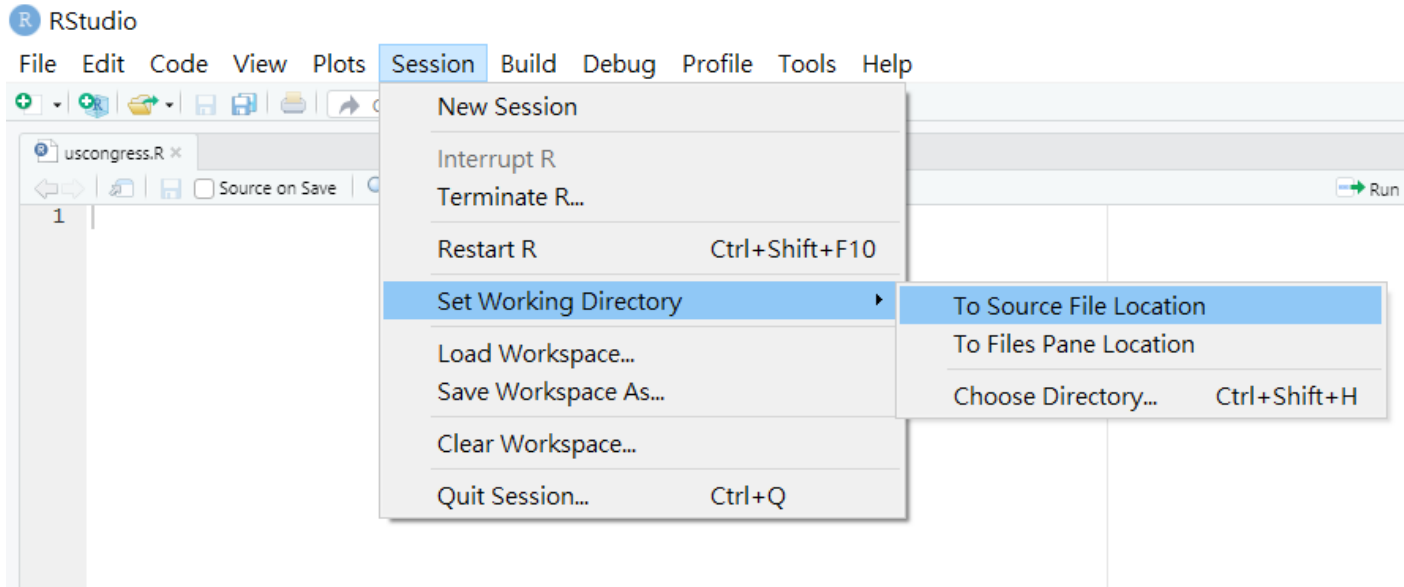
Create a folder called House



02

Import Data

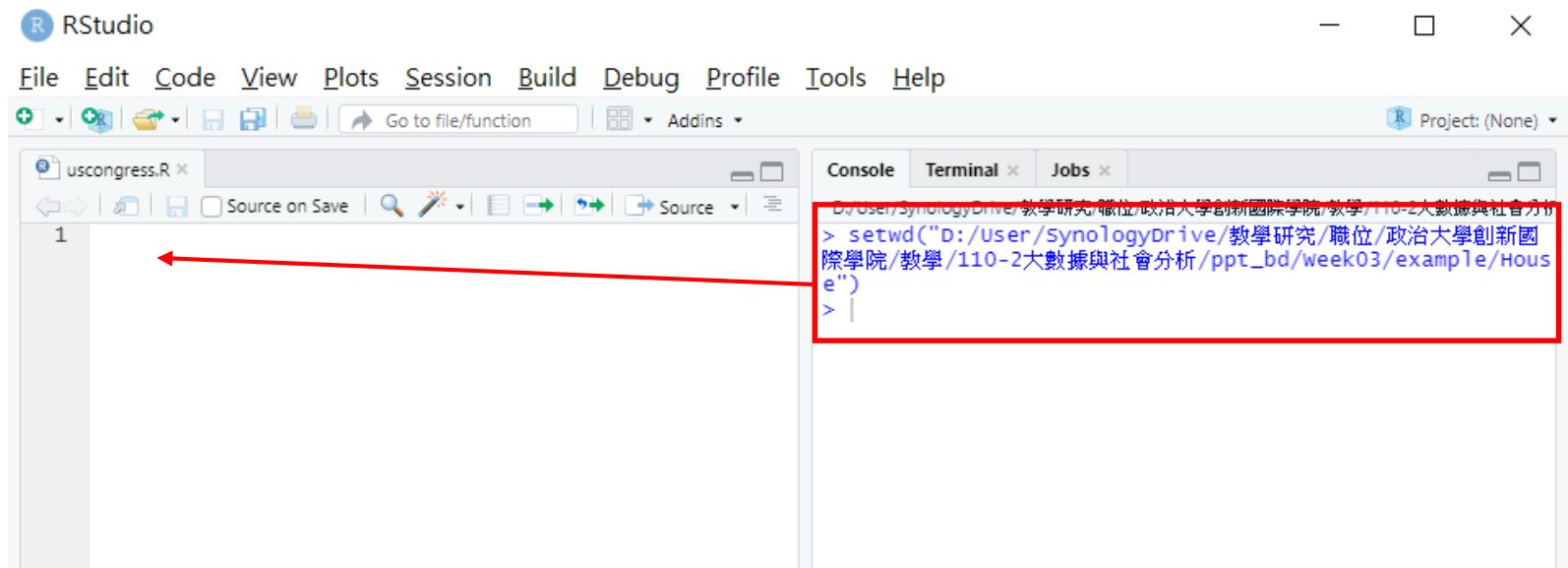
Set the working directory in R



02

Import Data

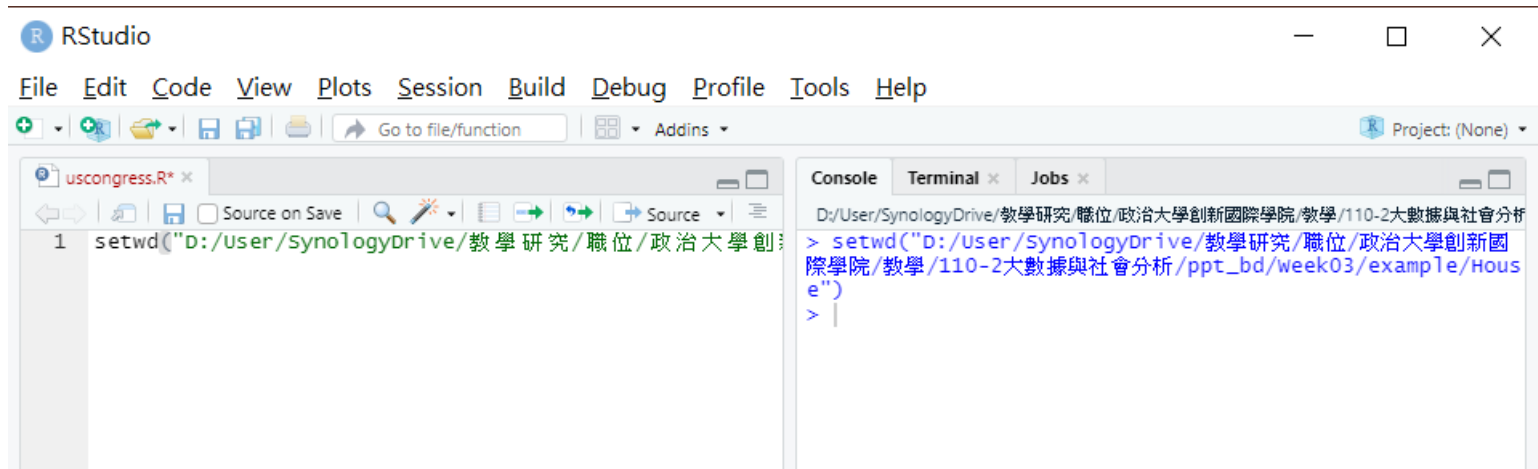
Select the codes in Console, and copy then paste it into uscongress.R



02

Import Data

Select the codes in Console, and copy then paste it into uscongress.R



The screenshot shows the RStudio interface. The script editor on the left contains a single line of code: `1 setwd("D:/User/SynologyDrive/教學研究/職位/政治大學創`. The console on the right shows the full path: `D:/User/SynologyDrive/教學研究/職位/政治大學創新國際學院/教學/110-2大數據與社會分析/ppt_bd/week03/example/House`. The code in the console is highlighted in blue, indicating it has been selected for copying.

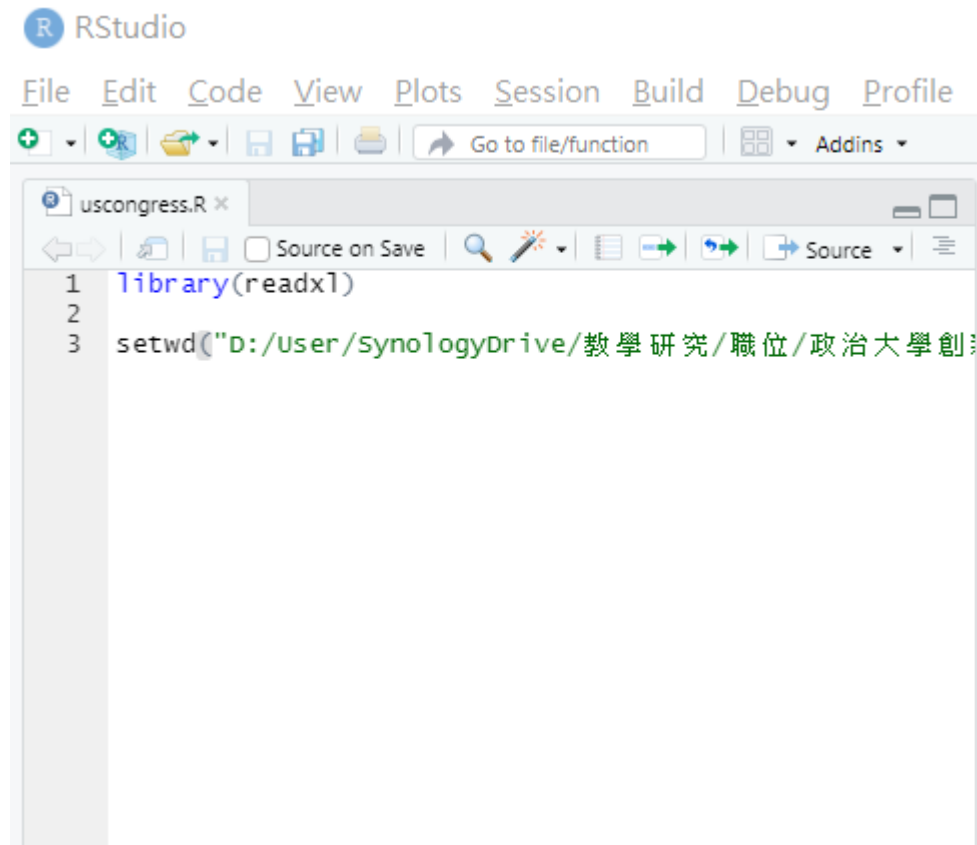
```
1 setwd("D:/User/SynologyDrive/教學研究/職位/政治大學創
```

```
> setwd("D:/User/SynologyDrive/教學研究/職位/政治大學創新國際學院/教學/110-2大數據與社會分析/ppt_bd/week03/example/House")  
> |
```


02

Import Data

Load readxl packages because we need to read xlsx files in this project



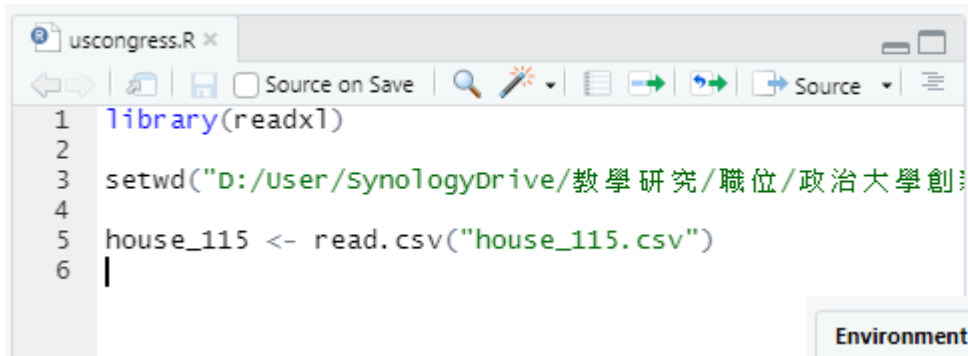
The screenshot shows the RStudio application window. The menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, and Profile. The toolbar contains icons for adding files, saving, and navigating. The source editor window, titled 'uscongress.R', displays the following R code:

```
1 library(readxl)
2
3 setwd("D:/User/SynologyDrive/教學研究/職位/政治大學創
```

02

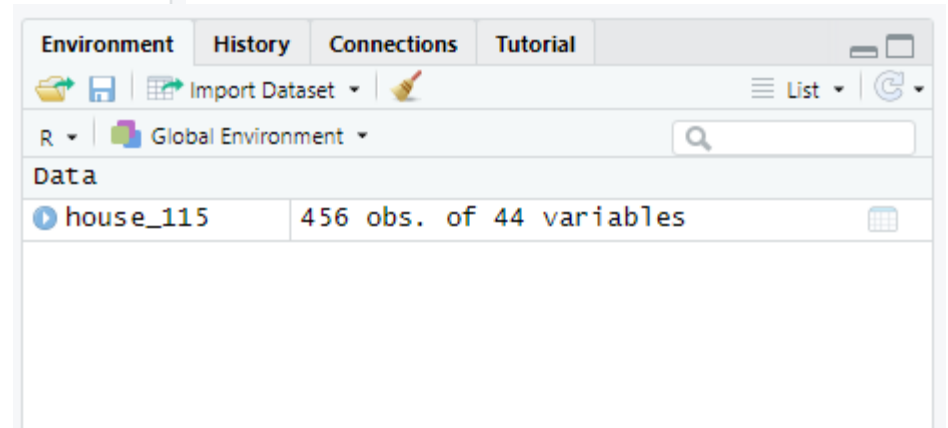
Import Data

```
house_115 <- read.csv("house_115.csv")
```



```
1 library(readxl)
2
3 setwd("D:/User/SynologyDrive/教學研究/職位/政治大學創")
4
5 house_115 <- read.csv("house_115.csv")
6 |
```

You will see the object **house_115** in the Environment window



02

Import Data

Click **house_115**, you will see the content of house_115

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

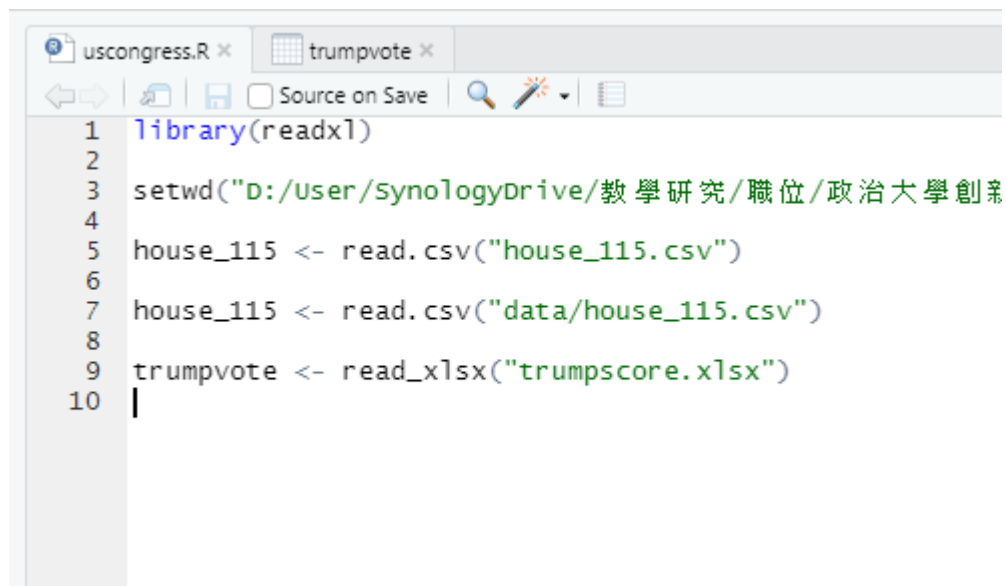
02

Import Data

Read other format files:

Copy files into folder

```
trumpvote <- read_xlsx("trumpscore.xlsx")
```



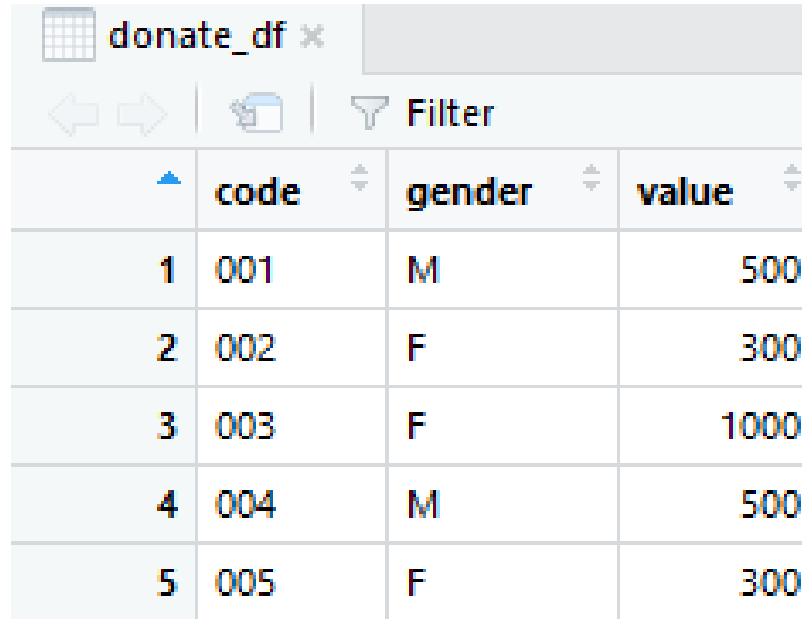
```
uscongress.R x trumpvote x
Source on Save
1 library(readxl)
2
3 setwd("D:/User/SynologyDrive/教學研究/職位/政治大學創業
4
5 house_115 <- read_csv("house_115.csv")
6
7 house_115 <- read_csv("data/house_115.csv")
8
9 trumpvote <- read_xlsx("trumpscore.xlsx")
10 |
```

Basic Cleaning of Dataframe

03

Basic Cleaning of Dataframe

Basic Skills



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

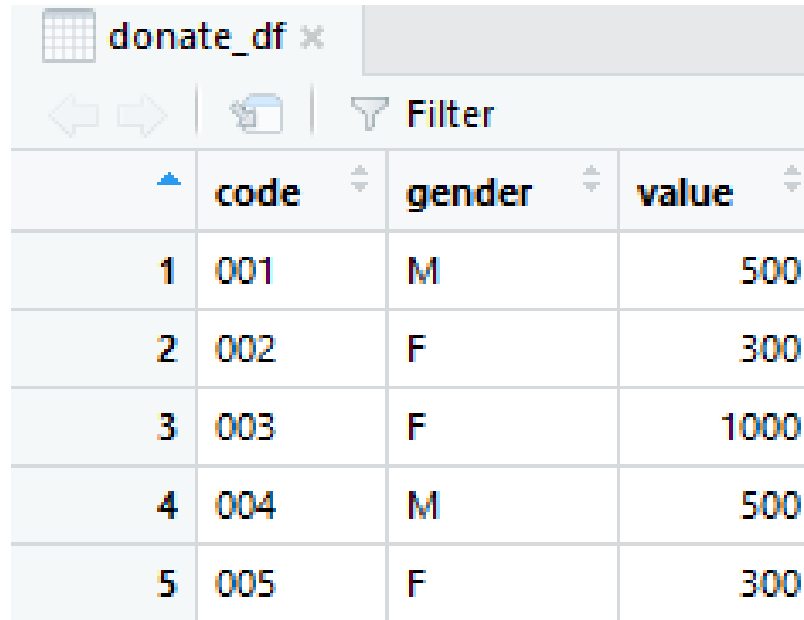
Base skills: Create new variables and slice the dataframe

1. Create a variable called `usd`: Exchange value (TW) to US dollar
2. Create a variable called `sex`: Female is 1 and male is 0
3. Create a variable called `donate`: value ≥ 500 is 2, and value small than 500 is 1
4. Create a table called `male_donate`: Only male observations

03

Basic Cleaning of Dataframe

Basic Skills



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

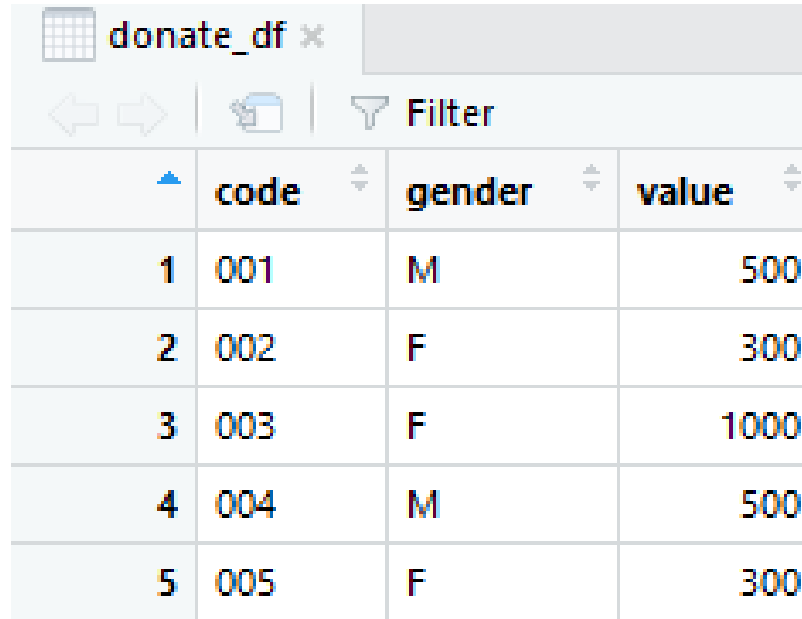
1. Create a variable called `usd`: Exchange value (TW) to US dollar

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

03

Basic Cleaning of Dataframe

Basic Skills



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

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

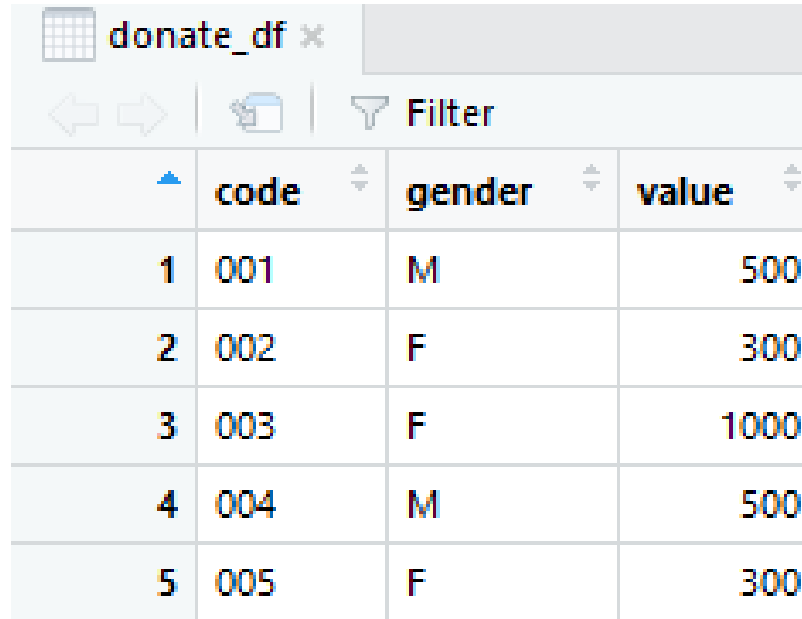
```
donate_df$sex <- 0
```

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


03

Basic Cleaning of Dataframe

Basic Skills



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

3. Create a variable called donate: value ≥ 500 is 2, and value small than 500 is 1

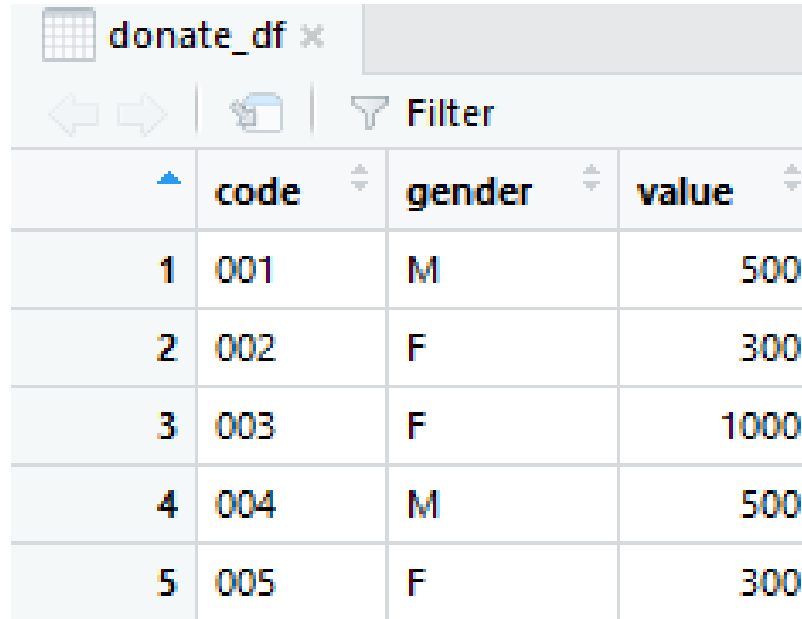
```
donate_df$donate <- 1
```

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

03

Basic Cleaning of Dataframe

Basic Skills



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

4. Create a table called male_donate: Only male observations

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

Only female observations and 2-3 columns

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

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

03

Basic Cleaning of Dataframe

Basic Skills

	id	title	short_title	api_uri	first_name	last_name	date_of_birth	gender
1	A000374	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000374.j...	Ralph	Abraham	1954-09-16	M
2	A000370	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000370.j...	Alma	Adams	1946-05-27	F
3	A000055	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000055.j...	Robert	Aderholt	1965-07-22	M
4	A000371	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000371.j...	Pete	Aguilar	1979-06-19	M
5	A000372	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000372.j...	Rick	Allen	1951-11-07	M
6	A000367	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000367.j...	Justin	Amash	1980-04-18	M
7	A000369	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000369.j...	Mark	Amodei	1958-06-12	M
8	A000375	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000375.j...	Jodey	Arrington	1972-03-09	M
9	B001291	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001291.j...	Brian	Babin	1948-03-23	M
10	B001298	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001298.j...	Don	Bacon	1963-08-16	M
11	B001306	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001306.j...	Troy	Balderson	1962-01-16	M
12	B001299	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001299.j...	Jim	Banks	1979-07-16	M
13	B001269	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001269.j...	Lou	Barletta	1956-01-28	M

Check the observations and variables:

```
nrow(house_115)
```

```
colnames(house_115)
```

03

Basic Cleaning of Dataframe

Basic Skills

	id	title	short_title	api_uri	first_name	last_name	date_of_birth	gender
1	A000374	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000374.j...	Ralph	Abraham	1954-09-16	M
2	A000370	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000370.j...	Alma	Adams	1946-05-27	F
3	A000055	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000055.j...	Robert	Aderholt	1965-07-22	M
4	A000371	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000371.j...	Pete	Aguilar	1979-06-19	M
5	A000372	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000372.j...	Rick	Allen	1951-11-07	M
6	A000367	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000367.j...	Justin	Amash	1980-04-18	M
7	A000369	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000369.j...	Mark	Amodei	1958-06-12	M
8	A000375	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000375.j...	Jodey	Arrington	1972-03-09	M
9	B001291	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001291.j...	Brian	Babin	1948-03-23	M
10	B001298	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001298.j...	Don	Bacon	1963-08-16	M
11	B001306	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001306.j...	Troy	Balderson	1962-01-16	M
12	B001299	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001299.j...	Jim	Banks	1979-07-16	M
13	B001269	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001269.j...	Lou	Barletta	1956-01-28	M

Don't want to key in this in the future?

house_115\$date_of_birth

You can change the columns' names:

```
colnames(house_115)[7] <- c("birth")
```

```
house_115$birth
```

03

Basic Cleaning of Dataframe

Basic Skills

	id	title	short_title	api_uri	first_name	last_name	date_of_birth	gender
1	A000374	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000374.j...	Ralph	Abraham	1954-09-16	M
2	A000370	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000370.j...	Alma	Adams	1946-05-27	F
3	A000055	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000055.j...	Robert	Aderholt	1965-07-22	M
4	A000371	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000371.j...	Pete	Aguilar	1979-06-19	M
5	A000372	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000372.j...	Rick	Allen	1951-11-07	M
6	A000367	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000367.j...	Justin	Amash	1980-04-18	M
7	A000369	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000369.j...	Mark	Amodei	1958-06-12	M
8	A000375	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000375.j...	Jodey	Arrington	1972-03-09	M
9	B001291	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001291.j...	Brian	Babin	1948-03-23	M
10	B001298	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001298.j...	Don	Bacon	1963-08-16	M
11	B001306	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001306.j...	Troy	Balderson	1962-01-16	M
12	B001299	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001299.j...	Jim	Banks	1979-07-16	M
13	B001269	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001269.j...	Lou	Barletta	1956-01-28	M

Change multiple column names?

```
colnames(house_115)[7, 10:11] <-  
c("birth", "twitter", "facebook")
```

03

Basic Cleaning of Dataframe

Basic Skills

	id	title	short_title	api_uri	first_name	last_name	date_of_birth	gender
1	A000374	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000374.j...	Ralph	Abraham	1954-09-16	M
2	A000370	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000370.j...	Alma	Adams	1946-05-27	F
3	A000055	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000055.j...	Robert	Aderholt	1965-07-22	M
4	A000371	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000371.j...	Pete	Aguilar	1979-06-19	M
5	A000372	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000372.j...	Rick	Allen	1951-11-07	M
6	A000367	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000367.j...	Justin	Amash	1980-04-18	M
7	A000369	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000369.j...	Mark	Amodei	1958-06-12	M
8	A000375	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000375.j...	Jodey	Arrington	1972-03-09	M
9	B001291	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001291.j...	Brian	Babin	1948-03-23	M
10	B001298	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001298.j...	Don	Bacon	1963-08-16	M
11	B001306	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001306.j...	Troy	Balderson	1962-01-16	M
12	B001299	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001299.j...	Jim	Banks	1979-07-16	M
13	B001269	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001269.j...	Lou	Barletta	1956-01-28	M

Select columns you want:

```
house_a <- house_115[, c(1, 5:7)]
```

Select rows and columns you want:

```
house_b <- house_115[100:200, c(1, 5:7)]
```

03

Basic Cleaning of Dataframe

Basic Skills

	id	title	short_title	api_uri	first_name	last_name	date_of_birth	gender
1	A000374	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000374.j...	Ralph	Abraham	1954-09-16	M
2	A000370	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000370.j...	Alma	Adams	1946-05-27	F
3	A000055	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000055.j...	Robert	Aderholt	1965-07-22	M
4	A000371	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000371.j...	Pete	Aguilar	1979-06-19	M
5	A000372	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000372.j...	Rick	Allen	1951-11-07	M
6	A000367	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000367.j...	Justin	Amash	1980-04-18	M
7	A000369	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000369.j...	Mark	Amodei	1958-06-12	M
8	A000375	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000375.j...	Jodey	Arrington	1972-03-09	M
9	B001291	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001291.j...	Brian	Babin	1948-03-23	M
10	B001298	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001298.j...	Don	Bacon	1963-08-16	M
11	B001306	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001306.j...	Troy	Balderson	1962-01-16	M
12	B001299	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001299.j...	Jim	Banks	1979-07-16	M
13	B001269	Representative	Rep.	https://api.propublica.org/congress/v1/members/B001269.j...	Lou	Barletta	1956-01-28	M

Select rows you want:

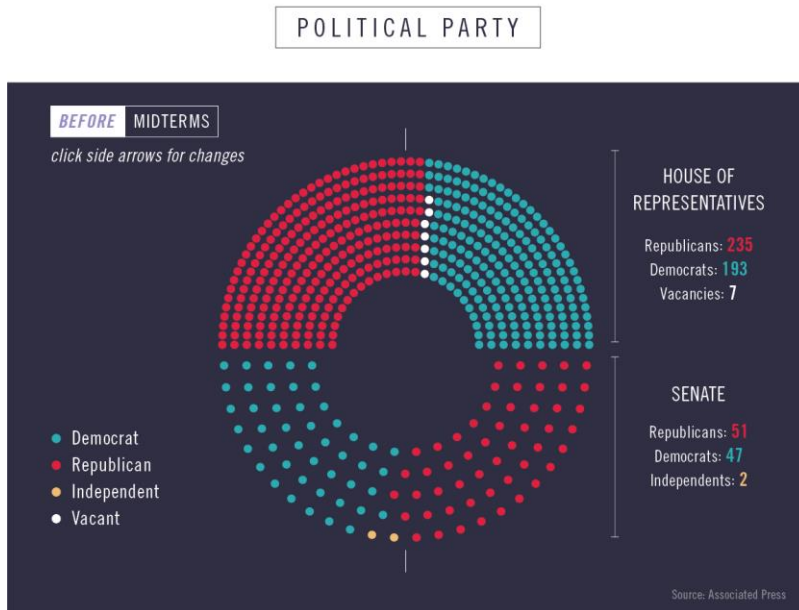
Create df only involving female lawmakers' data:

```
house_f <- house_115[house_115$gender == "F",]
```


03

Basic Cleaning of Dataframe

Example 1

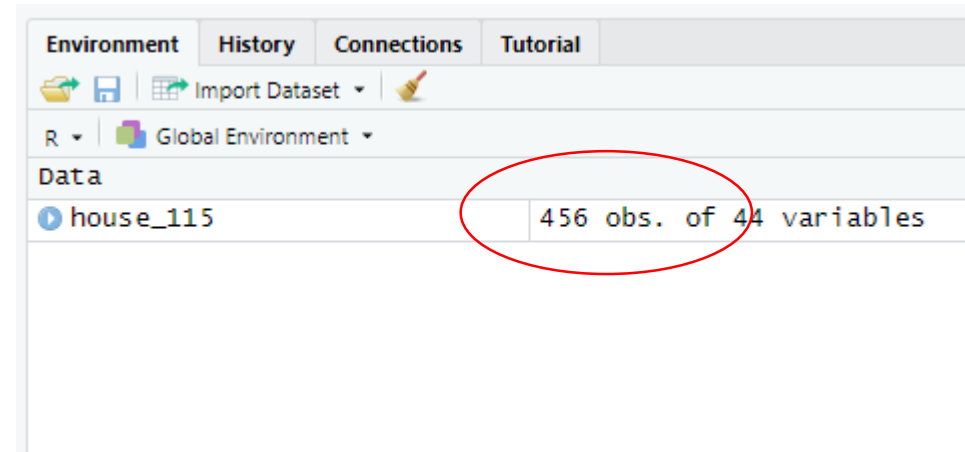


Why our data has 456 obs?????

2018

House: 435

Senate: 100



Please analyze house_115 and review literature to figure out:

1. Why there are two different number of observations?
2. How to produce a table with correct observations?

03

Basic Cleaning of Dataframe

Example 1

House of Representatives [\[edit\]](#)

See also: [List of special elections to the United States House of Representatives](#)

District	Vacated by	Reason for change	Successor
West Virginia 3	Evan Jenkins (R)	Resigned September 30, 2018, to become justice of the Supreme Court of Appeals of West Virginia . ^[76] Seat remained vacant until determined by general election.	Vacant until the next
Utah 3	Jason Chaffetz (R)	Resigned June 30, 2017, for personal reasons. ^[55] A special election was held November 7, 2017. ^[56]	John Curtis (R)
Texas 27	Blake Farenthold (R)	Resigned April 6, 2018. ^[22] A special election was held June 30, 2018. ^[69]	Michael Cloud (R)
South Carolina 5	Mick Mulvaney (R)	Resigned February 16, 2017, to become Director of the Office of Management and Budget . ^[52] A special election was held June 20, 2017. ^[53]	Ralph Norman (R)
Pennsylvania 18	Tim Murphy (R)	Resigned October 21, 2017. ^[57] A special election was held March 13, 2018. ^[58]	Conor Lamb (D)

Because house_115
involves vacating, success,
and non-voting
lawmakers

115th United States Congress

114th ← → 116th



United States Capitol (2017)

January 3, 2017 – January 3, 2019

Members

- 100 senators
- 435 representatives
- 6 non-voting delegates

In 2016, there were 435 lawmakers plus 6 non-voting ones.

There were 18 lawmakers who left the positions from 2016-2018

However, only 15 districts were reelected.

03

Basic Cleaning of Dataframe

Example 1

vacate	successor	non_voting
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	0
0	0	0

We are very lucky. House_115 has that information.

Variable	Coding	Description	No.
vacate	0	Stay in the office	-
	1	Leave the office	18
successor	0	Non-successor members	-
	1	Successor	15
non_voting	0	Formal members	-
	1	Non_voting members	6

03

Basic Cleaning of Dataframe

Example 1

Variable	Coding	Description	No.
vacate	0	Stay in the office	-
	1	Leave the office	18
successor	0	Non-successor members	-
	1	Successor	15
non_voting	0	Formal members	-
	1	Non_voting members	6

A table represents 2016's lawmakers:

All observations – (successor = 1) – (non_voting = 1) = **435**

A table represents lawmakers before the 2018 election:

All observations – (vacate = 1) - (non_voting = 1) = **432**

03

Basic Cleaning of Dataframe

Example 1

Variable	Coding	Description	No.
vacate	0	Stay in the office	-
	1	Leave the office	18
successor	0	Non-successor members	-
	1	Successor	15
non_voting	0	Formal members	-
	1	Non_voting members	6

```
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),]
```

03

Basic Cleaning of Dataframe

Example 2

Using house_115_2016 to calculate gender ratio, and partisan ratio in Moodle (Practice 2).

Create new variables based on your demands

1. names: Combine first name and last name
2. sex: Female is 1 and male is 0
3. party_cate: R is 1, D is 2, and others is 0
4. year: Lawmakers' birth year

Create new variables based on your demands

1. names: Combine first name and last name

```
house_115_2016 $name <-  
  paste(house_115_2016$first_name,  
        house_115_2016$last_name, sep = " ")
```

2. sex: Female is 1 and male is 0

3. party_cate: R is 1, D is 2, and others is 0

4. year: Lawmakers' birth year

Create new variables based on your demands

2. sex: Female is 1 and male is 0

```
unique(house_115_2016$gender)
house_115_2016$sex <- 0
house_115_2016$sex[house_115_2016$gender ==
"F"] <- 1
```

Create new variables based on your demands

3. party_cate: R is 1, D is 2, and others is 0

```
unique(house_115_2016$party)
house_115_2016$party_cate <- 2
house_115_2016$ party_cate
  [house_115_2016$party == "R"] <- 1
```

03

Basic Cleaning of Dataframe

Example 3

Create new variables based on your demands

4. year: Lawmakers' birth year

```
house_115_2016$year <-  
  substr(house_115_2016$birth, 1, 4)
```

Create new variables based on your demands

4. year: Lawmakers' birth year

```
minus <- regexpr("-",  
  house_115_2016$date_of_birth)
```

```
house_115_2016$year <-  
substr(house_115_2016$date_of_birth, 1, minus -  
1)
```

Read 116th US House lawmaker table: `house_116.csv`

There is a column called `name`.

Give me codes to create a new column which shows every lawmakers' last name in Moodle (Practice 3).

Assignment

Read 116th and 117th US House lawmaker table:
house_116.csv and house_117.csv

Clean and create tables of 116th and 117th US House lawmakers.