Lecture Notes

Prof. Bisbee, Vanderbilt University

2023-02-08

QUIZ 8

Password: 1576

Then load the data

- `Pres2020_PV.Rds` is on GitHub

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.3.6     ✓ purrr   0.3.4
## ✓ tibble  3.1.7     ✓ dplyr   1.0.9
## ✓ tidyr   1.2.0     ✓ stringr 1.4.0
## ✓ readr   2.1.2     ✓ forcats 0.5.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
poll <- read_rds('https://github.com/jbisbee1/DS1000_S2023/blob/main/Lectures/4_Uni_Mult
ivariate/data/Pres2020_PV.Rds?raw=true')
```

Quick Wrangling

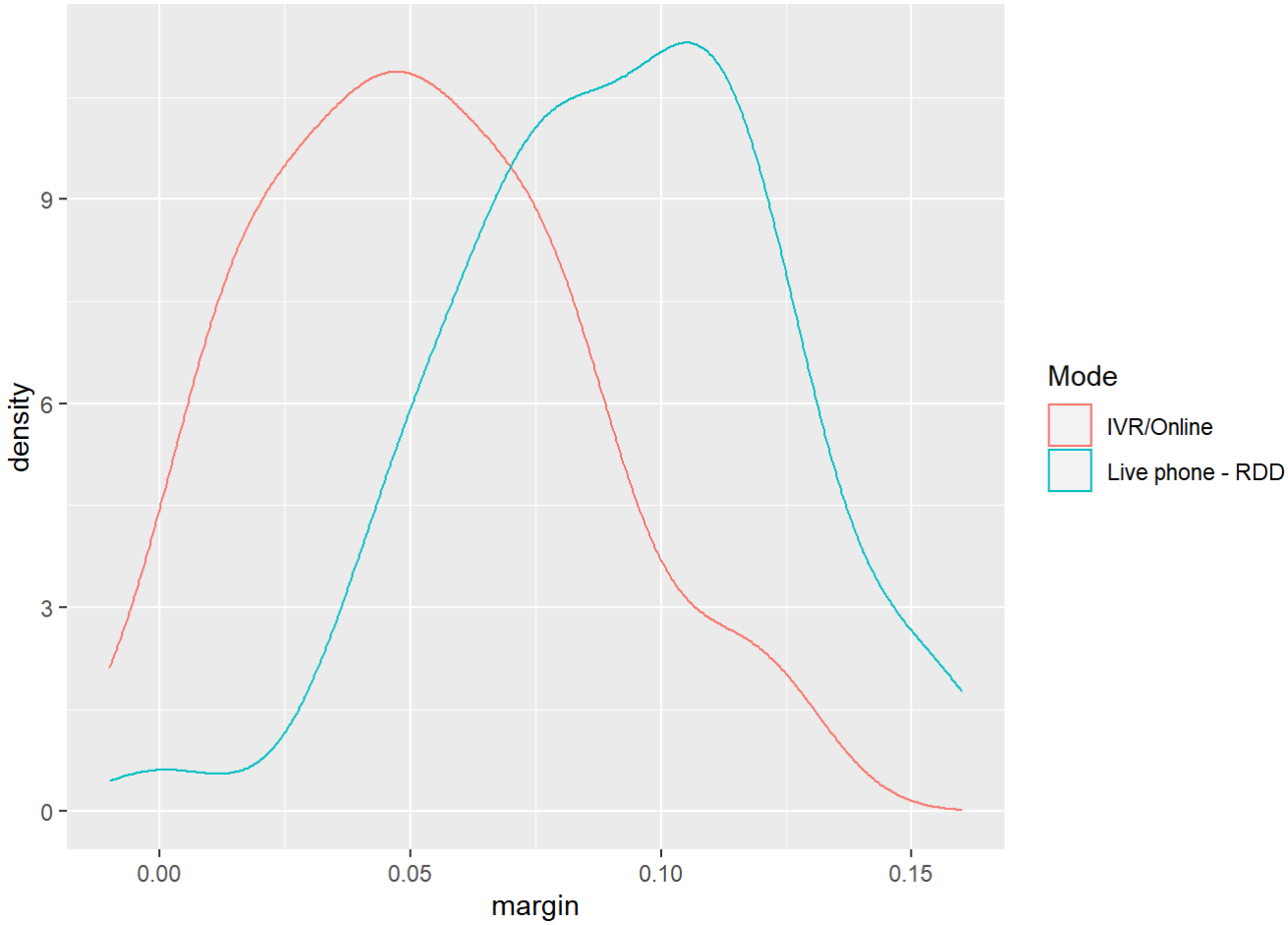
```
poll %>%
  select(Trump,Biden)
```

```
## # A tibble: 528 × 2
##    Trump Biden
##    <dbl> <dbl>
##  1    43    53
##  2    44    53
##  3    45    52
##  4    46    52
##  5    42    48
##  6    43    53
##  7    46    52
##  8    41    53
##  9    41    52
## 10    42    52
## # … with 518 more rows
```

```
poll <- poll %>%
  mutate(Trump = Trump/100,
         Biden = Biden/100,
         margin = round(Biden - Trump,2))
```

Bivariate Vis: Binary X Continuous

```
poll %>%
  filter(Mode == 'IVR/Online' | Mode == 'Live phone - RDD') %>%
  ggplot(aes(x = margin,color = Mode)) +
  geom_density()
```



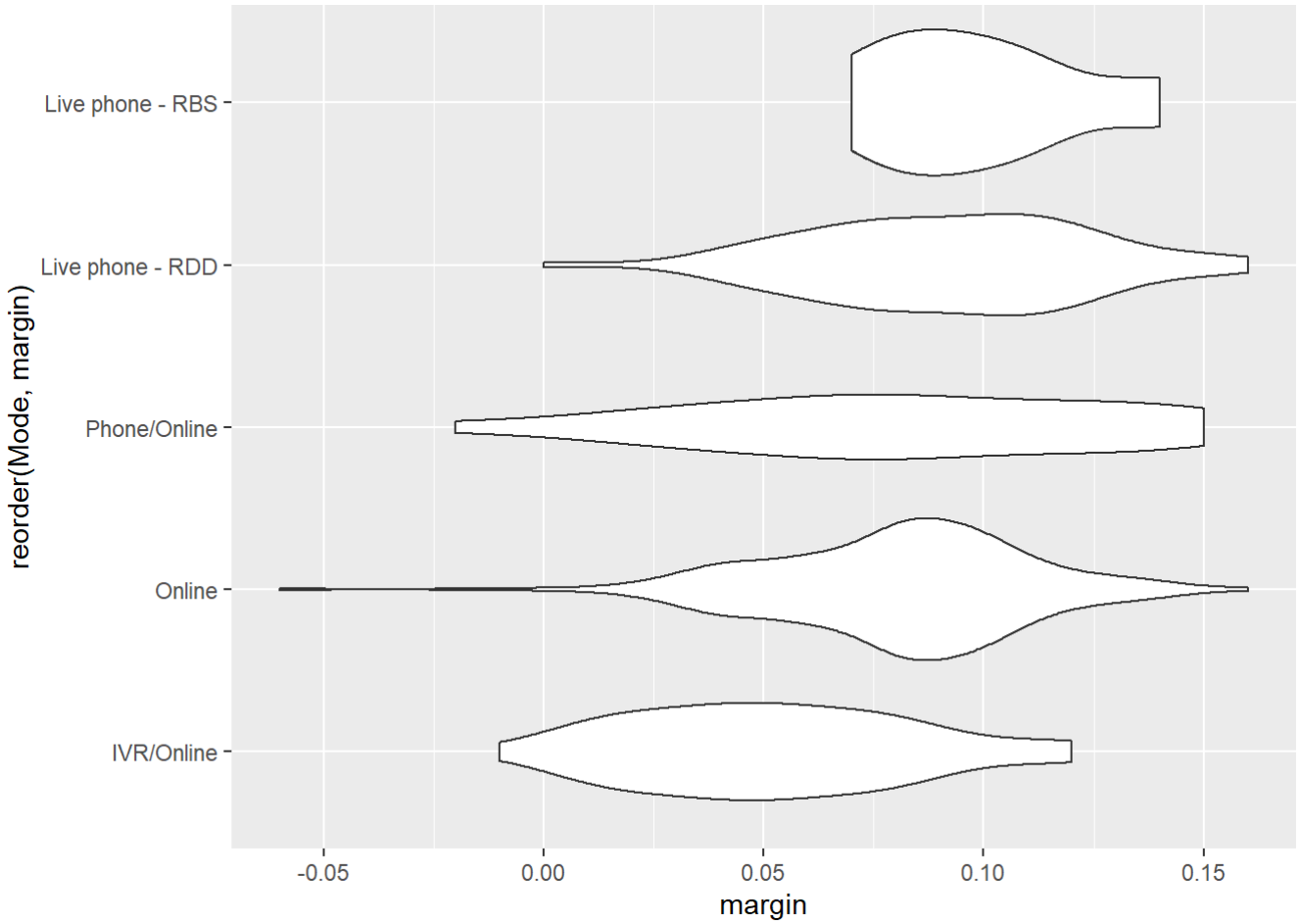

```
  # geom_histogram()
```

Bivariate Visualization Categorical X Continuous

```
poll %>%
  count(Mode)
```

```
## # A tibble: 9 × 2
##   Mode                 n
##   <chr>            <int>
## 1 IVR                  1
## 2 IVR/Online          47
## 3 Live phone - RBS    13
## 4 Live phone - RDD    51
## 5 Online             366
## 6 Online/Text          1
## 7 Phone - unknown      1
## 8 Phone/Online        19
## 9 <NA>                29
```

```
poll %>%
  filter(Mode != 'IVR' & Mode != 'Online/Text' & Mode != 'Phone - unknown') %>%
  ggplot(aes(x = margin,y = reorder(Mode,margin))) +
  geom_violin()
```

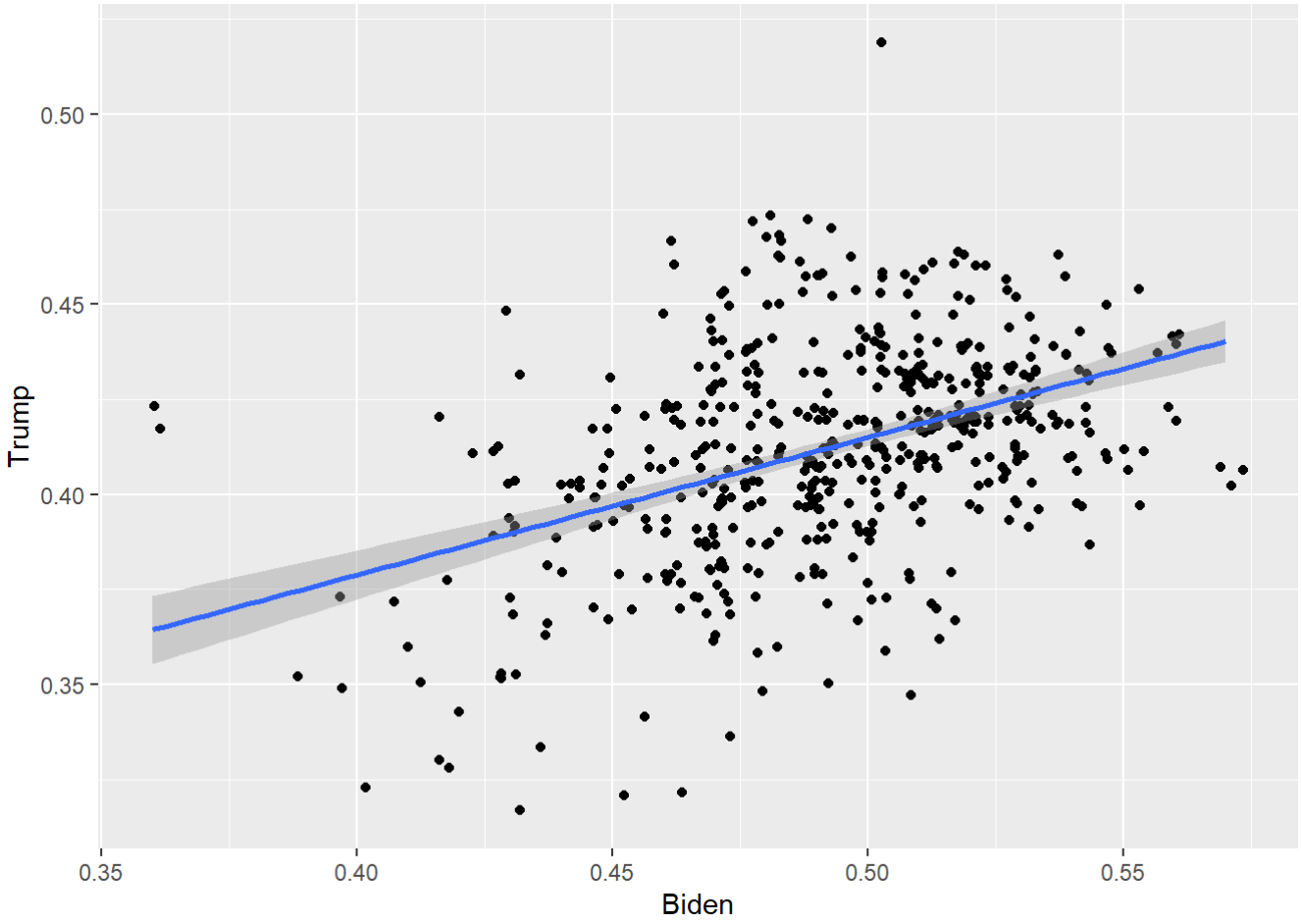



```
  # geom_boxplot()
```

Bivariate Visualization: Continuous X Continuous

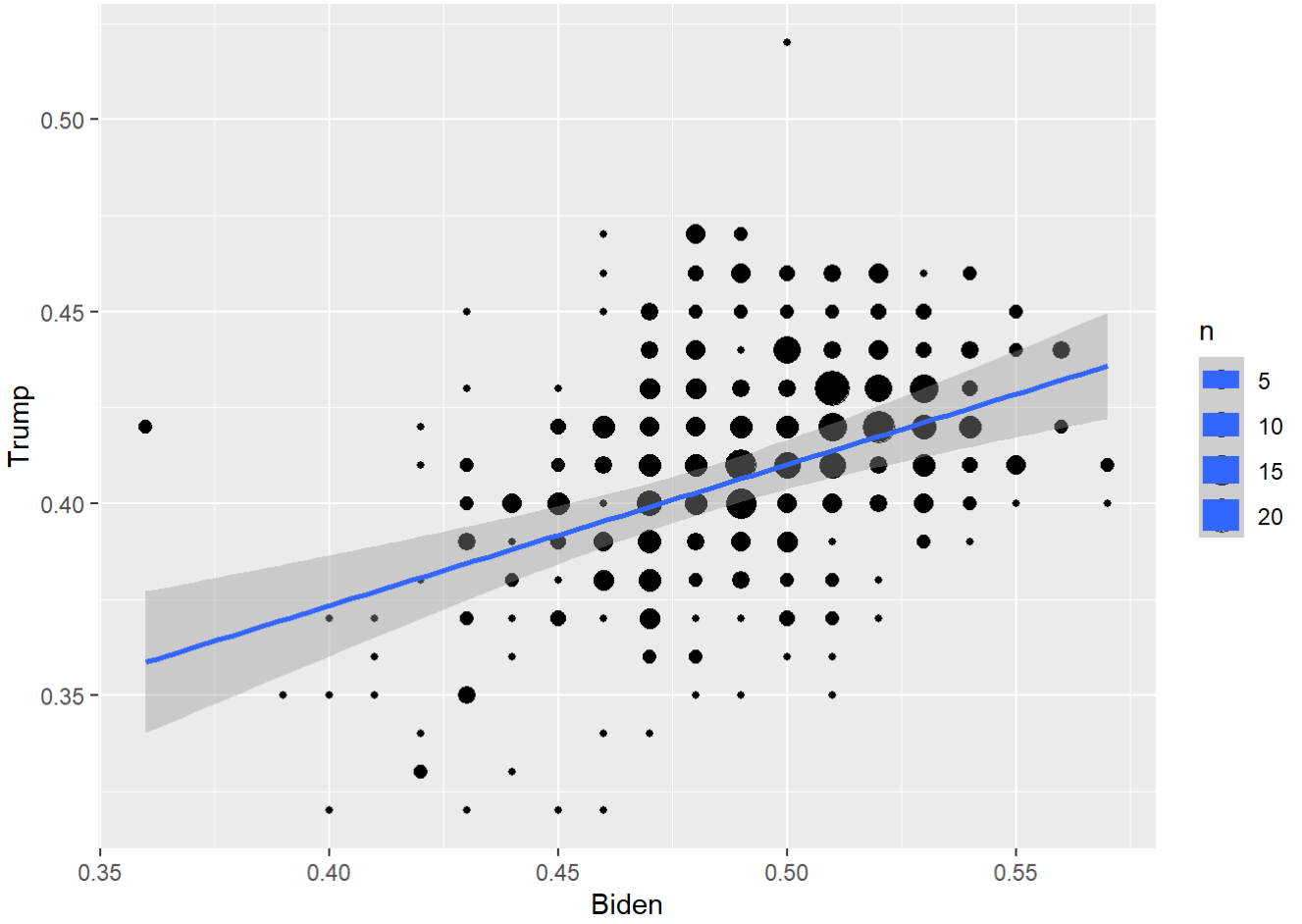
```
poll %>%
  ggplot(aes(x = Biden,y = Trump)) +
  # geom_point(alpha = .3) + # Approach #1 to revealing multiple data on the same point
  geom_jitter() + # Approach # 2 to revealining multiple data on the same point
  geom_smooth(method = 'lm')
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
poll %>%
  count(Biden,Trump) %>%
  ggplot(aes(x = Biden,y = Trump,size = n)) +
  geom_point() +
  geom_smooth(method = 'lm')
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Working with Dates

- `as.Date()` function converts characters to `date` class variables

```
d1 <- as.Date('02/08/2023',format = '%m/%d/%Y')
midterm <- as.Date('03/08/2023',format = '%m/%d/%Y')

midterm - d1
```

```
## Time difference of 28 days
```

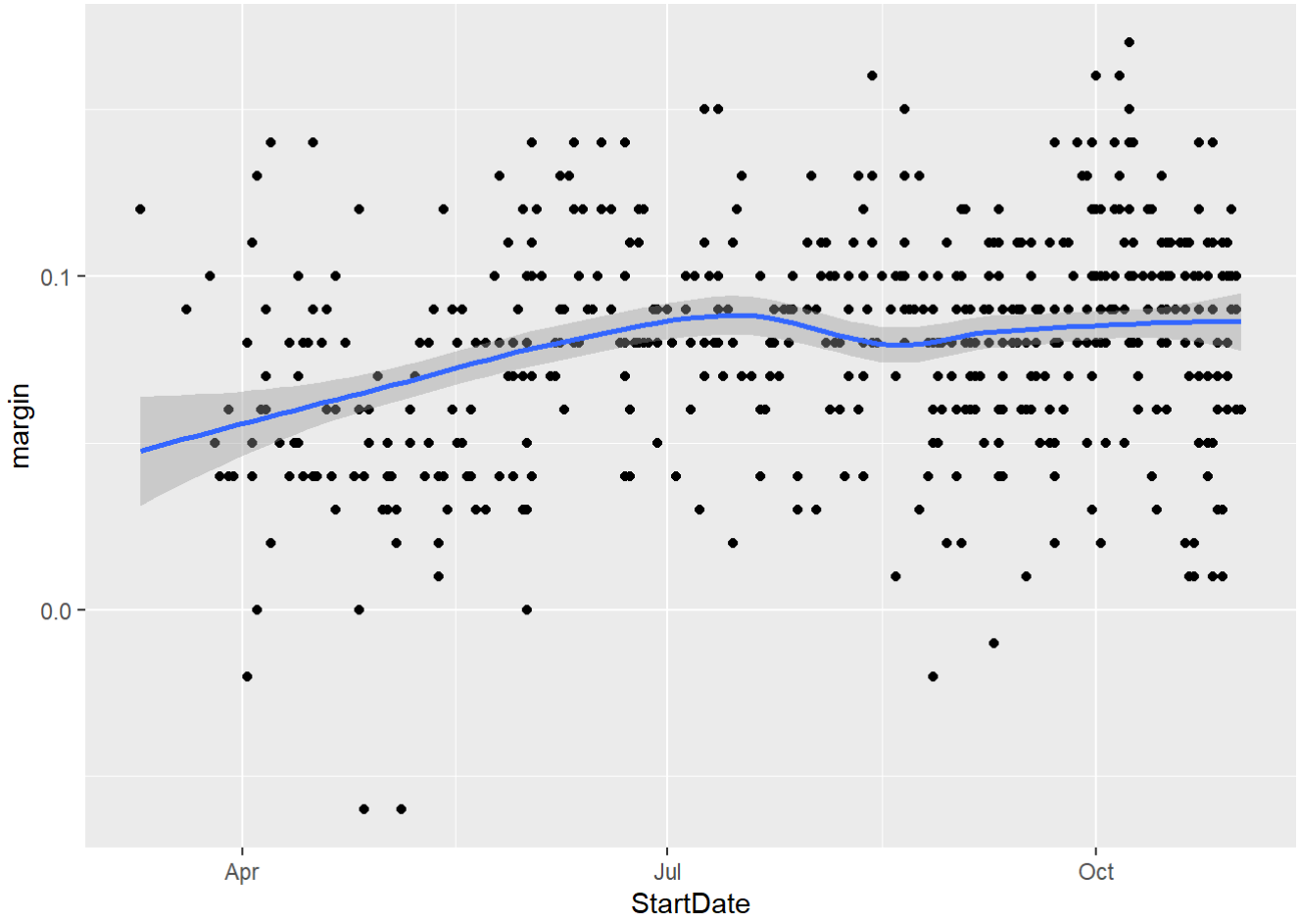
```
# Let's apply this to the poll data
poll <- poll %>%
  mutate(StartDate = as.Date(StartDate,format = '%m/%d/%Y'),
         EndDate = as.Date(EndDate,format = '%m/%d/%Y'))


poll %>% select(StartDate,EndDate)
```

```
## # A tibble: 528 × 2
##    StartDate  EndDate
##    <date>     <date>
##  1 2020-10-31 2020-11-02
##  2 2020-10-31 2020-11-02
##  3 2020-10-29 2020-11-02
##  4 2020-11-01 2020-11-01
##  5 2020-11-01 2020-11-01
##  6 2020-10-30 2020-11-01
##  7 2020-10-31 2020-11-02
##  8 2020-10-30 2020-11-01
##  9 2020-10-29 2020-11-01
## 10 2020-10-29 2020-11-01
## # … with 518 more rows
```

```
poll %>%
  ggplot(aes(x = StartDate,y = margin)) +
  # geom_bar(stat = 'identity')
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Look at each point individualls

```
poll %>%
  ggplot() +
  geom_point(aes(x = StartDate,y = Biden),color = 'blue') +
  geom_point(aes(x = StartDate,y = Trump),color = 'red')
```