Show me the data!

Week05: Data Mining and Visualization

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 Data Mining

2 Basic Data Visualization

3 Data Project Skills

Data Mining

Data Mining

Introduction

Now, you have basic R skills to explore undetected information inside data.

1. Create variables
2. Calcuate elements in vectors or variables
3. group_by and summarize
4. Merge tables

5. Establish statistical models

These skills are called “data mining.”

02

Data Mining

Introduction

Now, you have basic R skills to explore undetected information inside data.

1. Create variables
2. Calcuate elements in vectors or variables
3. group_by and summarize
4. Merge tables

~~5. Establish statistical models~~

These skills are called “data mining.”

02

Data Mining

Gender Issues in US Congress

We tried to analyze gender issues from 115th – 117th US Congress last few weeks.

However, if you search Google, discussing gender issues of US Congress is common. Many media, think tanks, and scholoars have already done that.

02 Data Mining

Gender Issues in US Congress

If I still want to analyze gender issues of US Congress, I can’t repeat arguments of previous analyses and make no contributions.

1. **Observe what previous analyses did**
2. **Find something new**

Data Mining

Gender Issues in US Congress

1. What previous analyses did:

   Trend of gender ratio
   Different parties' gender ratio

2. What previous analyses did no have?

   Brainstorm: data you have or it exists in somewhere can do……

02

Data Mining

Gender Issues in US Congress

I have two idea:

1. **Different parties' gender ratio at state level**
2. **Were female lawmakers less likely to support Trump?**

Challenges:

1. **How to calculate them?**
2. **How to interpret the results?**

02 Data Mining

Gender Issues in US Congress

1. Different parties' gender ratio at state level

The number of male and female lawmakers and their ratio.

```
house_115_g <- house_115_2016 %>%
  group_by(gender) %>%
  summarise(sex = n(),
                      ratio = n() / nrow(.))
```

02

Data Mining

Gender Issues in US Congress

1. Different parties’ gender ratio at state level

The number of male and female lawmakers and their ratio by party.

```
house_115_gp <- house_115_2016 %>%
  group_by(gender, party) %>%
  summarise(sex = n(),
            ratio = n() / nrow(.))
```

02

Data Mining

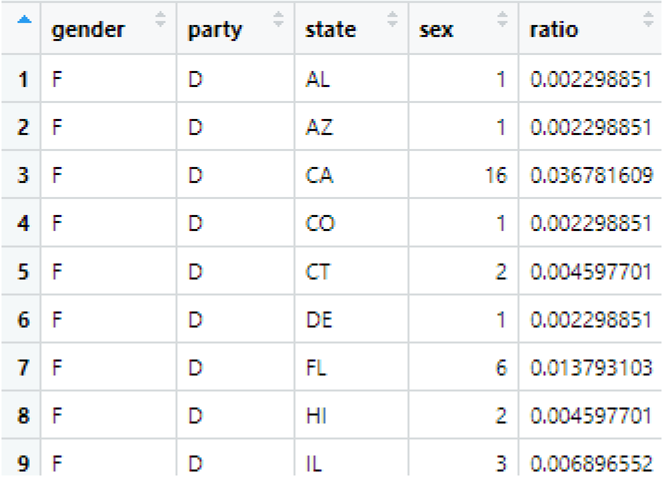
Gender Issues in US Congress

1. **Different parties' gender ratio at state level**

Every state's party gender ratio (Practice 1)

Every state’s party gender ratio (Practice 1)

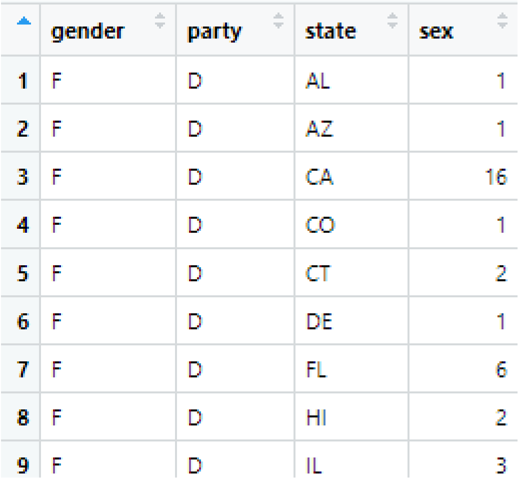
```
house_115_sp <- house_115_2016 %>%
  group_by(gender, party, state) %>%
  summarise(sex = n(),
                      ratio = n() / nrow(.))
```



	gender	party	state	sex	ratio
1	F	D	AL	1	0.002298851
2	F	D	AZ	1	0.002298851
3	F	D	CA	16	0.036781609
4	F	D	CO	1	0.002298851
5	F	D	CT	2	0.004597701
6	F	D	DE	1	0.002298851
7	F	D	FL	6	0.013793103
8	F	D	HI	2	0.004597701
9	F	D	IL	3	0.006896552

Every state's party gender ratio (Practice 1)

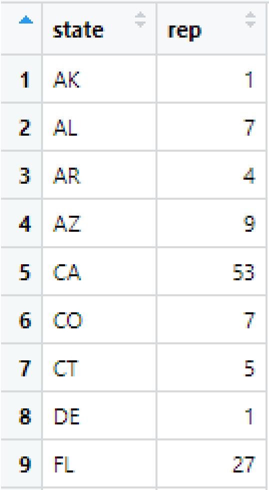
```
house_115_sp <- house_115_2016 %>%
  group_by(gender, party, state) %>%
  summarise(sex = n())
```



	gender	party	state	sex
1	F	D	AL	1
2	F	D	AZ	1
3	F	D	CA	16
4	F	D	CO	1
5	F	D	CT	2
6	F	D	DE	1
7	F	D	FL	6
8	F	D	HI	2
9	F	D	IL	3

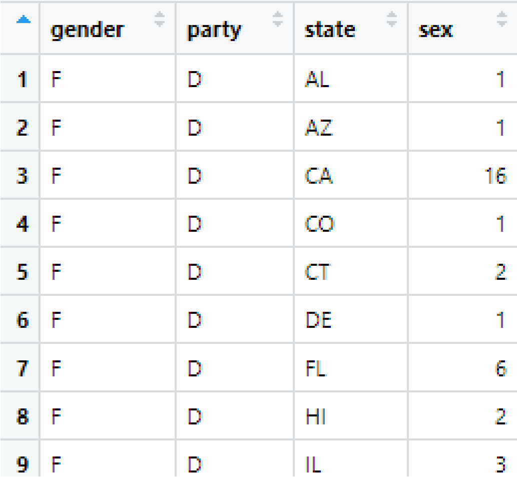
Every state’s party gender ratio (Practice 1)

```
house_115_state <- house_115_2016 %>%
  group_by(state) %>%
  summarise(rep  = n())
```

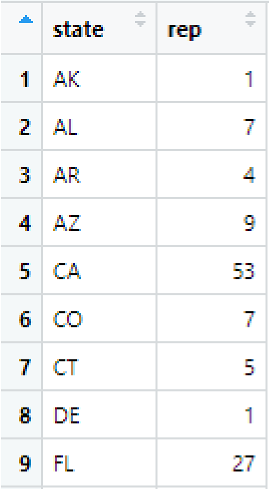


	state	rep
1	AK	1
2	AL	7
3	AR	4
4	AZ	9
5	CA	53
6	CO	7
7	CT	5
8	DE	1
9	FL	27

Every state’s party gender ratio (Practice 1)



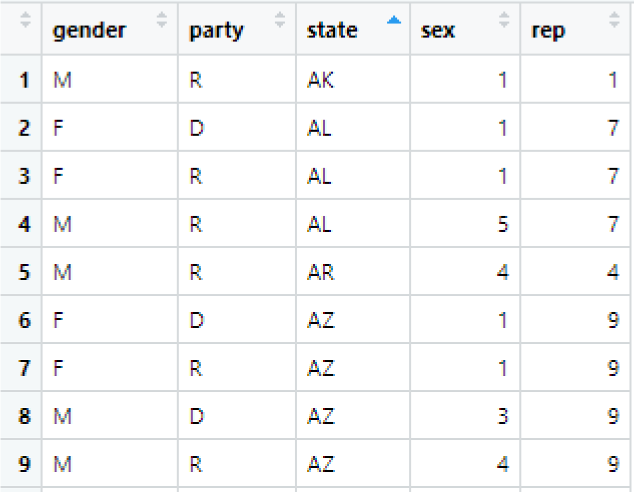
	gender	party	state	sex
1	F	D	AL	1
2	F	D	AZ	1
3	F	D	CA	16
4	F	D	CO	1
5	F	D	CT	2
6	F	D	DE	1
7	F	D	FL	6
8	F	D	HI	2
9	F	D	IL	3



	state	rep
1	AK	1
2	AL	7
3	AR	4
4	AZ	9
5	CA	53
6	CO	7
7	CT	5
8	DE	1
9	FL	27

```
house_115_sp <- house_115_sp %>%
  left_join(house_115_state, by = "state")
```

Every state’s party gender ratio (Practice 1)



	gender	party	state	sex	rep
1	M	R	AK	1	1
2	F	D	AL	1	7
3	F	R	AL	1	7
4	M	R	AL	5	7
5	M	R	AR	4	4
6	F	D	AZ	1	9
7	F	R	AZ	1	9
8	M	D	AZ	3	9
9	M	R	AZ	4	9

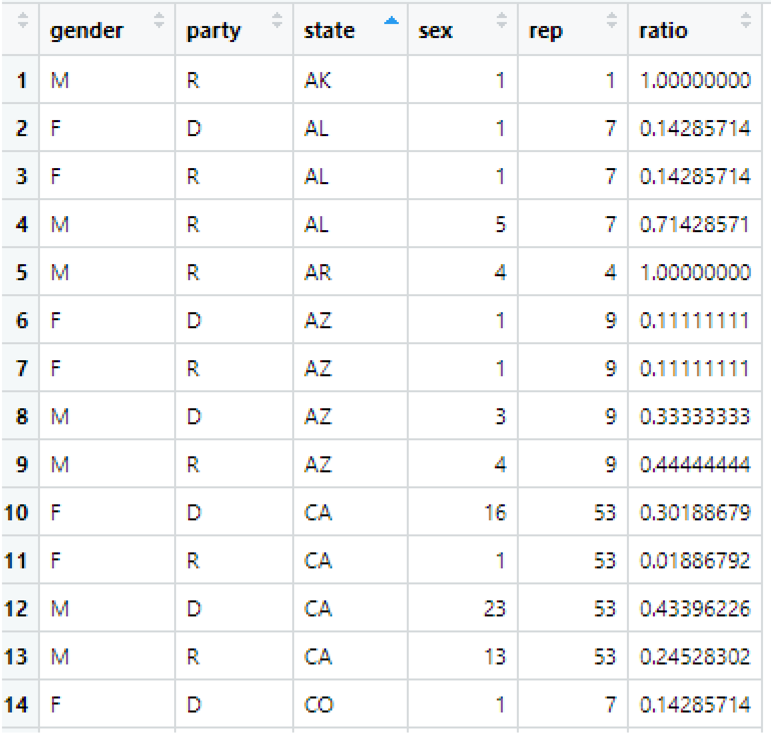
```
house_115_sp <- house_115_sp %>%
  mutate(ratio = sex / rep)
```

02

Data Mining

Gender Issues in US Congress

Every state’s party gender ratio (Practice 1)



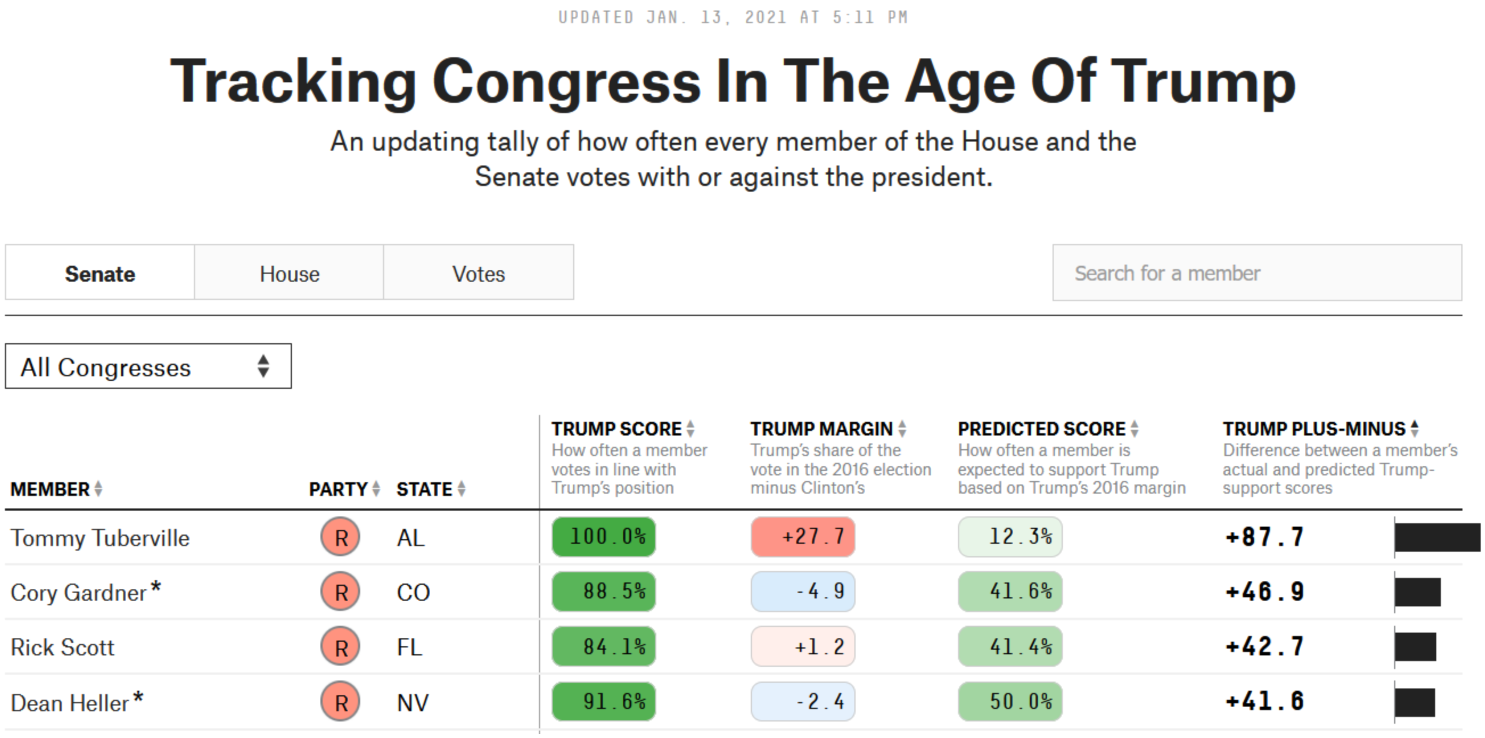
	gender	party	state	sex	rep	ratio
1	M	R	AK	1	1	1.00000000
2	F	D	AL	1	7	0.14285714
3	F	R	AL	1	7	0.14285714
4	M	R	AL	5	7	0.71428571
5	M	R	AR	4	4	1.00000000
6	F	D	AZ	1	9	0.11111111
7	F	R	AZ	1	9	0.11111111
8	M	D	AZ	3	9	0.33333333
9	M	R	AZ	4	9	0.44444444
10	F	D	CA	16	53	0.30188679
11	F	R	CA	1	53	0.01886792
12	M	D	CA	23	53	0.43396226
13	M	R	CA	13	53	0.24528302
14	F	D	CO	1	7	0.14285714

02 Data Mining

Gender Issues in US Congress

2. Were female lawmakers less likely to support Trump?

Read trumpscore.xlsx



UPDATED JAN. 13, 2021 AT 5:11 PM

Tracking Congress In The Age Of Trump

An updating tally of how often every member of the House and the Senate votes with or against the president.

MEMBER	PARTY	STATE	TRUMP SCORE How often a member votes in line with Trump's position	TRUMP MARGIN Trump's share of the vote in the 2016 election minus Clinton's	PREDICTED SCORE How often a member is expected to support Trump based on Trump's 2016 margin	TRUMP PLUS-MINUS Difference between a member's actual and predicted Trump-support scores
Tommy Tuberville	R	AL	100.0%	+27.7	12.3%	+87.7
Cory Gardner*	R	CO	88.5%	-4.9	41.6%	+46.9
Rick Scott	R	FL	84.1%	+1.2	41.4%	+42.7
Dean Heller*	R	NV	91.6%	-2.4	50.0%	+41.6

Data Mining

Gender Issues in US Congress

2. Were female lawmakers less likely to support Trump?

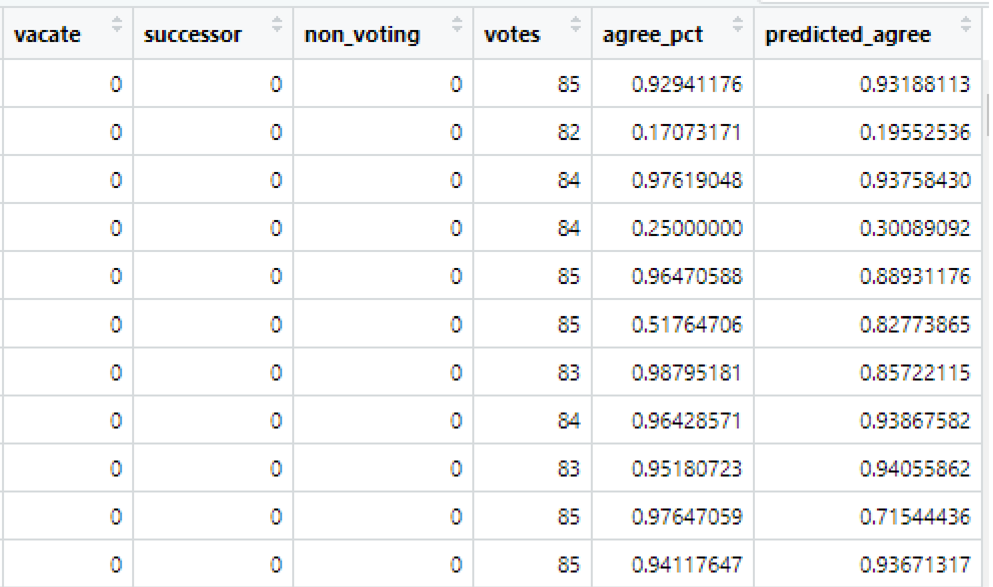
Post your codes in Moodle (Practice 2)

02 Data Mining

Gender Issues in US Congress

2. Were female lawmakers less likely to support Trump?

```
house_115_2016 <- house_115_2016 %>%
    left_join(trumpvote[, c(2, 7:10)],
    by = c("id" = "bioguide"))
```



vacate	successor	non_voting	votes	agree_pct	predicted_agree
0	0	0	85	0.92941176	0.93188113
0	0	0	82	0.17073171	0.19552536
0	0	0	84	0.97619048	0.93758430
0	0	0	84	0.25000000	0.30089092
0	0	0	85	0.96470588	0.88931176
0	0	0	85	0.51764706	0.82773865
0	0	0	83	0.98795181	0.85722115
0	0	0	84	0.96428571	0.93867582
0	0	0	83	0.95180723	0.94055862
0	0	0	85	0.97647059	0.71544436
0	0	0	85	0.94117647	0.93671317

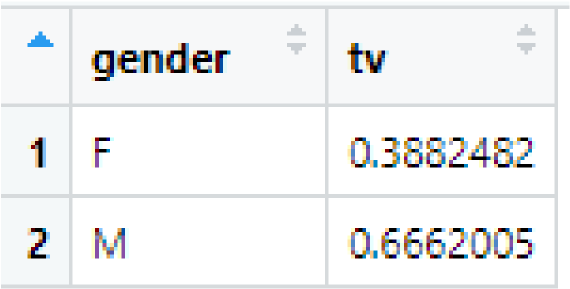
02

Data Mining

Gender Issues in US Congress

2. Were female lawmakers less likely to support Trump?

```
house_115_trump <- house_115_2016 %>%
  group_by(gender) %>%
  summarise(tv = mean(agree_pct, na.rm = TRUE))
```



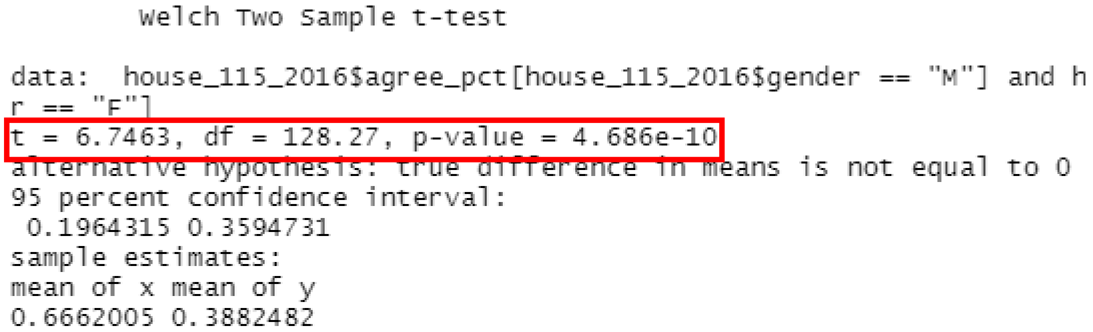
	gender	tv
1	F	0.3882482
2	M	0.6662005

02 Data Mining

Gender Issues in US Congress

2. Were female lawmakers less likely to support Trump?

t.test(house_115_2016$agree_pct[house_115_2016$gender == "M"], house_115_2016$agree_pct[house_115_2016$gender == "F"])



```
        Welch Two Sample t-test

data:  house_115_2016$agree_pct[house_115_2016$gender == "M"] and h
r == "F"]
t = 6.7463, df = 128.27, p-value = 4.686e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1964315 0.3594731
sample estimates:
mean of x mean of y
0.6662005 0.3882482
```

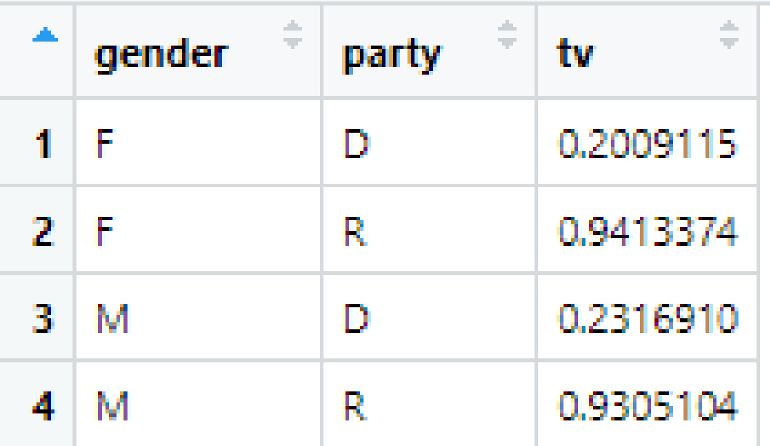
02

Data Mining

Gender Issues in US Congress

2. Were female lawmakers less likely to support Trump?

```
house_115_trump_p <- house_115_2016 %>%
  group_by(gender, party) %>%
  summarise(tv = mean(agree_pct, na.rm = TRUE))
```



	gender	party	tv
1	F	D	0.2009115
2	F	R	0.9413374
3	M	D	0.2316910
4	M	R	0.9305104

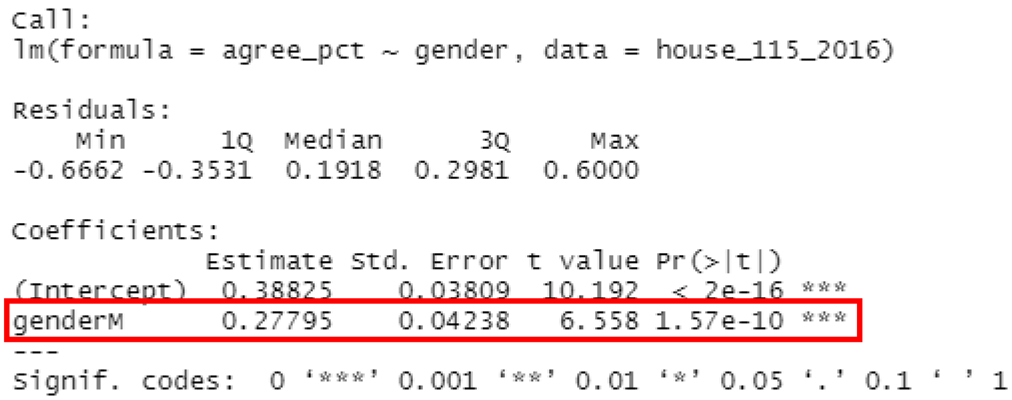
02

Data Mining

Gender Issues in US Congress

2. Were female lawmakers less likely to support Trump?

trump_support_1 <- lm(agree_pct ~ gender, data = house_115_2016)
summary(trump_support_1)



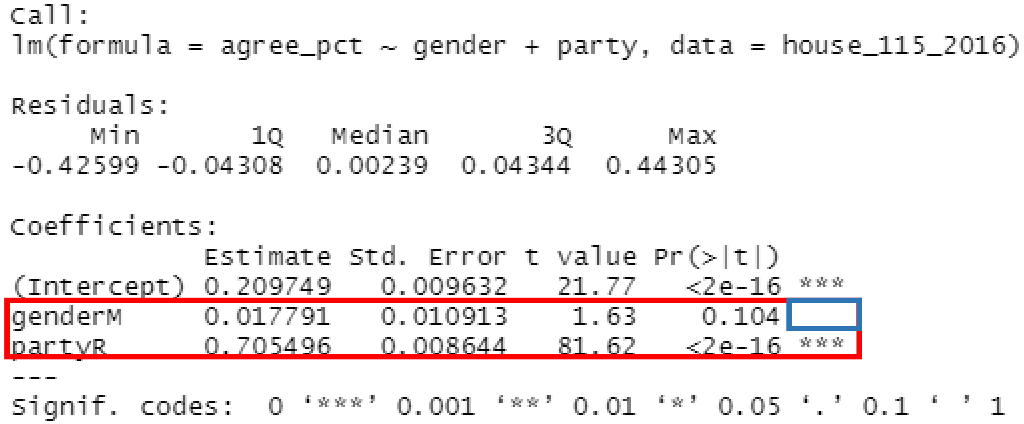
```
Call:
lm(formula = agree_pct ~ gender, data = house_115_2016)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.6662 -0.3531  0.1918  0.2981  0.6000 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.38825    0.03809  10.192  < 2e-16 ***
genderM      0.27795    0.04238   6.558 1.57e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

02 Data Mining

Gender Issues in US Congress

2. Were female lawmakers less likely to support Trump?

trump_support_2 <- lm(agree_pct ~ gender + party, data = house_115_2016)
summary(trump_support_2)



```
Call:
lm(formula = agree_pct ~ gender + party, data = house_115_2016)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.42599 -0.04308  0.00239  0.04344  0.44305 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.209749   0.009632   21.77   <2e-16 ***
genderM     0.017791   0.010913    1.63    0.104    
partyR      0.705496   0.008644   81.62   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Basic Data Visualization

02

Basic Data Visualization

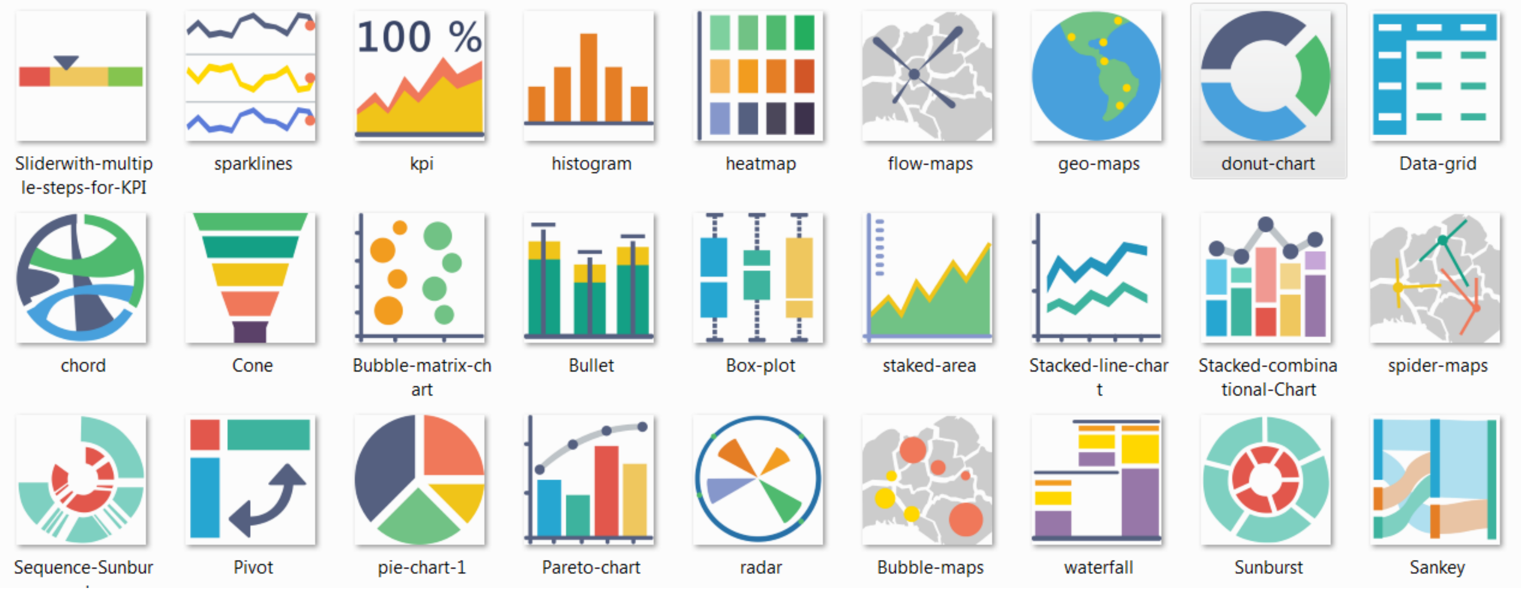
Introduction

Data visualization:

The graphical representation of information and data by using charts, graphs, maps, and other data visualization tools (tableau.com)

02 Basic Data Visualization

Introduction



Choosing a way to present your data is depended on your data's content and audience.

Good news is that R can produce all above plots.

02

Basic Data Visualization

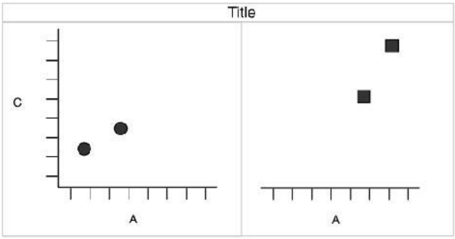
ggplot2

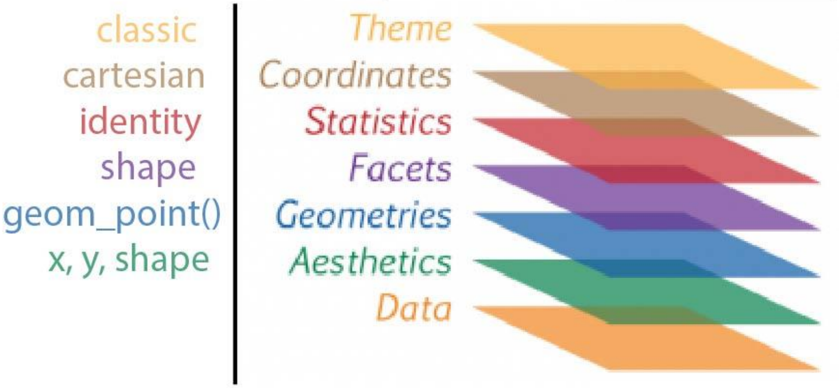
ggplot2 is called the grammar of graphics.

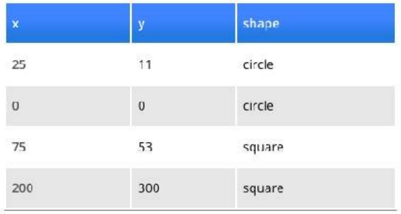
```
Install.packages(“ggplot2”)
library(ggplot2)
```

02 Basic Data Visualization

ggplot2







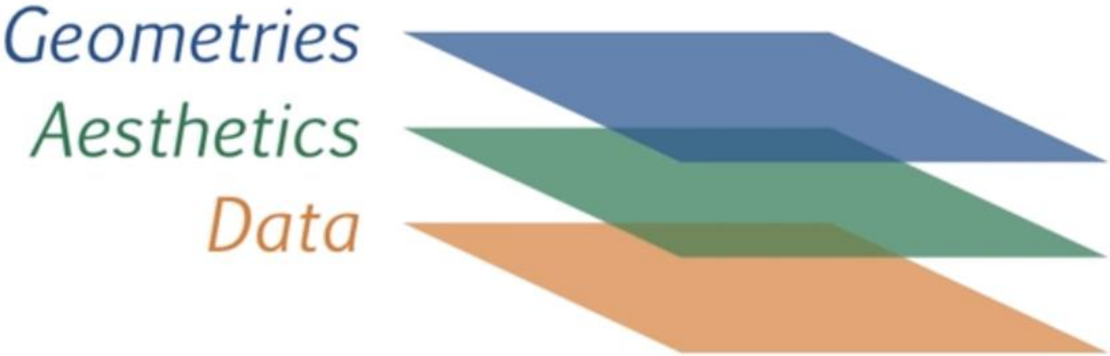
x	y	shape
25	11	circle
0	0	circle
75	53	square
200	300	square

ggplot2 creates plots by layers

02

Basic Data Visualization

ggplot2



The first three are essential layers.

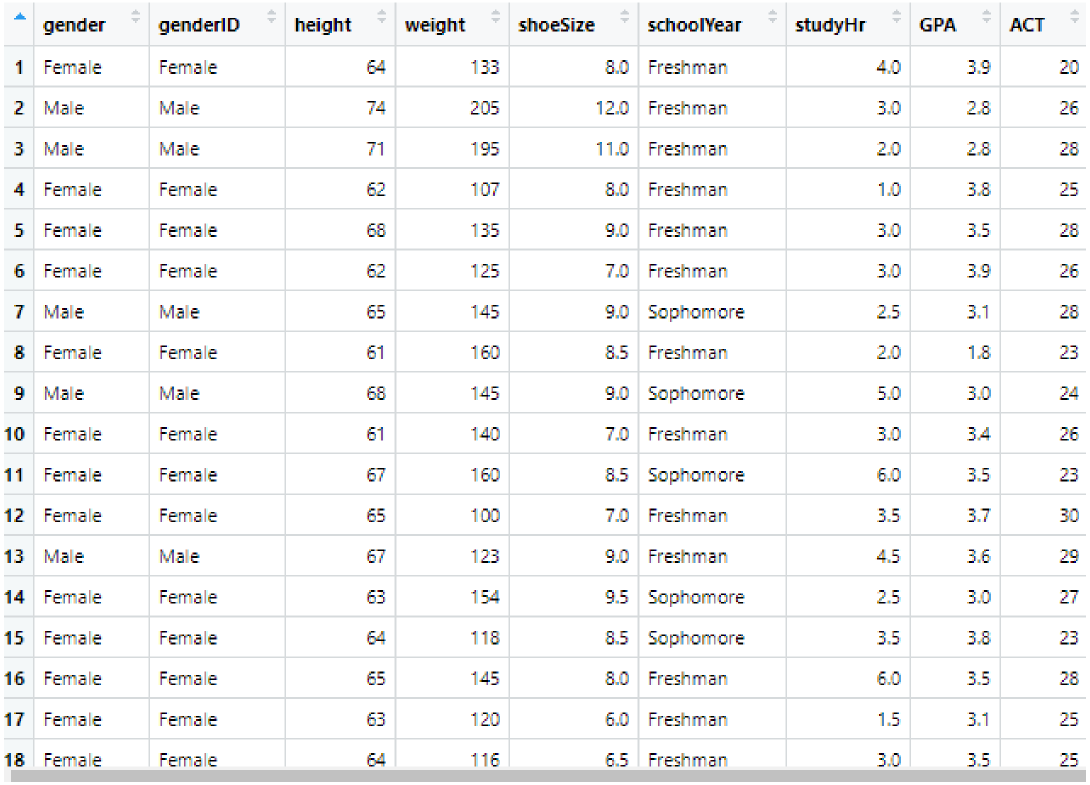
They are the data layer, aesthetics layer, and geometrics layer.

02

Basic Data Visualization

ggplot2 Basics

Read the data of USA GPA scores
gpa <- read.csv("gpa.csv")
To see how to produce ggplot charts

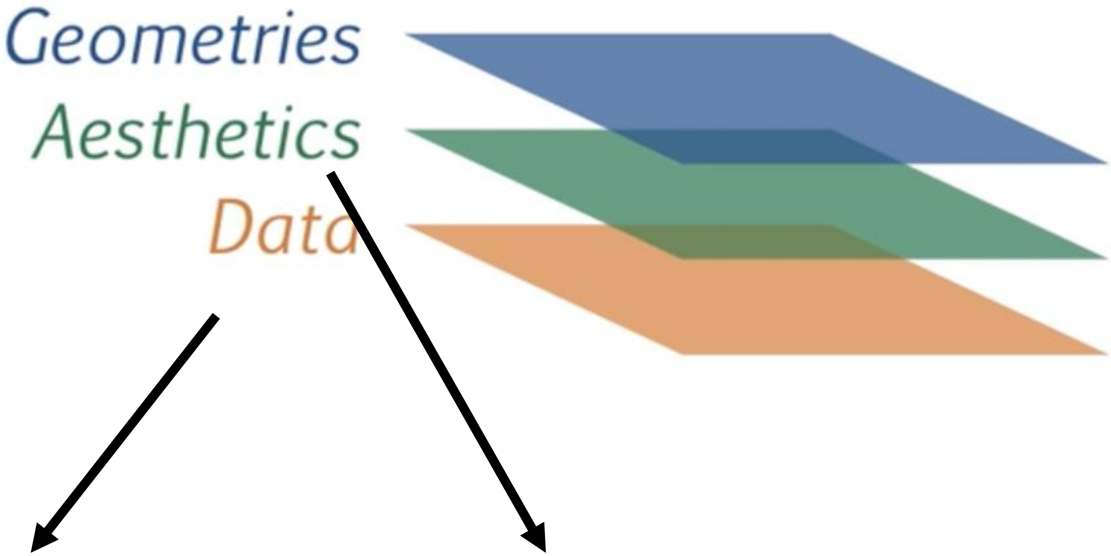


	gender	genderID	height	weight	shoeSize	schoolYear	studyHr	GPA	ACT
1	Female	Female	64	133	8.0	Freshman	4.0	3.9	20
2	Male	Male	74	205	12.0	Freshman	3.0	2.8	26
3	Male	Male	71	195	11.0	Freshman	2.0	2.8	28
4	Female	Female	62	107	8.0	Freshman	1.0	3.8	25
5	Female	Female	68	135	9.0	Freshman	3.0	3.5	28
6	Female	Female	62	125	7.0	Freshman	3.0	3.9	26
7	Male	Male	65	145	9.0	Sophomore	2.5	3.1	28
8	Female	Female	61	160	8.5	Freshman	2.0	1.8	23
9	Male	Male	68	145	9.0	Sophomore	5.0	3.0	24
10	Female	Female	61	140	7.0	Freshman	3.0	3.4	26
11	Female	Female	67	160	8.5	Sophomore	6.0	3.5	23
12	Female	Female	65	100	7.0	Freshman	3.5	3.7	30
13	Male	Male	67	123	9.0	Freshman	4.5	3.6	29
14	Female	Female	63	154	9.5	Sophomore	2.5	3.0	27
15	Female	Female	64	118	8.5	Sophomore	3.5	3.8	23
16	Female	Female	65	145	8.0	Freshman	6.0	3.5	28
17	Female	Female	63	120	6.0	Freshman	1.5	3.1	25
18	Female	Female	64	116	6.5	Freshman	3.0	3.5	25

02

Basic Data Visualization

ggplot2 Basics

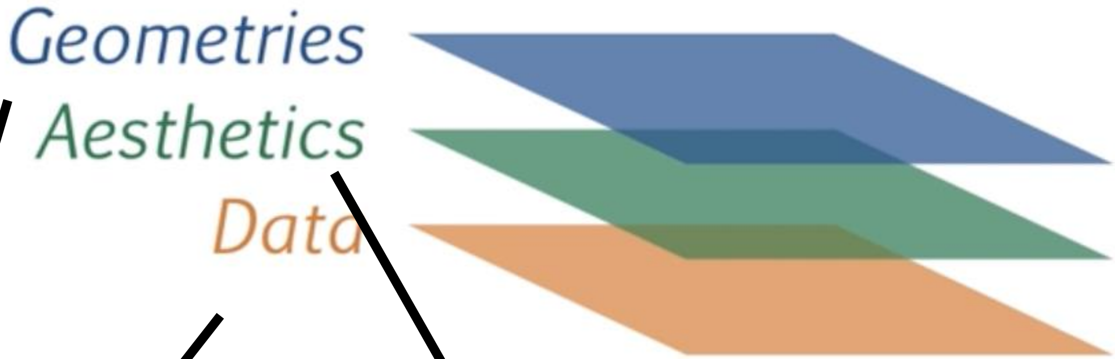


```
ggplot(gpa, aes(x = studyHr, y = GPA))
```

02

Basic Data Visualization

ggplot2 Basics



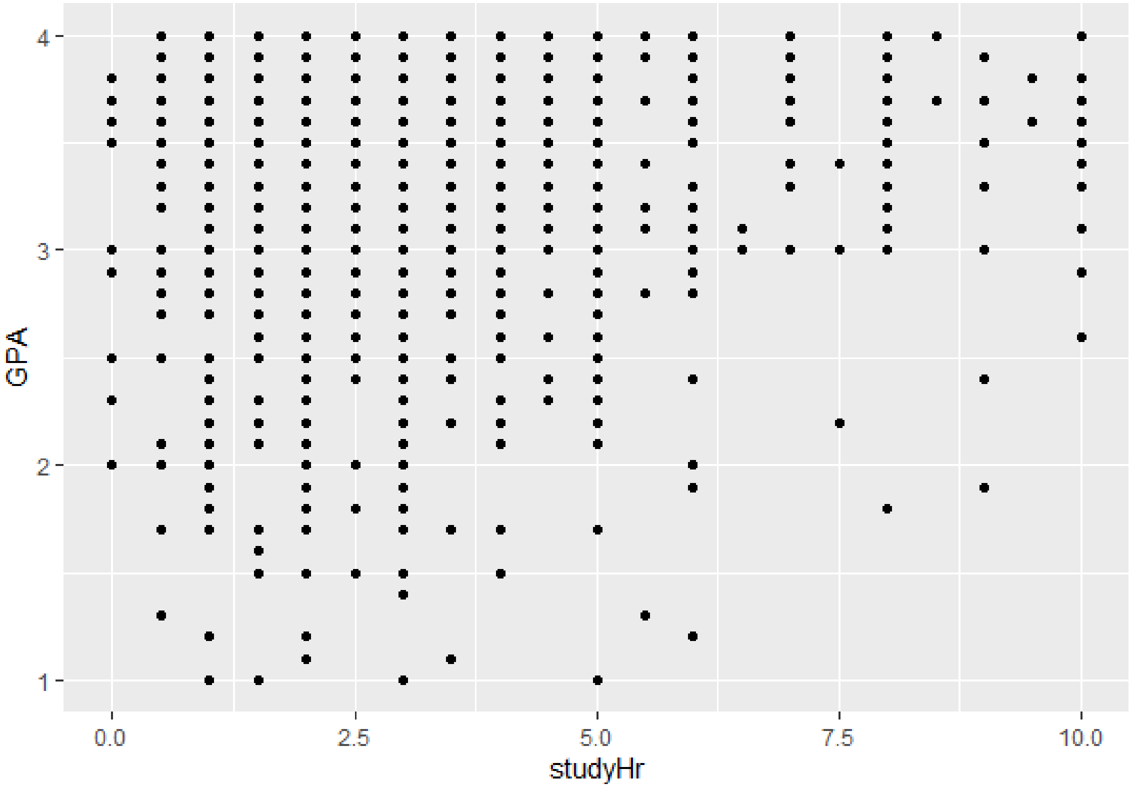
```
ggplot(gpa, aes(x = studyHr, y = GPA)) +
    geom_point()
```

02

Basic Data Visualization

ggplot2 Basics

```
ggplot(gpa, aes(x = studyHr, y = GPA)) +
    geom_point()
```

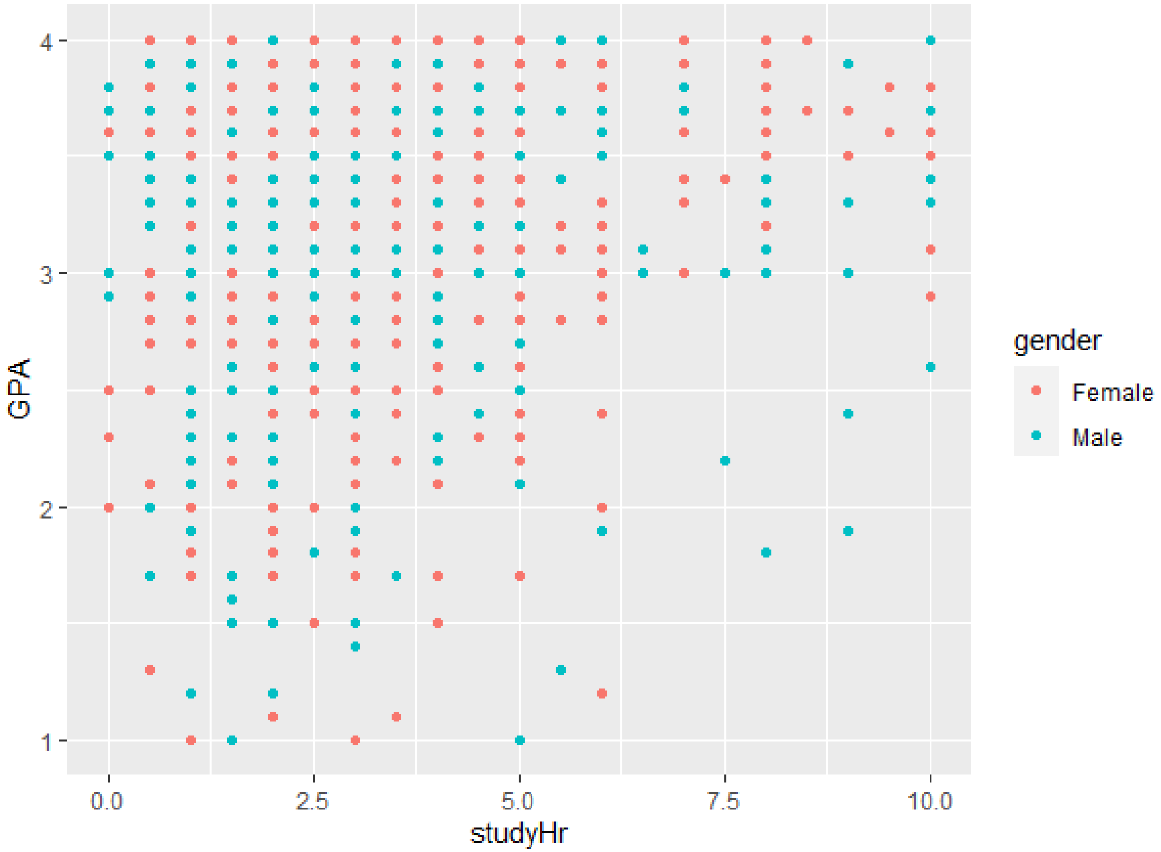


02

Basic Data Visualization

ggplot2 Basics

ggplot(gpa, aes(x = studyHr, y = GPA, color = gender)) + geom_point()

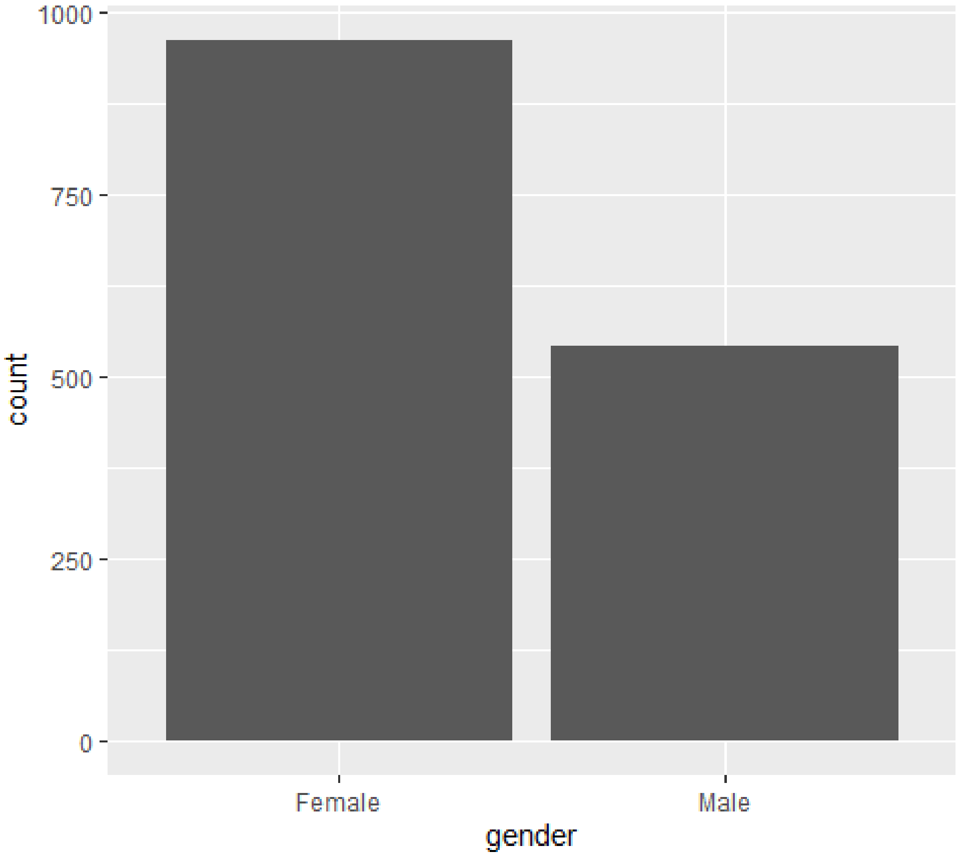


02

Basic Data Visualization

ggplot2 Basics

```
ggplot(gpa, aes(gender)) +
  geom_bar()
```



02

Basic Data Visualization

ggplot2 Basics

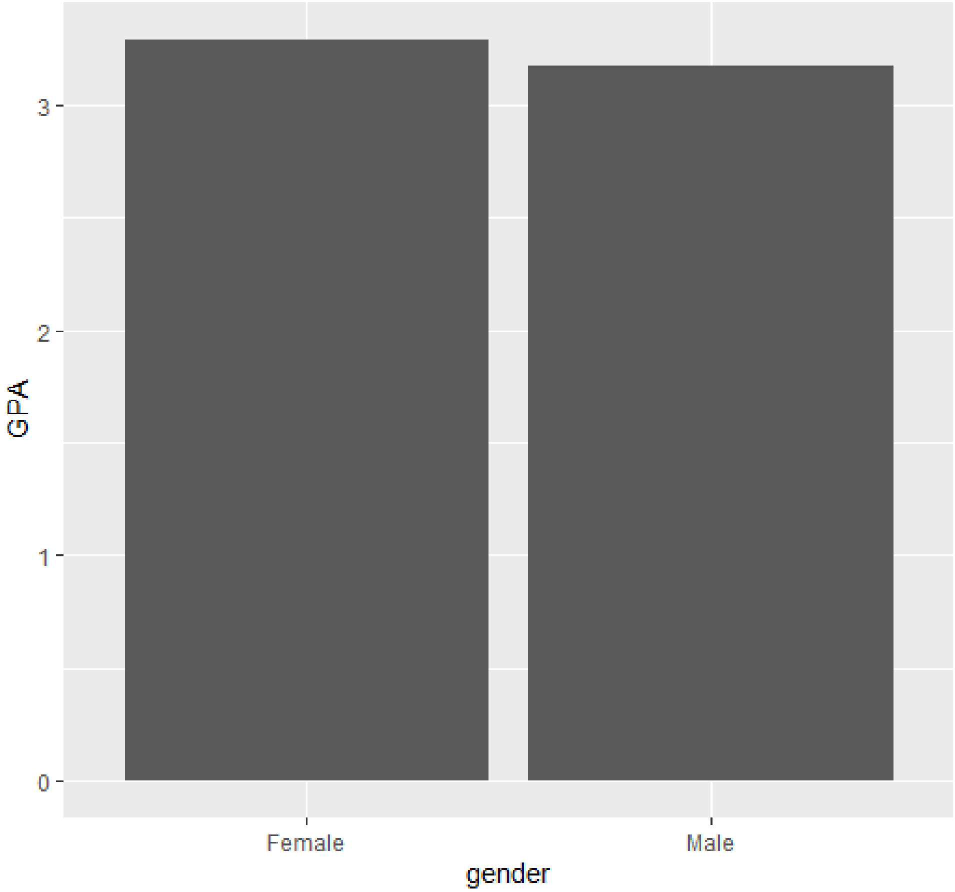
Male and female have different performance of GPA?

```
gender <- gpa %>%
  group_by(gender) %>%
  summarise(GPA = mean(GPA))
```

Basic Data Visualization

```
ggplot(gender, aes(gender, GPA)) +
  geom_bar(stat = "identity")
```

identity means that you have Y value

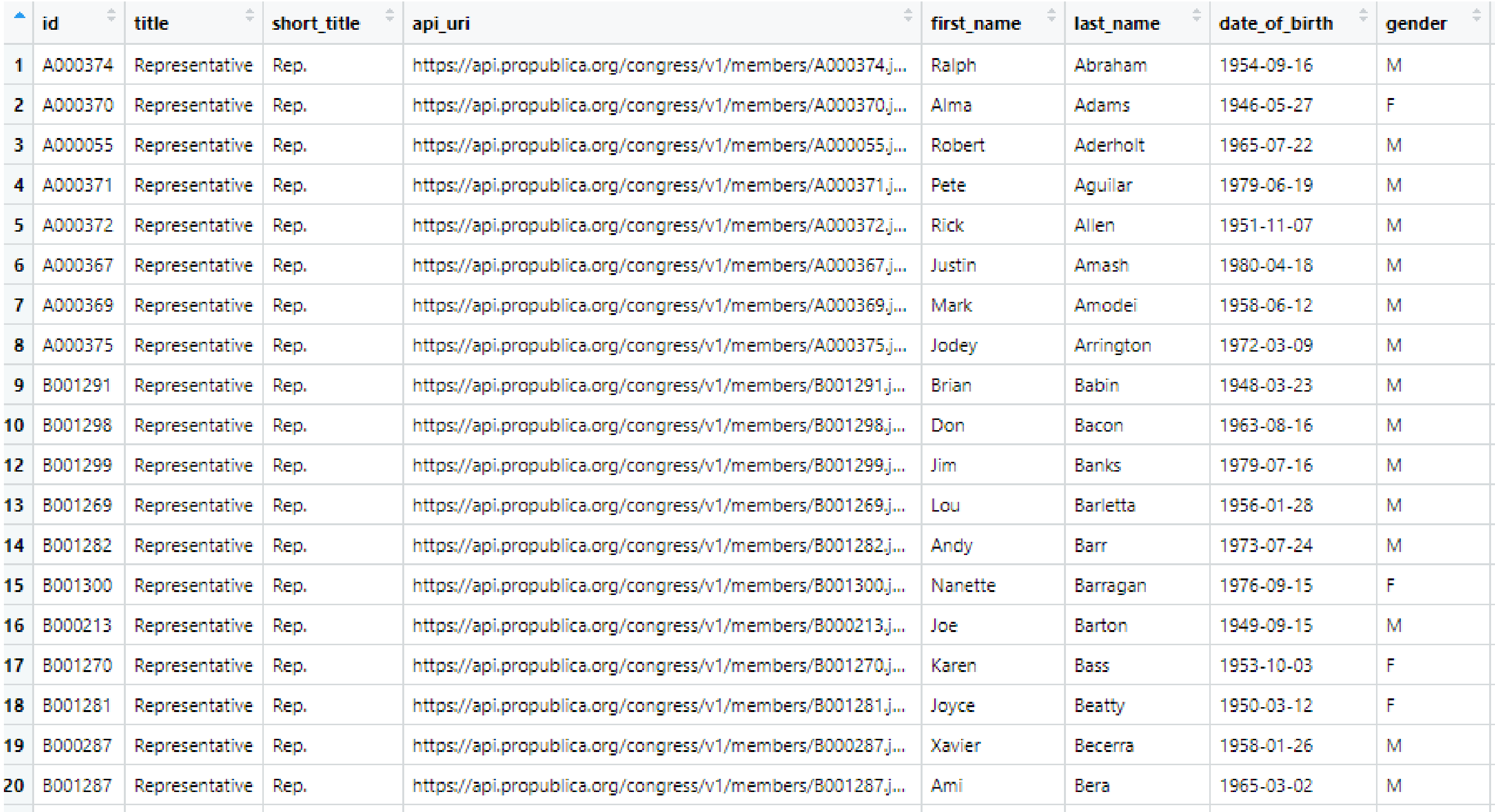


02

Basic Data Visualization

US Congress

Let’s go back to our house_115_2016, house_116_2019, and house_117_2021

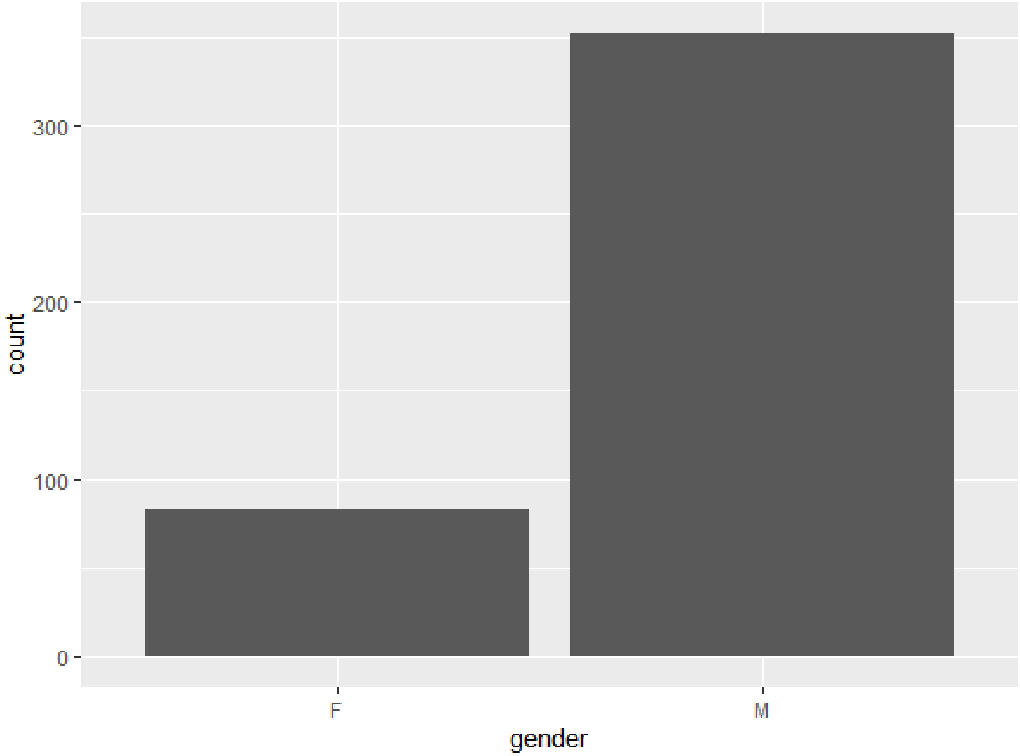


	id	title	short_title	api_uri	first_name	last_name	date_of_birth	gender
1	A000374	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000374.j...	Ralph	Abraham	1954-09-16	M
2	A000370	Representative	Rep.	https://api.propublica.org/congress/v1/members/A000370.j...	Alma	Adams	1946-05-27	F
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
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	Barragan	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

Basic Data Visualization

house_115_2016’s gender distribution

```
ggplot(house_115_2016, aes(gender)) +
  geom_bar()
```



02

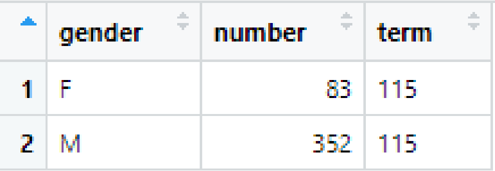
Basic Data Visualization

US Congress

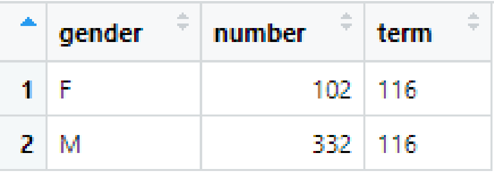
One plot to show three terms’ gender distribution

Create a new dataframe that includes three terms’ gender information

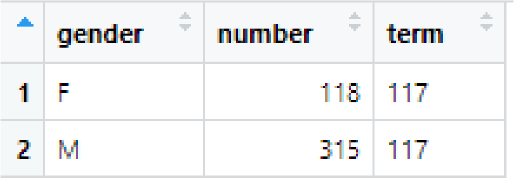
```
house_gender_115 <- house_115_2016 %>%
  group_by(gender) %>%
  summarise(number = n()) %>%
  mutate(term = "115")
```



	gender	number	term
1	F	83	115
2	M	352	115



	gender	number	term
1	F	102	116
2	M	332	116



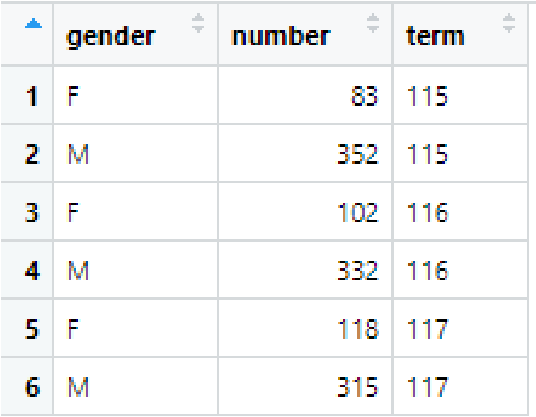
	gender	number	term
1	F	118	117
2	M	315	117

rbind() them!

02 Basic Data Visualization

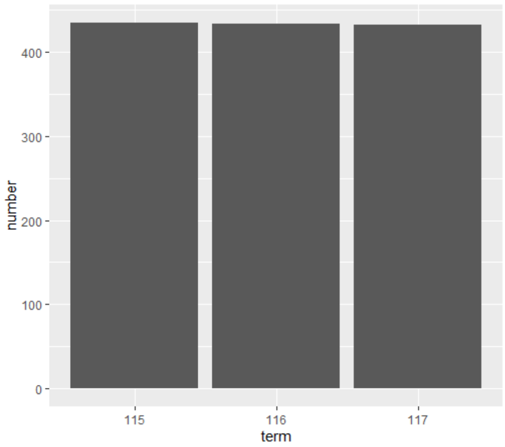
US Congress

```
house_gender <- rbind(house_gender_115, house_gender_116,
house_gender_117)
```



	gender	number	term
1	F	83	115
2	M	352	115
3	F	102	116
4	M	332	116
5	F	118	117
6	M	315	117

```
ggplot(house_gender, aes(term, number)) +
  geom_bar(stat = "identity")
```

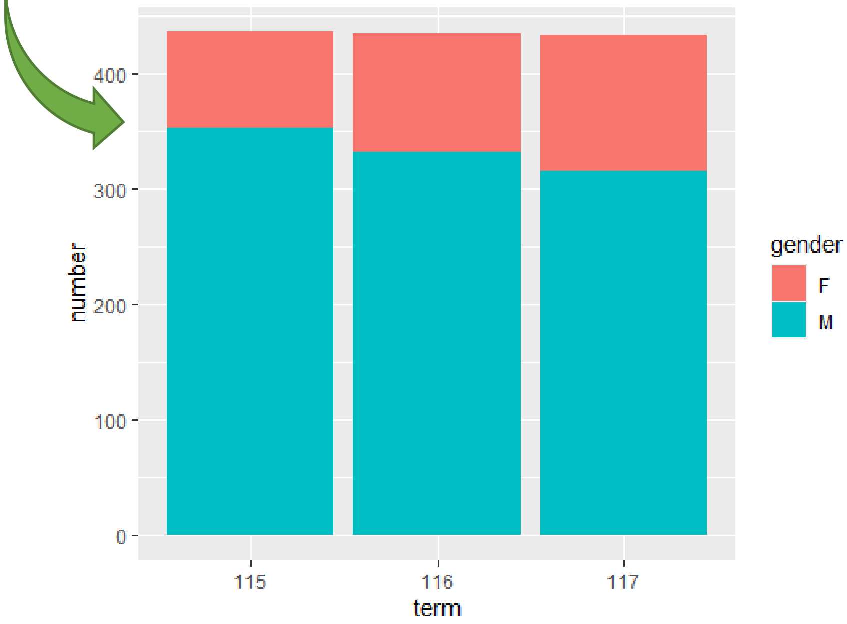


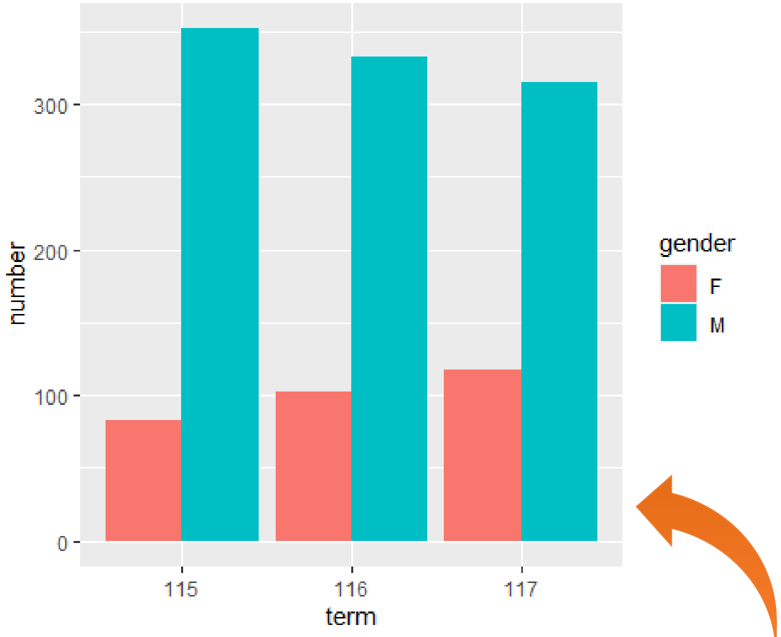
02

Basic Data Visualization

US Congress

```
ggplot(house_gender, aes(term, number,  fill = gender)) +
  geom_bar(stat = "identity")
```





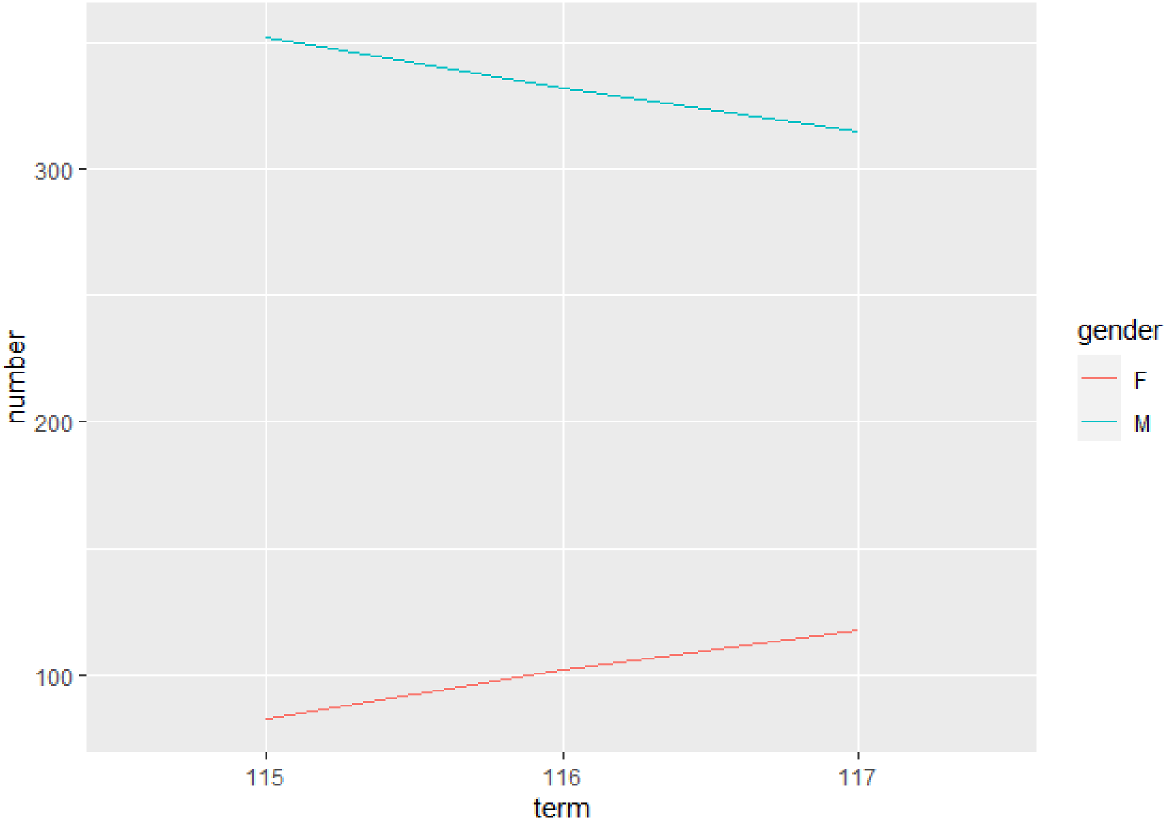
```
ggplot(house_gender, aes(term, number,  fill = gender)) +
  geom_bar(stat = "identity", position = position_dodge())
```

02

Basic Data Visualization

US Congress

```
ggplot(house_gender, aes(term, number, group = gender)) +
  geom_line(aes(color = gender))
```



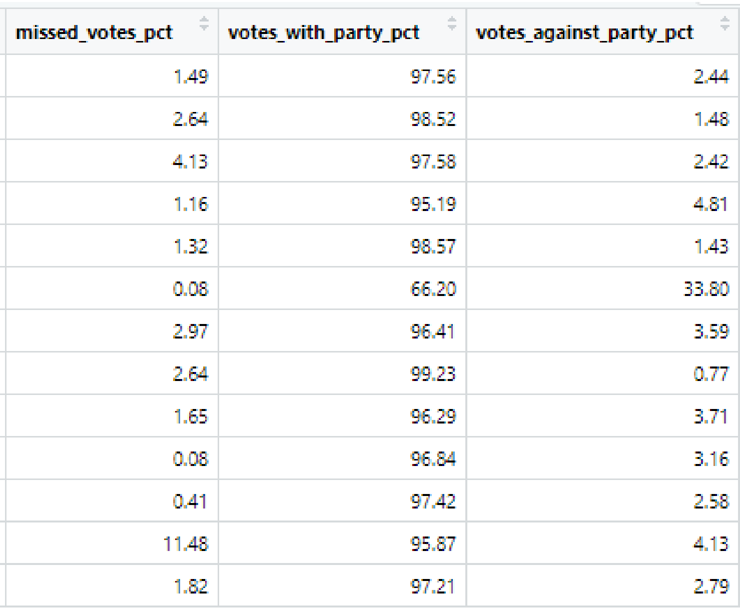
02

Basic Data Visualization

US Congress

The votes_against_party_pct column show the percentage of a lawmaker's votes which didn't follow party's orders.

Draw a plot to show the degree of party controls on lawmakers change from 115-117 terms (Practice 3)

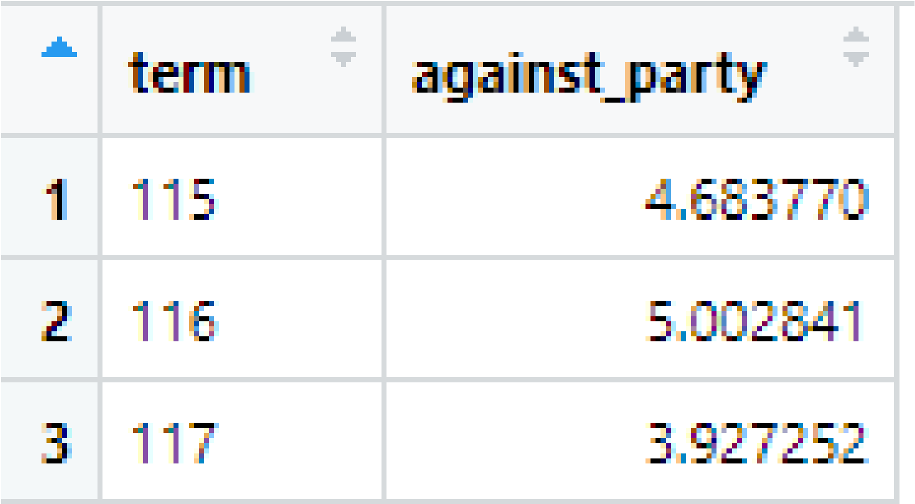


missed_votes_pct	votes_with_party_pct	votes_against_party_pct
1.49	97.56	2.44
2.64	98.52	1.48
4.13	97.58	2.42
1.16	95.19	4.81
1.32	98.57	1.43
0.08	66.20	33.80
2.97	96.41	3.59
2.64	99.23	0.77
1.65	96.29	3.71
0.08	96.84	3.16
0.41	97.42	2.58
11.48	95.87	4.13
1.82	97.21	2.79

02 Basic Data Visualization

US Congress

```
party_vote <- data.frame(term = c("115", "116", "117"),
                         against_party =
c(mean(house_115_2016$votes_against_party_pct),
  mean(house_116_2019$votes_against_party_pct, na.rm = TRUE),
  mean(house_117_2021$votes_against_party_pct)))
```



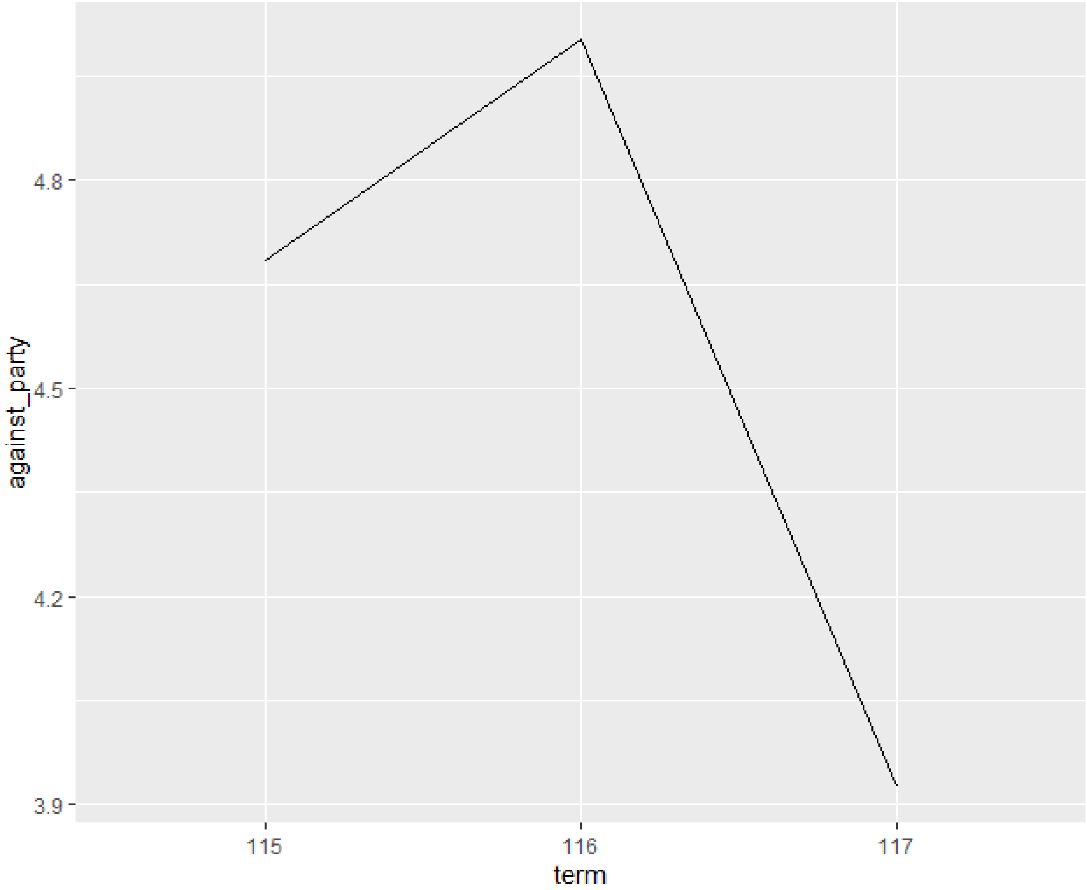
	term	against_party
1	115	4.683770
2	116	5.002841
3	117	3.927252

Basic Data Visualization

02

US Congress

```
ggplot(party_vote, aes(x = term, y = against_party, group = 1)) +
  geom_line()
```



Basic Data Visualization

US Congress

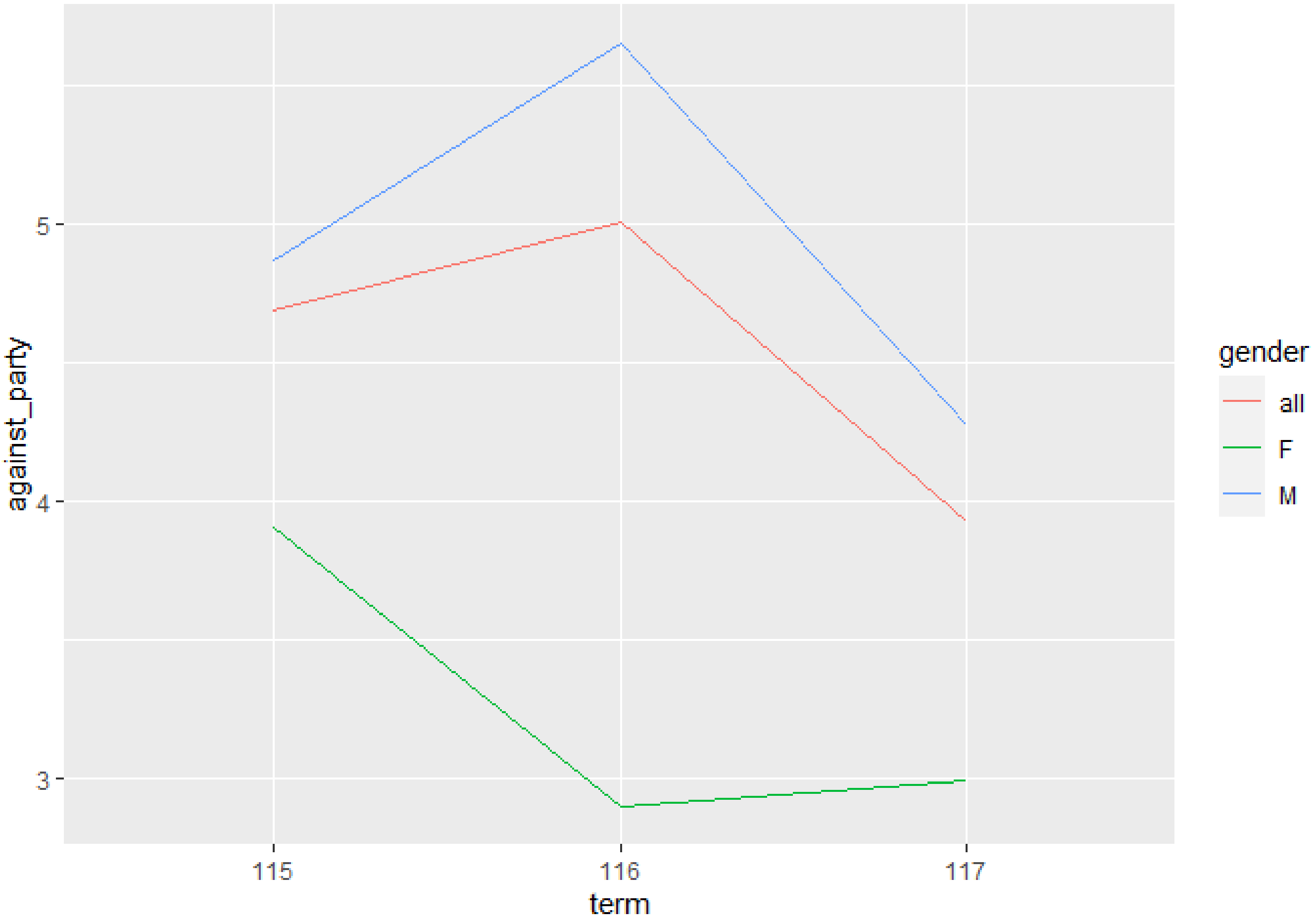
Draw a line plot that show all, male, and female laymakers' every term's mean of votes_against_party_pct

Practice 4

02

Basic Data Visualization

US Congress



Data Project Skills

03 Data Project Skills

Introduction

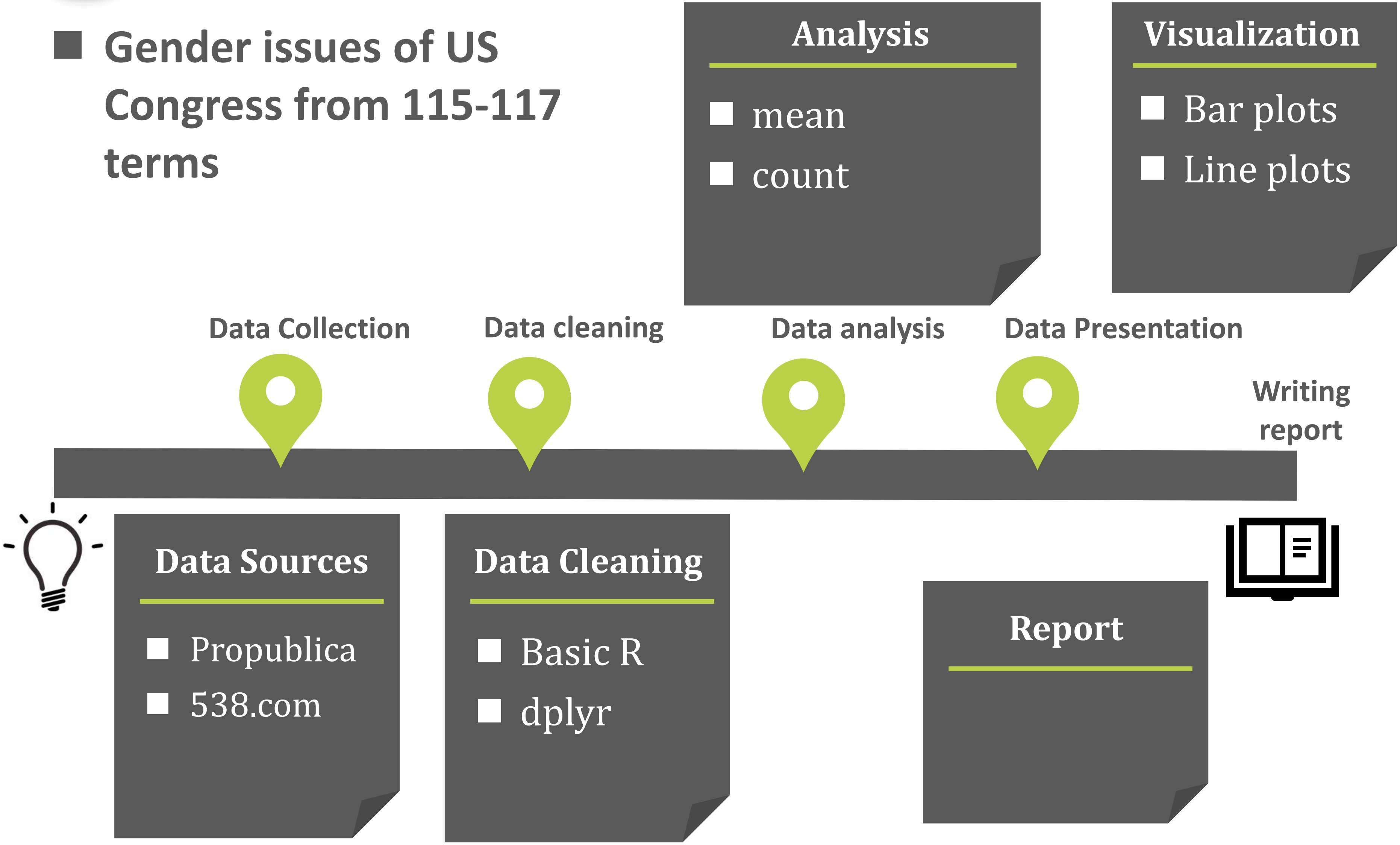
In general, you have learned basic skills of data projects.

03

Data Project Skills

Introduction

- Gender issues of US Congress from 115-117 terms



Data Collection

Data cleaning

Data analysis

Data Presentation

Writing report

Data Sources

- Propublica
- 538.com

Data Cleaning

- Basic R
- dplyr

Analysis

- mean
- count

Visualization

- Bar plots
- Line plots

Report

03

Data Project Skills

Introduction

R is your one-stop service of data analysis

Improve your R skills to:

1. **Finish all tasks in R**
2. **Never open Excel after you conduct data analysis**

03

Data Project Skills

Introduction

Basic R skills for data projects

1. Read dataset
2. Clean data
3. Data mining
4. Data analysis
5. Data visualization

All tasks can be repeated and you can insert new tasks anytime.

03

Data Project Skills

Introduction

For example, image you use excel to open house_115, delete non-voting and successors, and save the file as house_115_2016.csv.

You find you have to create house_115_2018. You have to open house_115 again and to do all things.

But in R, just insert the codes and change a bit of codes, them create another object called house_115_2018.