POLI210: Political Science Research Methods

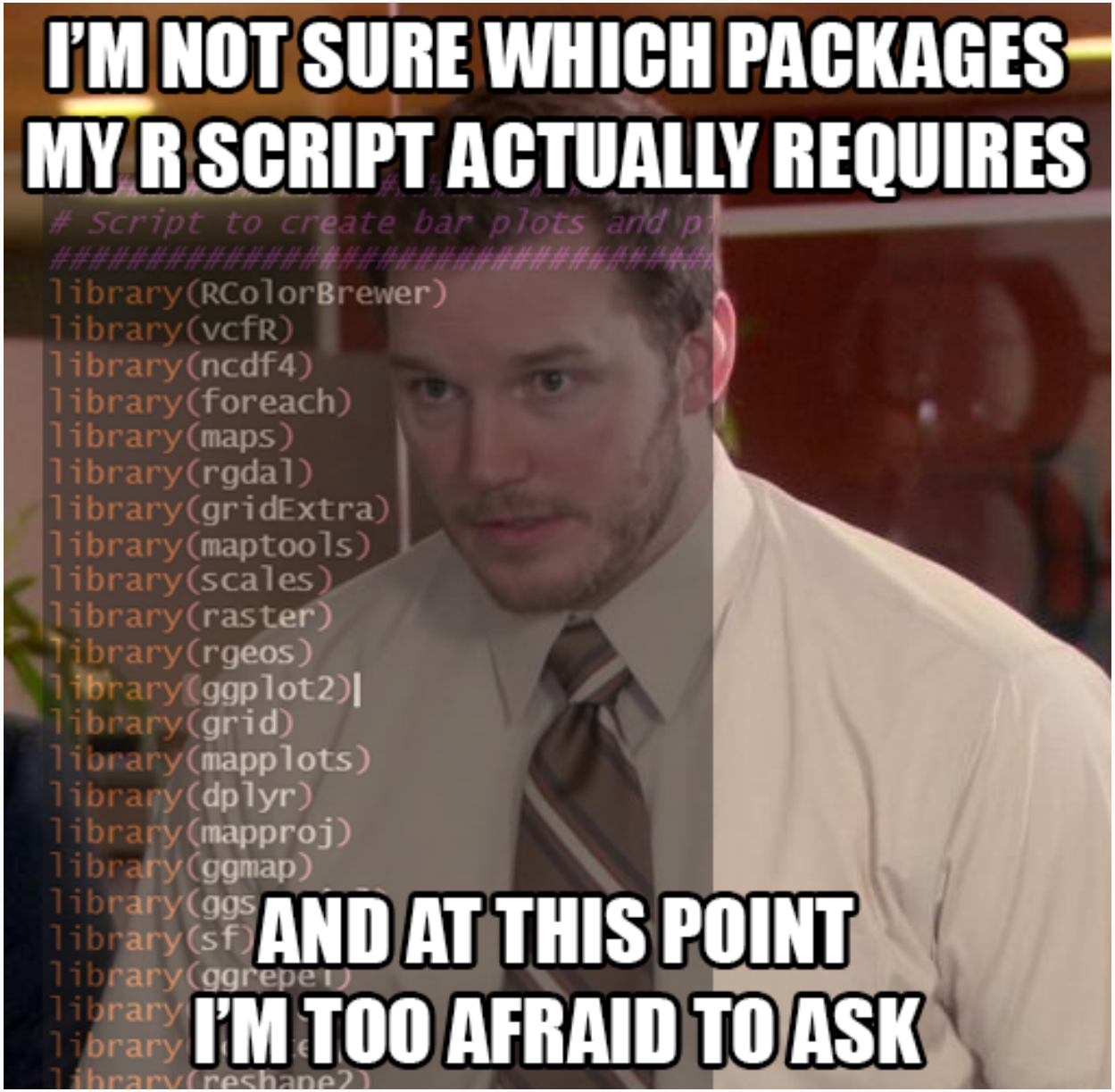
Lecture 10.2: Measures of association

Olivier Bergeron-Boutin

November 4th, 2021

```r
# taken from Andrew Heiss' website
library(ggtext)
theme_custom <- function(){
  theme_minimal(base_size = 19,
                base_family = "Fira Sans") %+replace%
  theme(legend.position = "none",
        panel.grid.minor = element_blank(),
        plot.title = element_markdown(face = "bold", size = r
        plot.subtitle = element_markdown(face = "plain", size
        axis.title = element_text(face = "bold"),
        axis.title.x = element_text(margin = margin(t = 10),
        axis.title.y = element_text(margin = margin(r = 10),
}
```

Some harmless fun



Boring admin stuff

- Problem set 4 has been posted
  - Do it: I take the 3 best grades out of 4 psets; 13.3% each
  - Don't do it: I take the 3 pset grades; 13.3% each
  - Due November 15th
- Midterm next week
  - A combination of paragraph-length answers and essays
  - Don't lose the forest for the trees!
  - Focus on the broad issues, not on specifics

Where we're going

We should now be able to describe the distribution of one variable

- The next step: describe how two variables move together
- We will speak of **correlations**
    - When one variable is big/small, does that give me a clue about whether some other variable is big/small?
- We want to judge correlations according to two criteria:
    - Direction
        - Positive correlation: when x is big, y is also big
        - Negative correlation: when x is big, y is small
    - Strength
        - How well can I guess the value of y if you give me x?
- The **correlation coefficient** summarizes both of these
    - It's a value between -1 and 1
    - Closer to -1 or 1: stronger relationship
        - Correlation of 0: no (linear) relationship
    - The sign indicates the direction

Different correlations in scatterplots

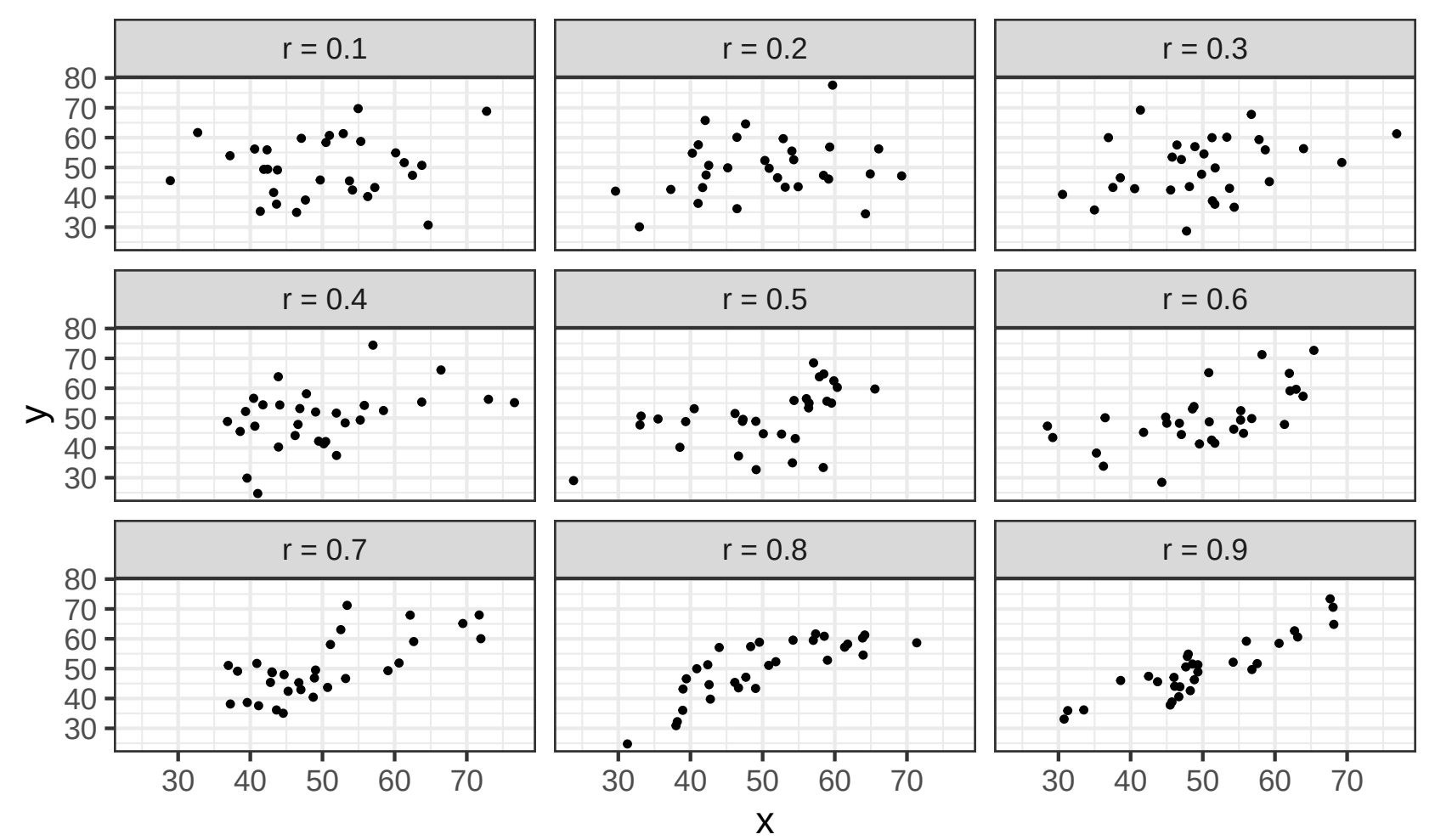


Figure 1: Scatterplots with different correlations

The scatterplot as a visual tool: economic voting

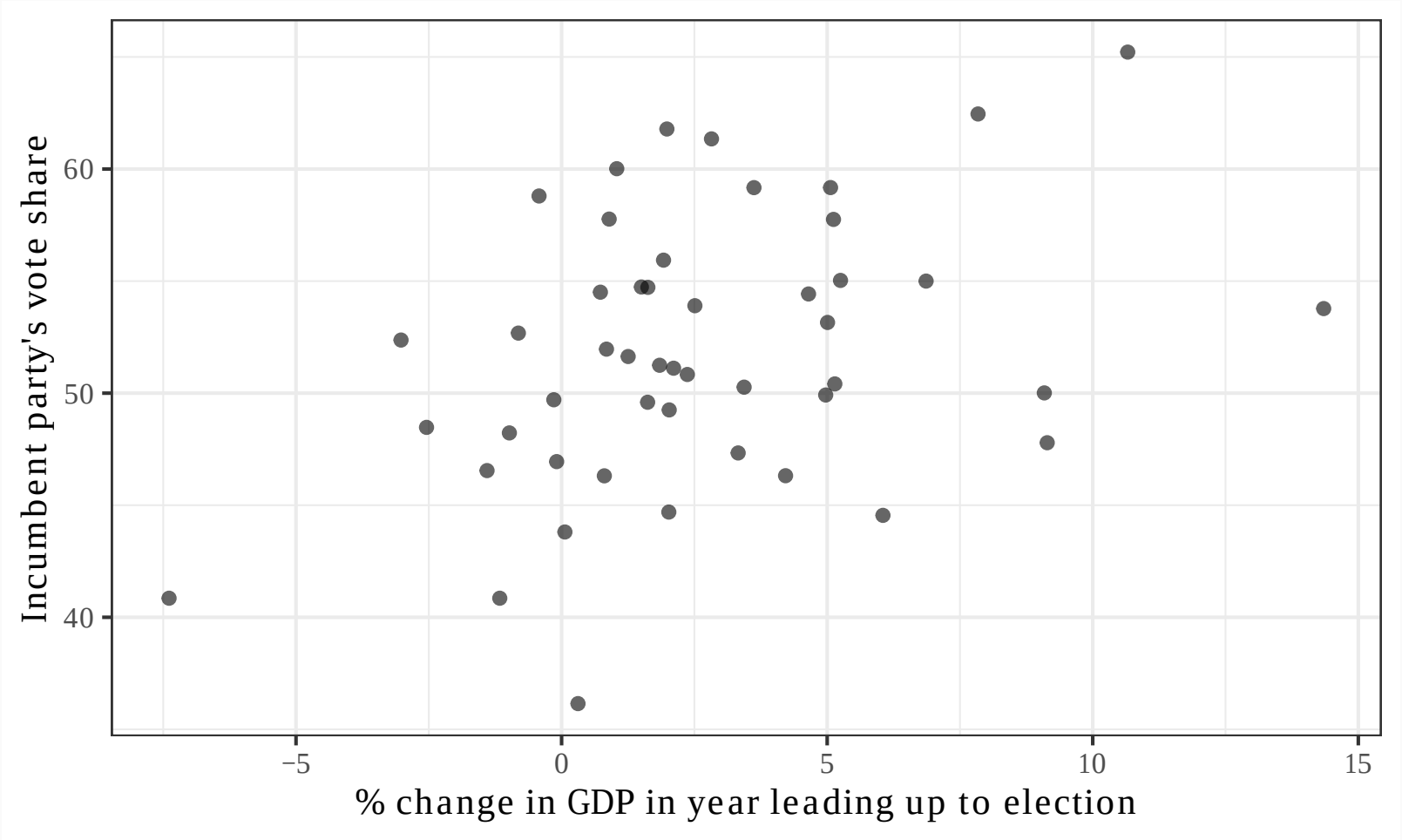


Figure 2: Relationship between economic growth and incumbent vote share in the United States, 1792-2016. Data from Guntermann, Lenz, and Myers (2021).

Economic voting

The Pearson correlation coefficient:

```
cor(economy$gdpchangeyr3, economy$partyincshr, use = "pairwise")
```

```
## [1] 0.3763856
```

A positive, moderately strong relationship

- As GDP growth increases, vote share for the incumbent tends to increase as well

r	Rough meaning
+/-0.1-0.3	Modest
+/-0.3-0.5	Moderate
+/-0.5-0.8	Strong
+/-0.8-1	Very strong

Economic voting for each party

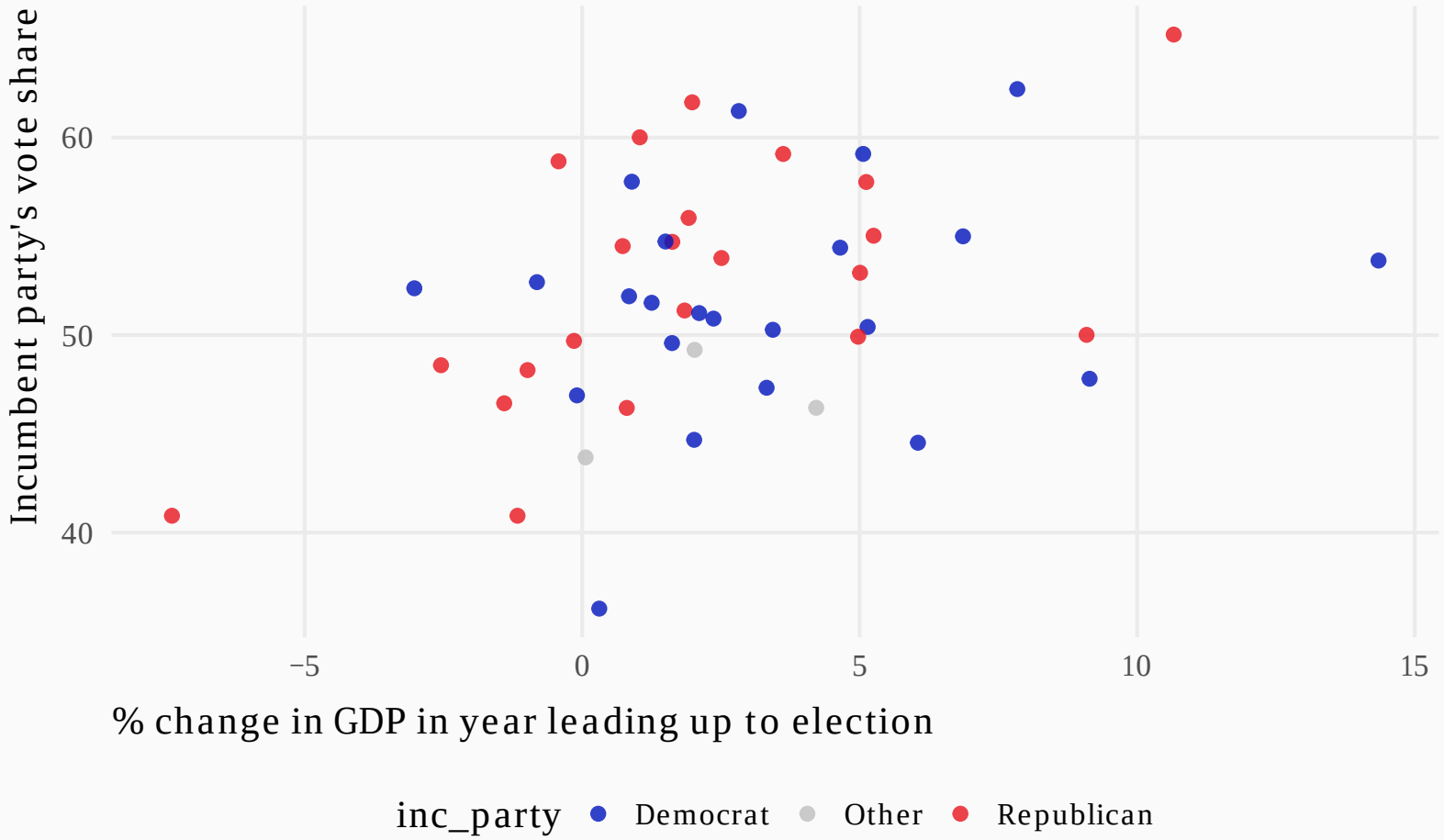


Figure 3: Relationship between economic growth and incumbent vote share in the United States, 1792-2016. Data from Guntermann, Lenz, and Myers (2021).

Economic voting for each party

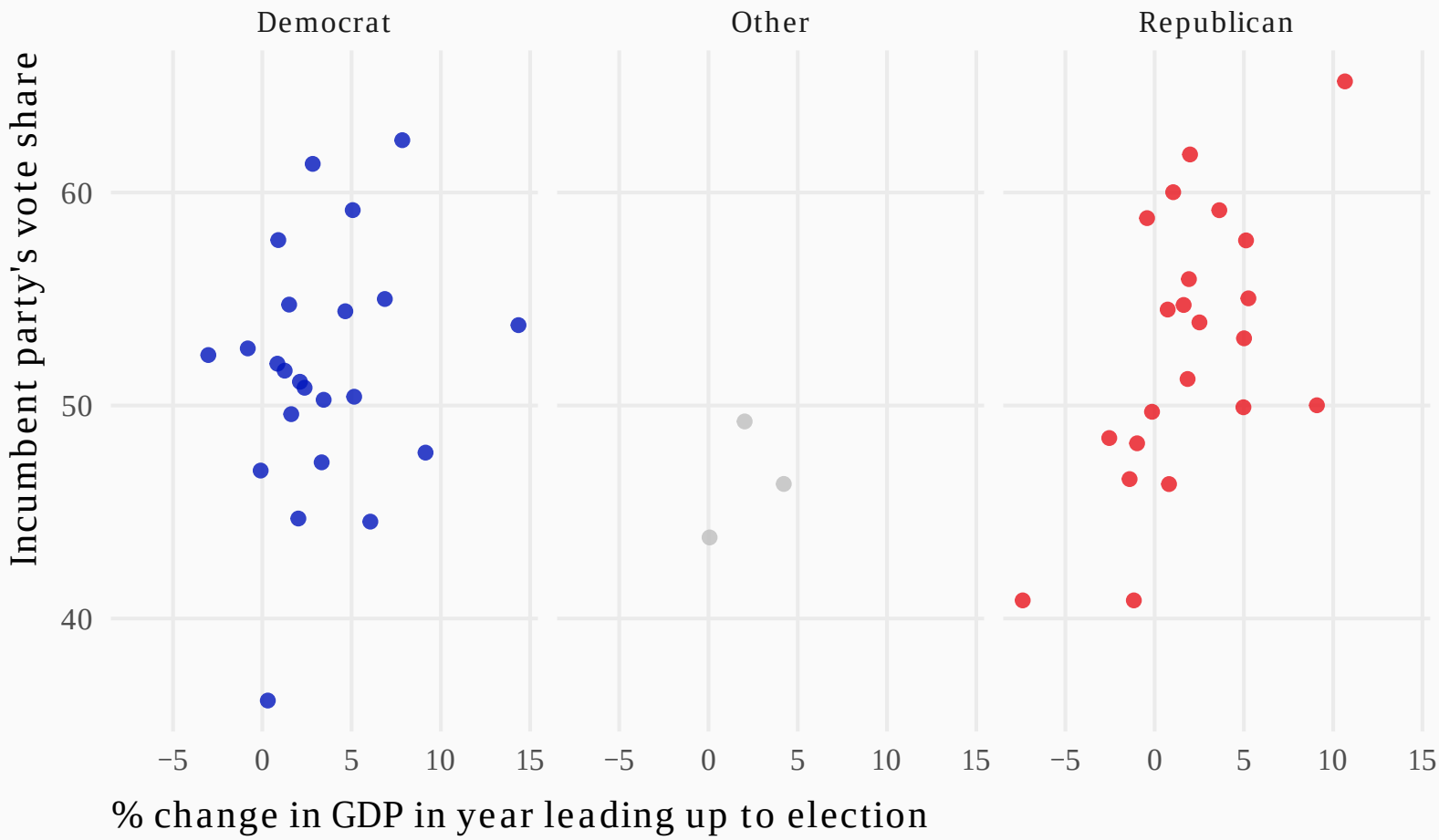


Figure 4: Relationship between economic growth and incumbent vote share in the United States, 1792-2016. Data from Guntermann, Lenz, and Myers (2021).

Economic voting for each party

```
library(tidyverse)
economy %>%
  group_by(inc_party) %>%
  summarise(cor = cor(gdpchangeyr3, partyincshr, use = "pairwise"))
```

```
## # A tibble: 3 x 2
##   inc_party    cor
##   <chr>      <dbl>
## 1 Democrat   0.206
## 2 Other      0.432
## 3 Republican 0.593
```

It looks like the correlation is stronger for Republican incumbents!

Is this a causal relationship?

Economic voting for each party

```
library(tidyverse)
economy %>%
  group_by(inc_party) %>%
  summarise(cor = cor(gdpchangeyr3, partyincshr, use = "pairwise"))
```

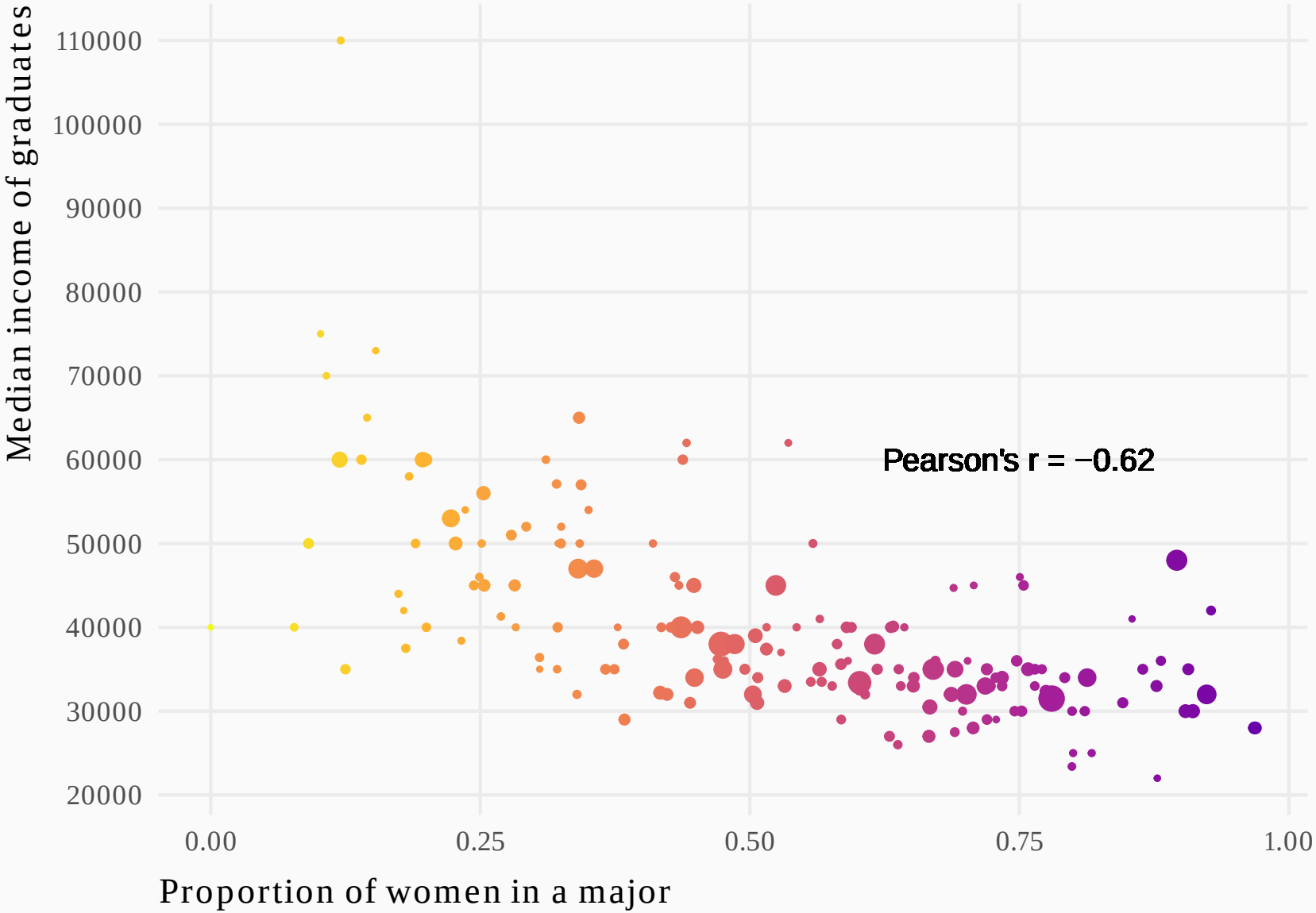
```
## # A tibble: 3 x 2
##   inc_party    cor
##   <chr>      <dbl>
## 1 Democrat   0.206
## 2 Other      0.432
## 3 Republican 0.593
```

It looks like the correlation is stronger for Republican incumbents!

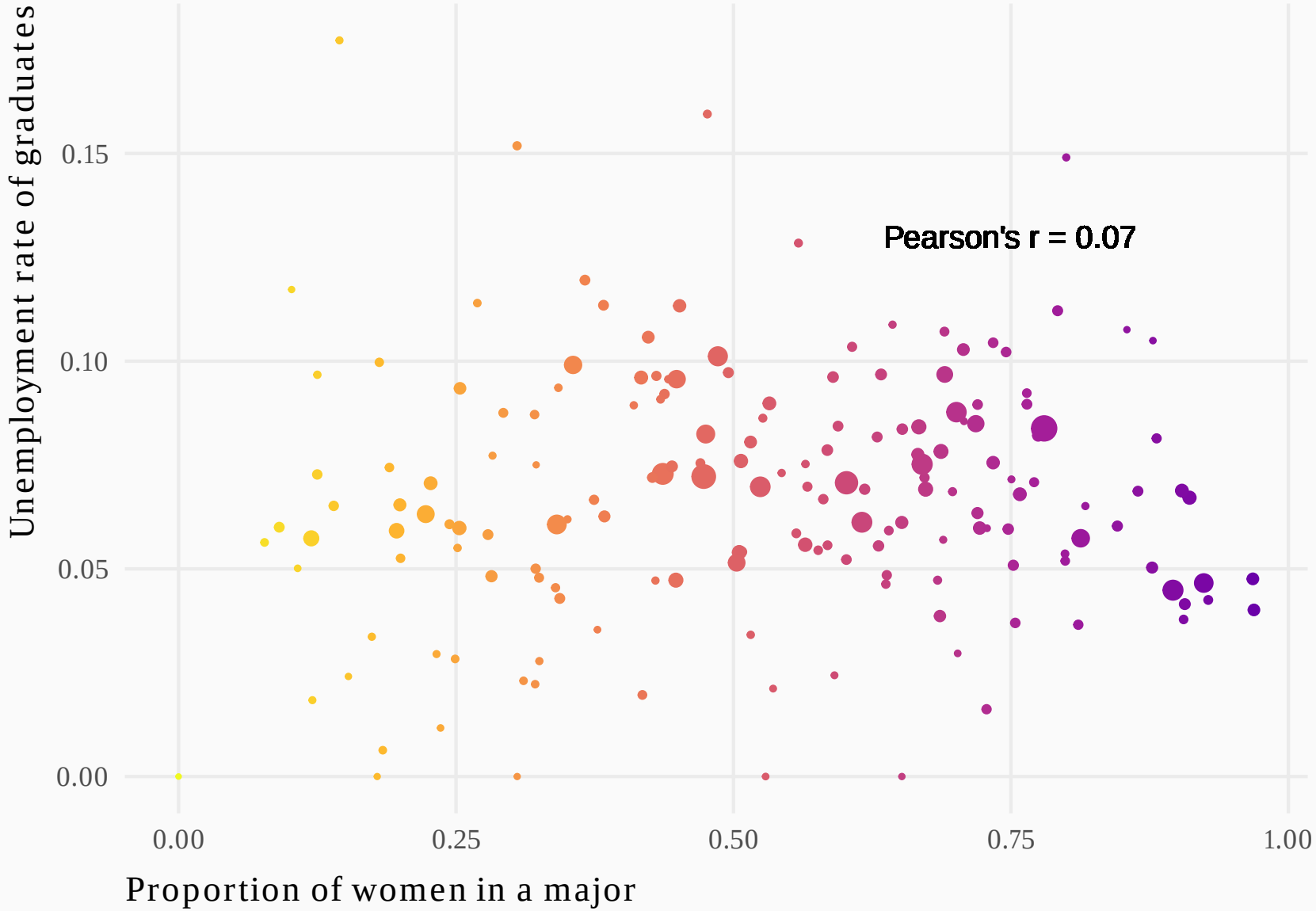
Is this a causal relationship?

- Maybe...maybe not!
- We could think of many **confounders**
  - A confounders is related to both X and Y
  - International economy, partisan control of Congress...

College majors: women and income



College majors: women and unemployment



College majors: correlation coefficients

```
cor(majors$ShareWomen, majors$Median,
    use = "pairwise")
```

```
## [1] -0.6186898
```

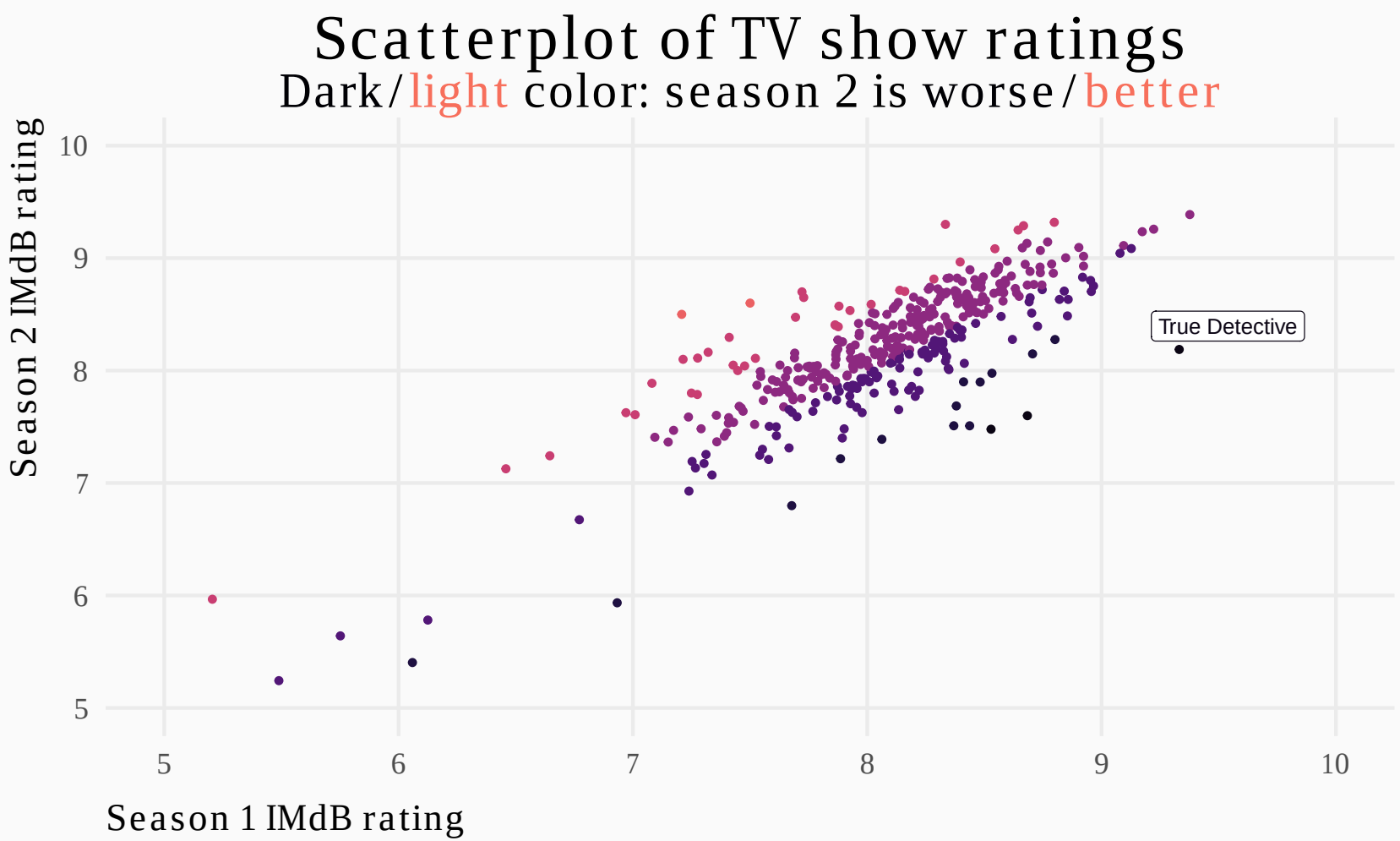
```
cor(majors$ShareWomen, majors$Unemployment_rate,
    use = "pairwise")
```

```
## [1] 0.07320458
```

`Share of women` and `Median salary`: a strong negative correlation

`Share of women` and `Unemployment`: basically no association

TV shows


Scatterplot of TV show ratings
Dark/light color: season 2 is worse/better
Season 2 IMdB rating
10
9
8
7
6
5
5
6
7
8
9
10
Season 1 IMdB rating
True Detective

TV shows

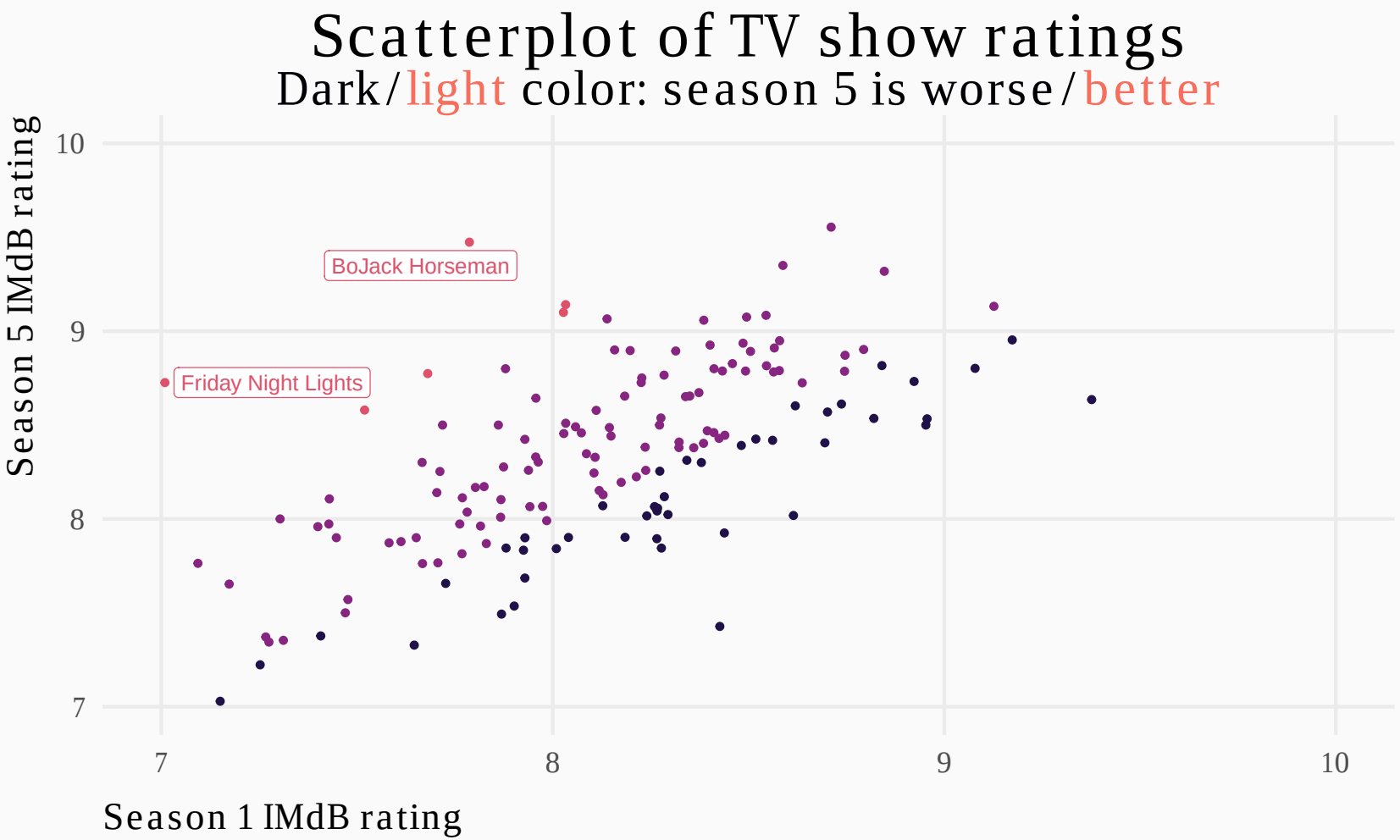
```
cor(show_level$`1`, show_level$`2`, use = "pairwise")
```

```
## [1] 0.8274108
```

Wow, that's a really strong correlation!

- How to interpret?
- Knowing how well-rated the first season is, you can make a very good guess as to the rating of the second season
- Do you think the relationship is as strong between season 1 and season 5?

Seasons 1 and 5



Seasons 1 and 5

```
cor(show_level_1_5$`1`, show_level_1_5$`5`, use = "pairwise")
```

```
## [1] 0.6334758
```

```
# you can change the order; doesn't matter
cor(show_level_1_5$`5`, show_level_1_5$`1`, use = "pairwise")
```

```
## [1] 0.6334758
```

The correlation is weaker, but still quite strong

- Scatterplots are very useful – **always** plot your data
- But must be careful in how you interpret them
- The scale for seasons 1 and 5 is different ⇝ correlation looks weaker than it is

Linearity

The correlation coefficient evaluates **linear** covariation

- What is a linear relationship?
- In response to a change in X, Y behaves in a particular way, no matter the value of X
- Non-linear relationship: the association between X and Y differs based on the value of X

Non-linearity: London Airbnb listings

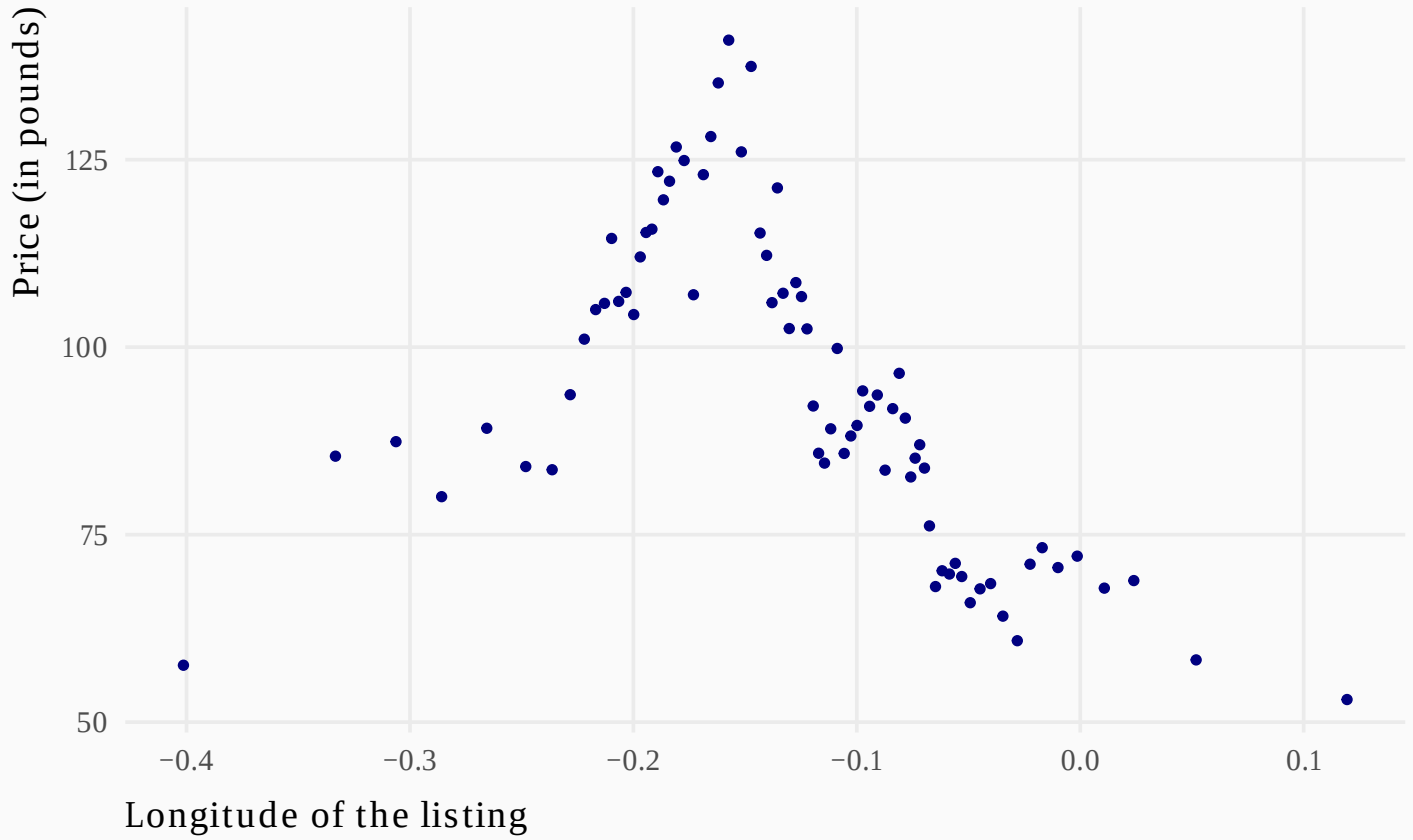


Figure 5: Longitude and price of London (UK) Airbnb listings on March 4th, 2017

```
cor(london$price, london$longitude, use = "pairwise")
```

```
## [1] -0.1262614
```

Equivalent relationships

Navigate to this link

- For all of these scatterplots, the summary stats are the same!
    - Same mean, same correlation, etc.
- And yet, looking at the scatterplots, the relationships are very different
- Always plot your data!
- Before doing any fancy statistics...
    - Look at the distribution of X
        - Do any cases stand out?
    - Look at the distribution of Y
        - Do any cases stand out?
    - Look at a scatterplot of X and Y
        - Do any cases stand out?

Not plotting your data? You might screw up

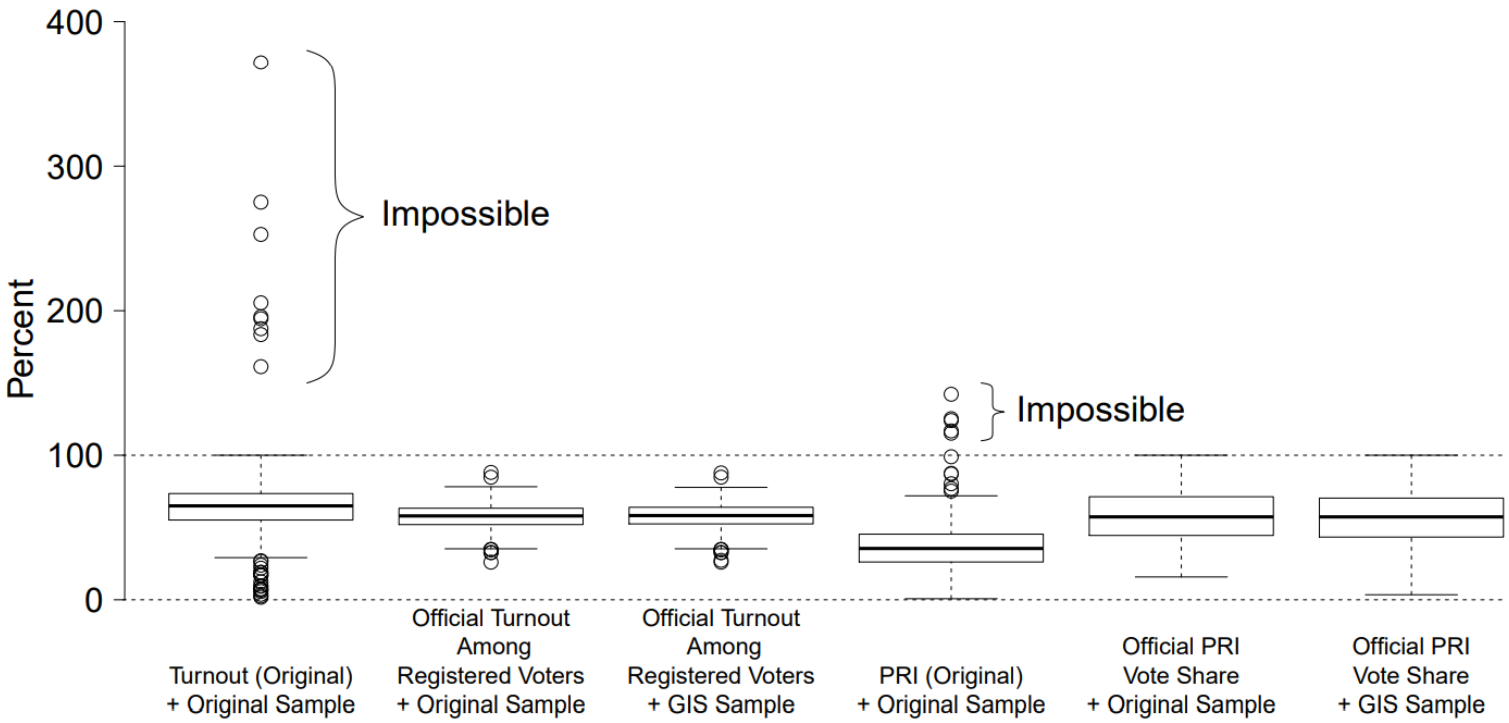
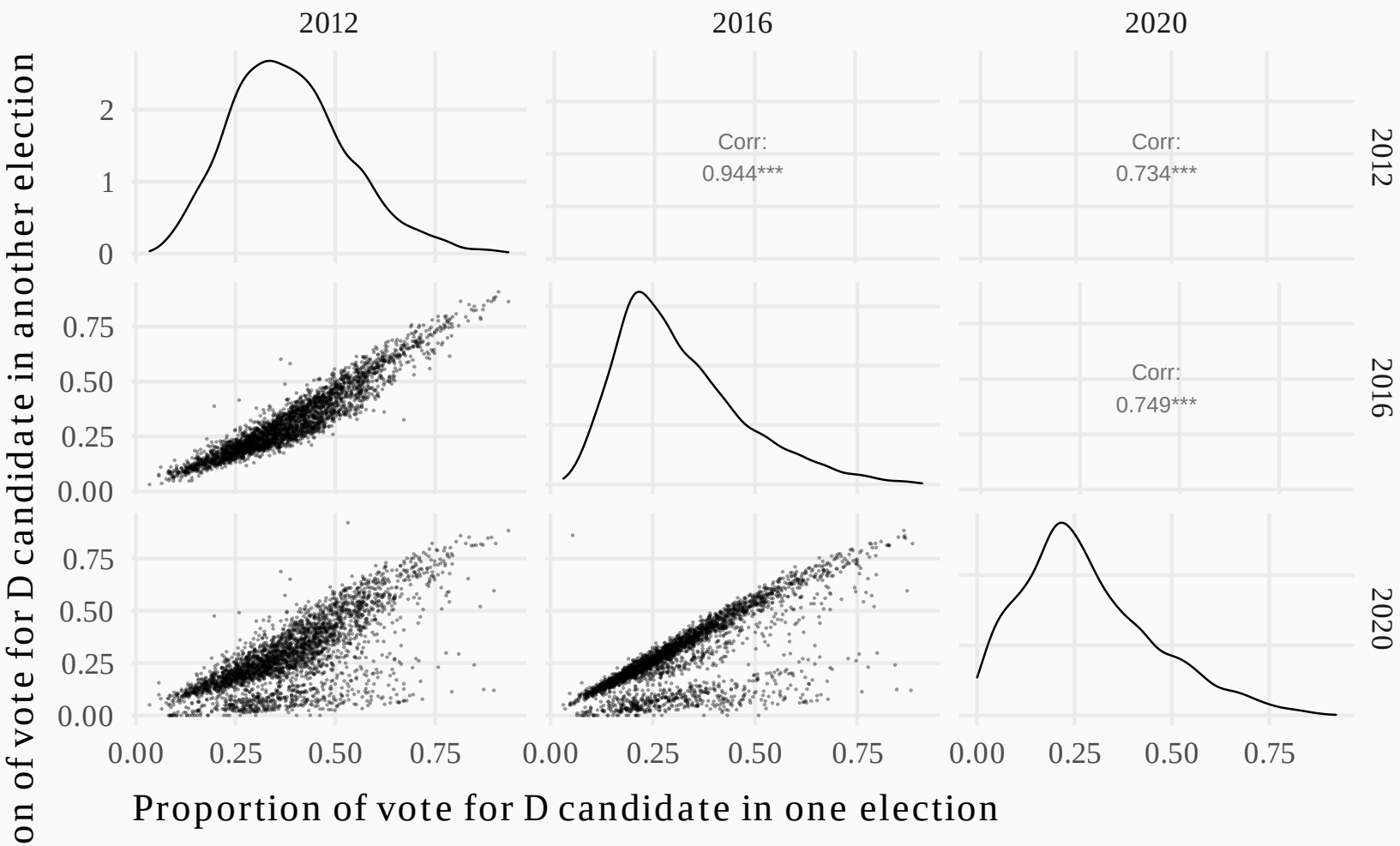


Figure 7: **Univariate Distribution of Turnout and Incumbent Party Vote in 2000.** This figure compares the variables originally constructed in De La O (2013) via name matching (in columns 1 and 4), with the official turnout among registered voters and PRI vote share in the name-matching sample (columns 2 and 5) and in the GIS sample (columns 3 and 6).

Scatterplot matrices



References i