GOING BEYOND R

DATA VISUALIZATION

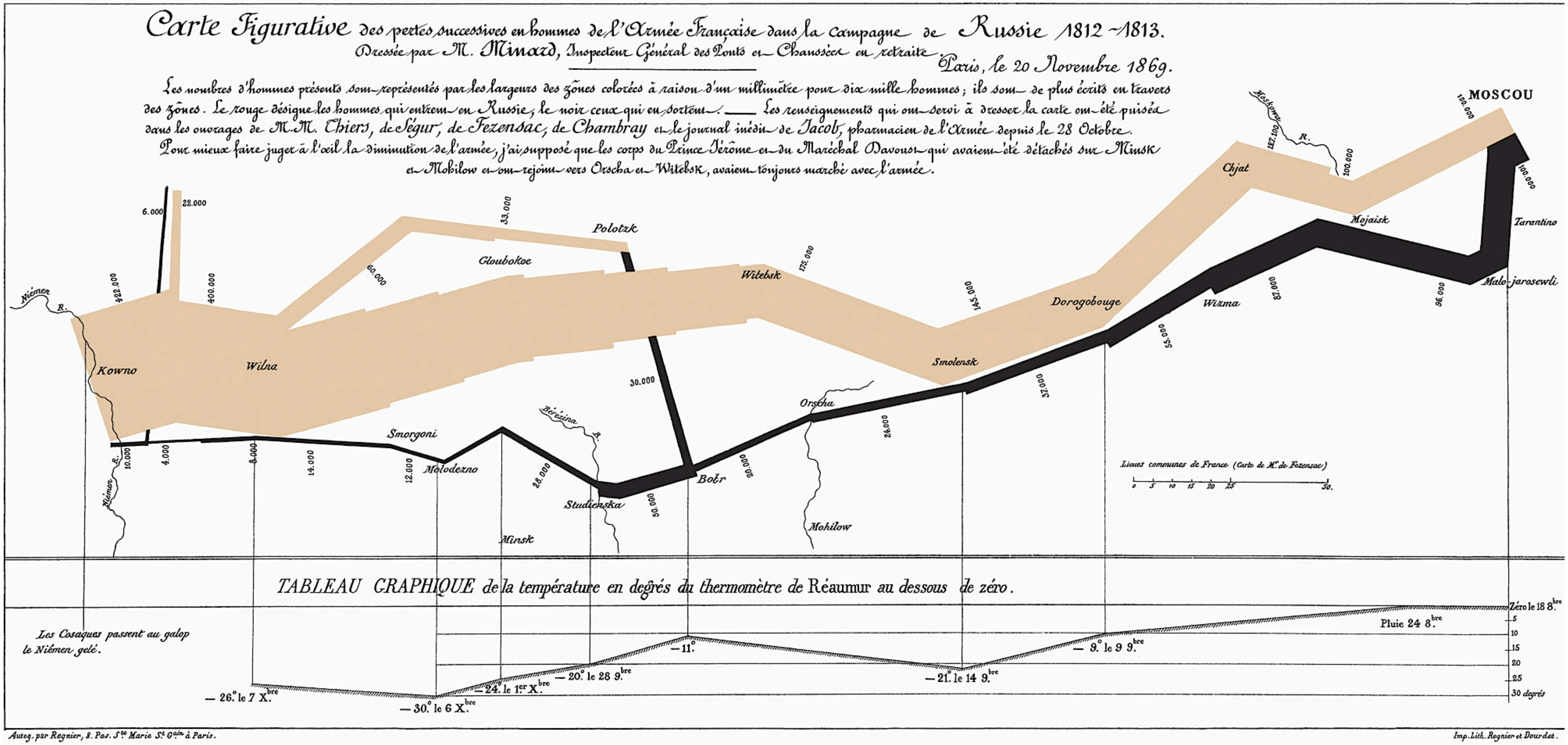
THANKS TO…

- This presentation draws liberally on examples from Edward Tufte's work on data visualization, Lee Wilkinson's work on the grammar of graphics, and some examples from my own work.
- Graphics are cited throughout, but due to space constraints, not all ideas are, though I try my best to use references for the core concepts.

WHAT'S THE PURPOSE OF VISUALIZING OUR DATA?

- Humans are highly visual learners and communicators
  - We can often recognize in a picture what would take months of advanced training to understand statistically or textually
  - Remember the adage "A picture is worth a thousand words?"
    - **Your visuals should be worth a lot more than that**

TUFTE: “THE MOST INFORMATIVE DATA VISUALIZATION EVER”



TWO PURPOSES

- There are two major reasons we want to visualize our data:
  - Exploratory data analysis (getting to know our own data)
  - Story-telling (helping others get to know our data)
- **Using graphics to explore your own data is a best practice**
  - *In <u>quantitative</u> analysis*, focus on statistical significance can blind us to the real findings, and can cause inadvertent "p-hacking"
  - *In <u>qualitative</u> analysis*, the huge scope of the data (and often inductive nature of the research) can lead to a similar phenomenon: "cherry-picking" of particular quotations or events that support one's arguments

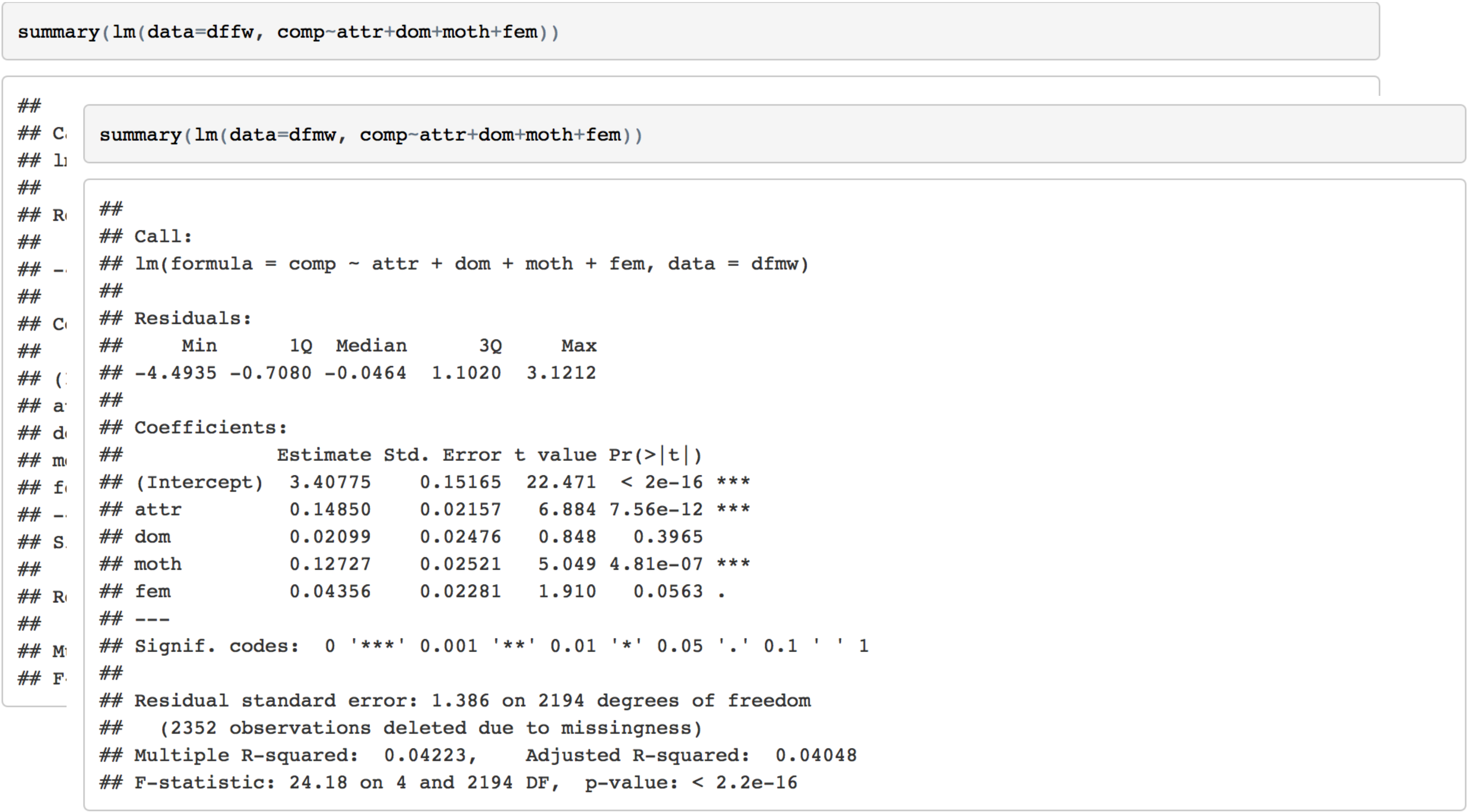
EXAMPLE (1)

EXAMPLE (1)

```
summary(lm(data=dffw, comp~attr+dom+moth+fem))
```

```
##
## Call:
## lm(formula = comp ~ attr + dom + moth + fem, data = dffw)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5720 -0.7787  0.0671  0.9869  3.5191
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.94855    0.17546  16.805  < 2e-16 ***
## attr         0.12764    0.02070   6.165 8.36e-10 ***
## dom          0.02084    0.02454   0.849    0.396
## moth         0.17507    0.02644   6.622 4.44e-11 ***
## fem          0.12545    0.02363   5.308 1.22e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.345 on 2191 degrees of freedom
##   (2355 observations deleted due to missingness)
## Multiple R-squared:  0.05982,    Adjusted R-squared:  0.05811
## F-statistic: 34.85 on 4 and 2191 DF,  p-value: < 2.2e-16
```

EXAMPLE (1)



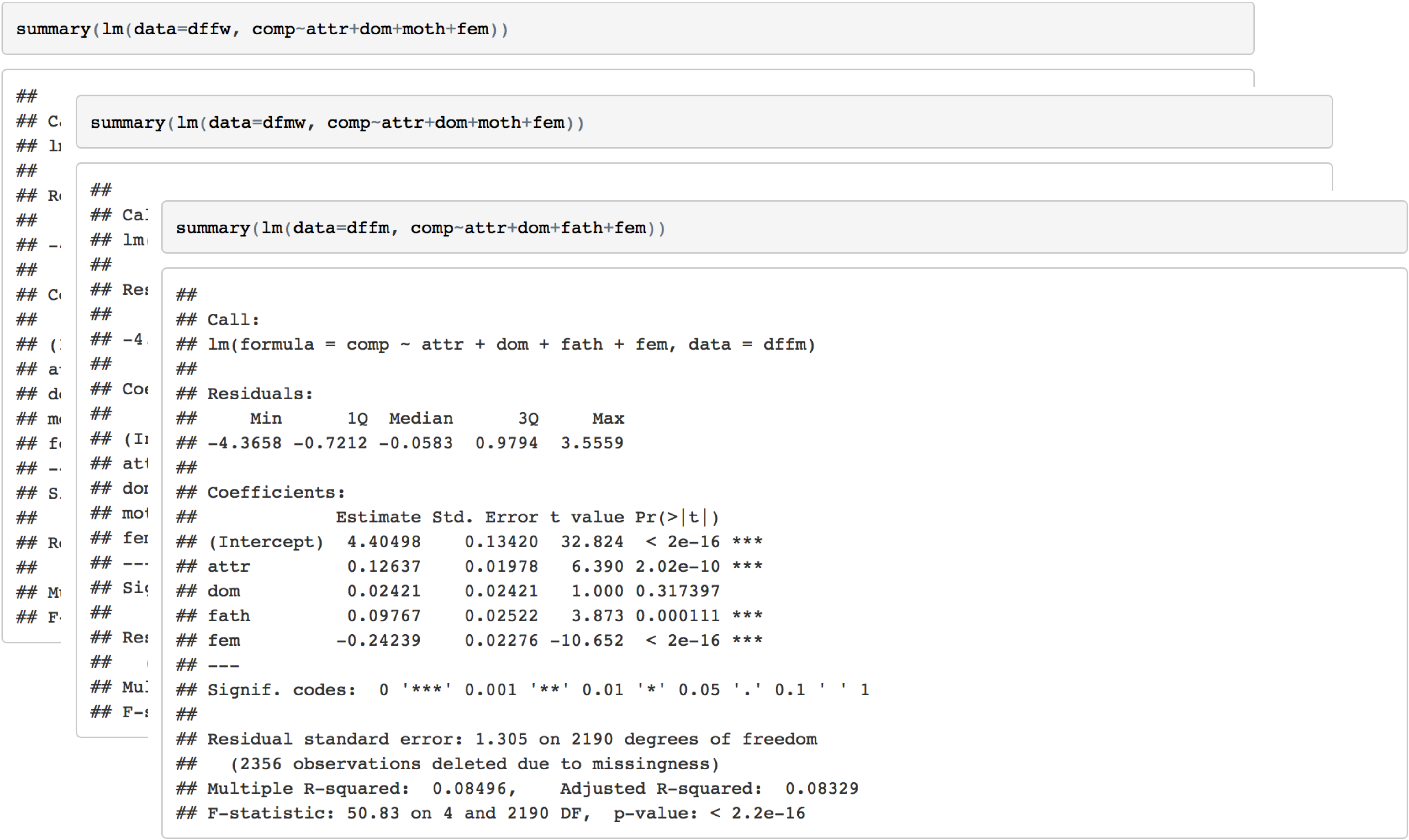
```
summary(lm(data=dffw, comp~attr+dom+moth+fem))
```

```
##
## C
## l
##
## R
##
## -
##
## C
##
## (
## a
## d
## m
## f
## -
## S
##
## R
##
## M
## F
```

```
summary(lm(data=dfmw, comp~attr+dom+moth+fem))
```

```
##
## Call:
## lm(formula = comp ~ attr + dom + moth + fem, data = dfmw)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4935 -0.7080 -0.0464  1.1020  3.1212
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.40775    0.15165  22.471  < 2e-16 ***
## attr         0.14850    0.02157   6.884 7.56e-12 ***
## dom          0.02099    0.02476   0.848   0.3965
## moth         0.12727    0.02521   5.049 4.81e-07 ***
## fem          0.04356    0.02281   1.910   0.0563 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.386 on 2194 degrees of freedom
##   (2352 observations deleted due to missingness)
## Multiple R-squared:  0.04223,    Adjusted R-squared:  0.04048
## F-statistic: 24.18 on 4 and 2194 DF,  p-value: < 2.2e-16
```

EXAMPLE (1)



```
summary(lm(data=dffw, comp~attr+dom+moth+fem))
```

```
##
## C
## l
##
## R
##
## -
##
## C
##
## (
## a
## d
## m
## f
## -
## S
##
## R
##
## M
## F
```

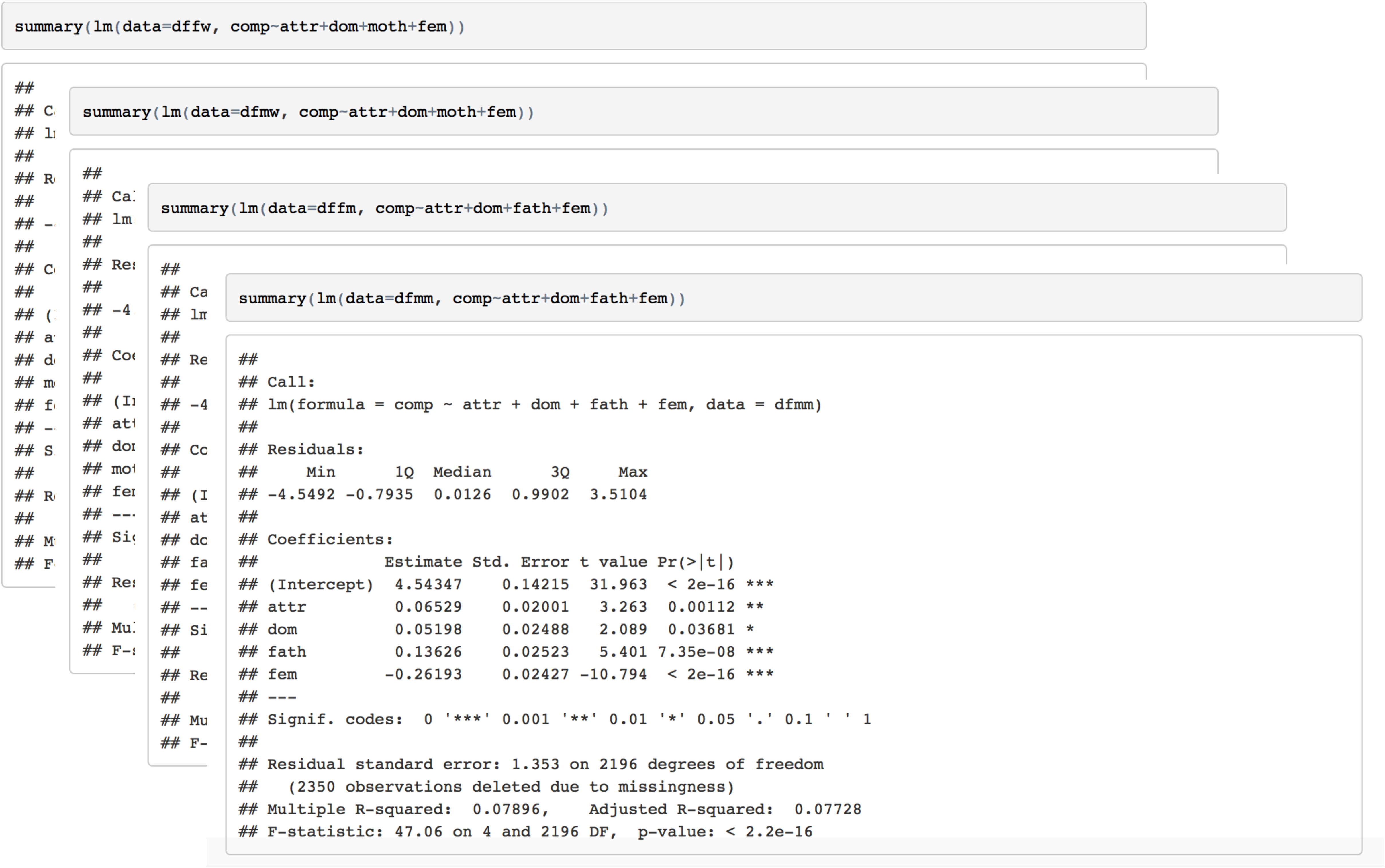
```
summary(lm(data=dfmw, comp~attr+dom+moth+fem))
```

```
##
## Ca
## lm
##
## Re
##
## -4
##
## Coe
##
## (I
## att
## dom
## mot
## fem
## ---
## Sig
##
## Re
##
## Mu
## F-
```

```
summary(lm(data=dffm, comp~attr+dom+fath+fem))
```

```
##
## Call:
## lm(formula = comp ~ attr + dom + fath + fem, data = dffm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3658 -0.7212 -0.0583  0.9794  3.5559
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.40498    0.13420  32.824  < 2e-16 ***
## attr         0.12637    0.01978   6.390 2.02e-10 ***
## dom          0.02421    0.02421   1.000 0.317397
## fath         0.09767    0.02522   3.873 0.000111 ***
## fem         -0.24239    0.02276 -10.652  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.305 on 2190 degrees of freedom
##   (2356 observations deleted due to missingness)
## Multiple R-squared:  0.08496,    Adjusted R-squared:  0.08329
## F-statistic: 50.83 on 4 and 2190 DF,  p-value: < 2.2e-16
```

EXAMPLE (1)



```
summary(lm(data=dffw, comp~attr+dom+moth+fem))
```

```
##
## C
## l
##
## R
##
## -
##
## C
##
## (
## a
## d
## m
## f
## -
## S
##
## R
##
## M
## F
```

```
summary(lm(data=dfmw, comp~attr+dom+moth+fem))
```

```
##
## Ca
## lm
##
## Re
##
## -4
##
## Co
##
## (I
## at
## do
## mo
## fe
## --
## Si
##
## Re
##
## Mu
## F-
```

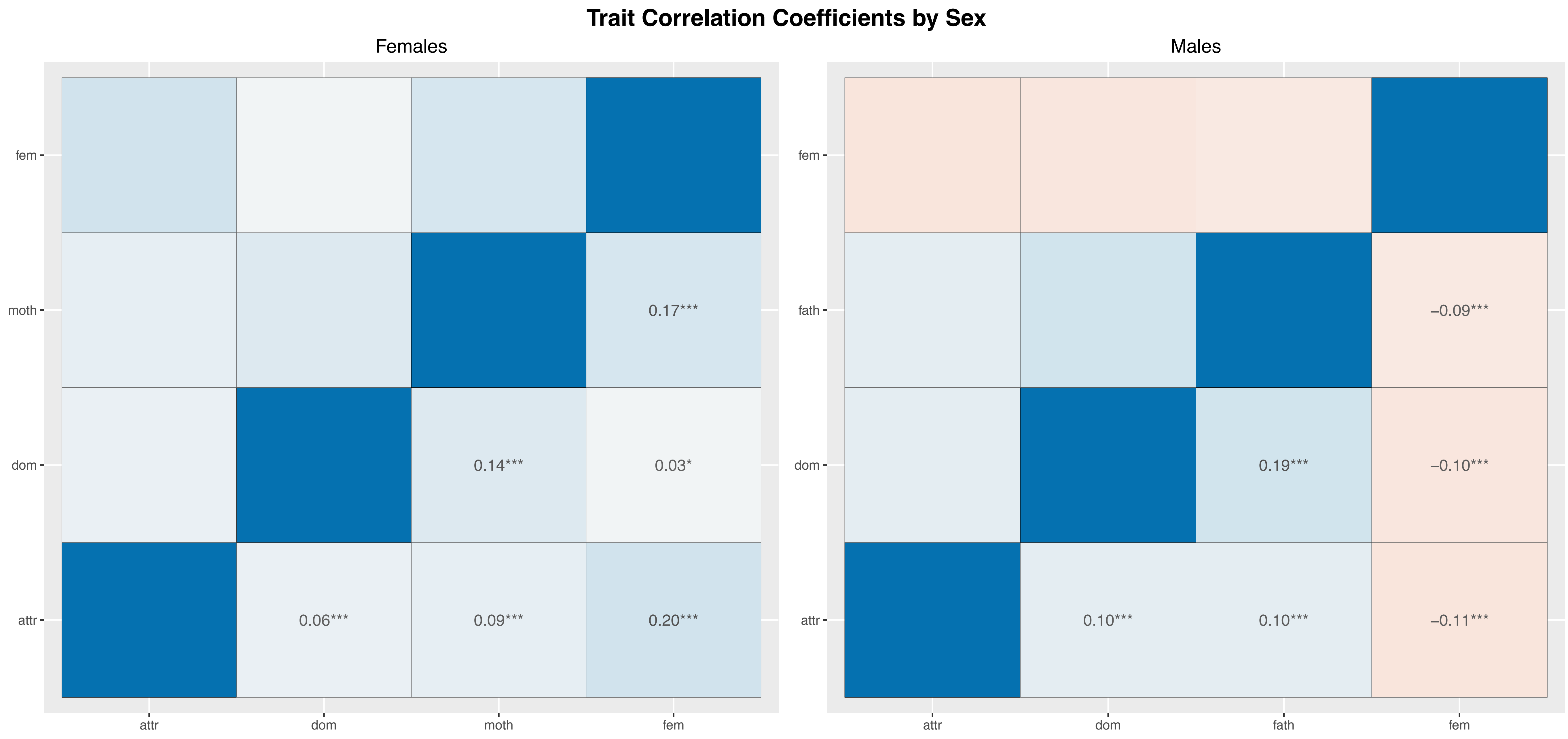
```
summary(lm(data=dffm, comp~attr+dom+fath+fem))
```

```
##
## Ca
## lm
##
## Re
##
## -4
##
## Co
##
## (I
## at
## do
## fa
## fe
## --
## Si
##
## Re
##
## Mu
## F-
```

```
summary(lm(data=dfmm, comp~attr+dom+fath+fem))
```

```
##
## Call:
## lm(formula = comp ~ attr + dom + fath + fem, data = dfmm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5492 -0.7935  0.0126  0.9902  3.5104
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.54347    0.14215  31.963  < 2e-16 ***
## attr         0.06529    0.02001   3.263  0.00112 **
## dom          0.05198    0.02488   2.089  0.03681 *
## fath         0.13626    0.02523   5.401 7.35e-08 ***
## fem         -0.26193    0.02427 -10.794  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.353 on 2196 degrees of freedom
##   (2350 observations deleted due to missingness)
## Multiple R-squared:  0.07896,    Adjusted R-squared:  0.07728
## F-statistic: 47.06 on 4 and 2196 DF,  p-value: < 2.2e-16
```

EXAMPLE (1)


Trait Correlation Coefficients by Sex
Females
fem
moth
dom
attr
0.17***
0.14***
0.03*
0.06***
0.09***
0.20***
attr
dom
moth
fem
Males
fem
fath
dom
attr
−0.09***
0.19***
−0.10***
0.10***
0.10***
−0.11***
attr
dom
fath
fem

EXAMPLE (2)

Afraid of my Debt, and not having money to live or someone to take care of my daughter

Mostly because of negative experiences within the local Democratic Party. Also, the office I was originally interested in (School Board) is less attractive and unpaid and the other position I am interested in (County Commissioner) is currently held by someone I admire and who does great work. /

(1) Pay for City Council (contemplated office) too low to do as only job, but job too large to do well on part-time basis (esp. with children still at home) / (2) Disdain for ratio of fundraising/political shmoozing to engagement with actual policy issues

1. Friendly incumbents in office / 2. I hold elected positions in my local, regional, state, and national union. So, my time is spread pretty thin. / / However, I hope that when an opening develops I will be able to stop some of my other work and run for office.

1. Haven't had the right opportunity in my districts yet. / / 2. Need to be more financially stable

1. It's a blood sport where I live and the process of running feels too daunting. 2. I like to have some time for myself and with my kids, spouse and friends. 3. While I think I would be a good legislator, my current job (in policy advocacy) allows me to fight for the things I care about without having to horse trade with people I don't respect.

1. Time Commitment. I don't feel like I can give the time my potential constituents deserve given that I work full time in a city 60+ miles from where I live. I'm the primary household earner. I have to work. Also concerns about giving my family the time it deserves. / 2. Nastiness of politics. I'm not confident I have thick enough skin. People are awful to elected officials regardless of whether the elected official deserves it.

1. Too intrusive in my personal life. / 2. Too much focus on fundraising. / 3. Am not 100% aligned with Democratic Party platform (I'm more liberal)

Afraid of the personal scrutiny, dislike fundraising and not good at answering questions on the fly. Also hard to keep integrity when so much compromise is required.

After participating in the Emerge Program I better understand the complexities of fundraising. I live in a big city and the money needed is enormous and I don't have name recognition yet.

At 27 years old, I am working to establish myself in the community and build connections. I am also working to gain political experience through Emerge and working with active campaigns. In two years, I plan on running and feel I will have the experience and name recognition needed to run at that point.

At 28, I've considered myself too young to seriously run for office and have dedicated myself to learning the nuts and bolts of both campaigns and public policy by working on campaigns or for the government.

Because I am 23-years-old

Because I don't want my past, private, and personal life in the spotlight. I also don't subscribe to the democratic party nor my union status quo. It's too constricting and short sided. And the fundraising portion is horrible. If there were someone else in charge of the funding--it would be a lot easier.

Because I had a baby the year i would have probably run for office and now am not sure if i plan to stay in Arizona long term. /

Because I was turned off by the amount of self I felt like I had to sacrifice. I was also discouraged by how difficult it is to break into city politics. Finally, I would like to have impact on issues that are larger that street pavement and light repair.

Because the Democratic party does not seek or encourage women with disabilities to run. Because when my name is floated the first question is always "Is she healthy enough" not "is she qualified."

Because the seat that I want is occupied by a party leader/fellow Emerge alum. I have been pushed to run for other offices but I find them less of a fit with my interests, skills and schedule.

Been building professional career

Been in medical school so don't have time yet.

Busy raising a family and working. I live in a rural, Republican dominated district so a Dem is unlikely to win. I can only run for county or statewide office.

Can't balance family with legislative session, not interested in any local office

Childcare

Children too young, financial not yet independent. But am transitioning into being ED of the non-profit I founded

Concerned about the economic impact on my family as I am the primary wage earner. Secondary concern at having my personal life or private medical decisions publicly trashed by negative campaigning.

Conflict with my job.

current democrat will not step down from office held for 30 years

Currently running.

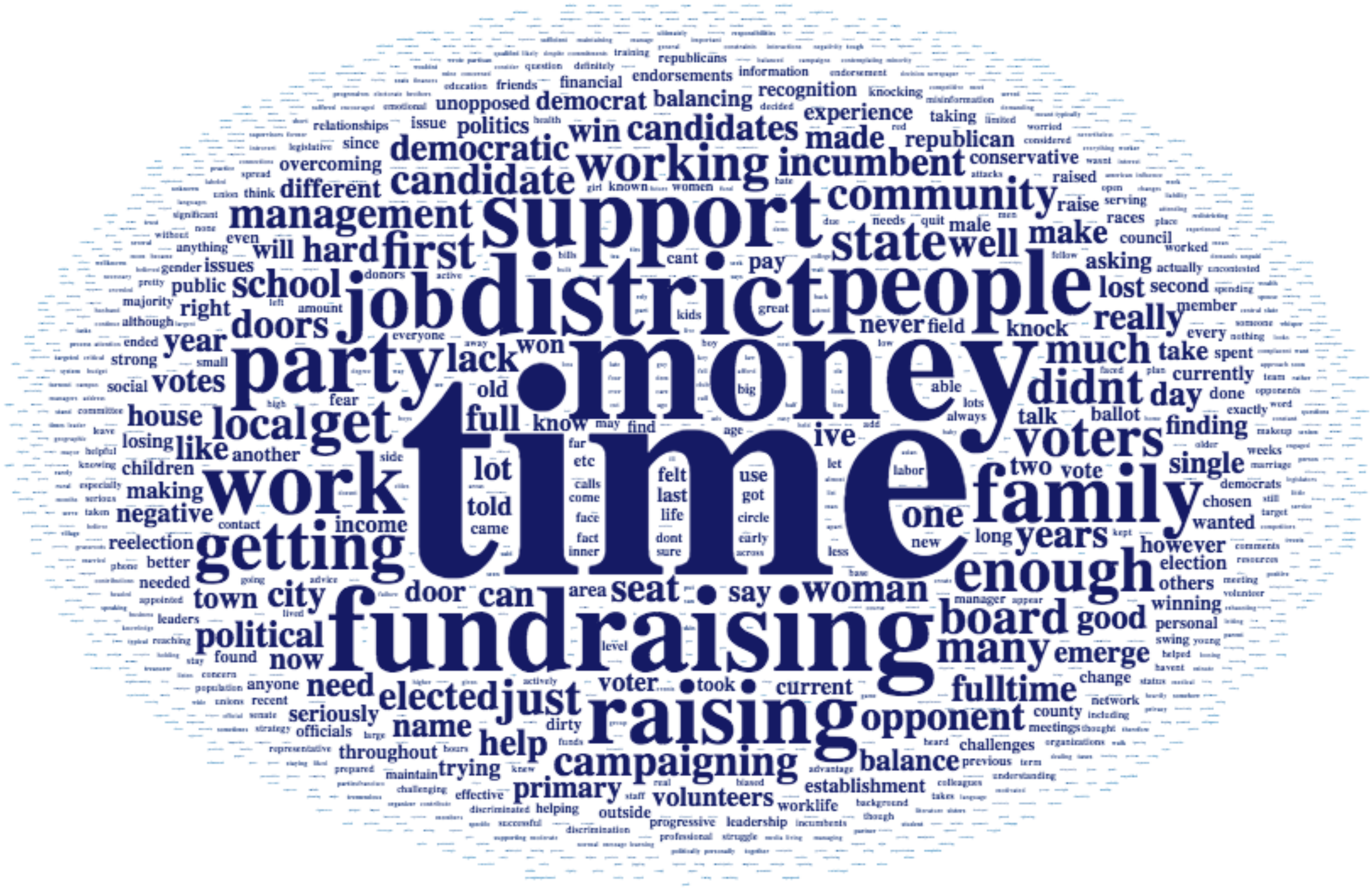
Decided not for me

Decided to dedicate more time to my family.

Determined that I'm better behind the scenes and don't want to run. Also DC does not have actual voting rights and the City Council is a mess.

Didn't think I had what it takes to run for office. Just finished emerge and there were no positions for me to run for during this election cycle.

EXAMPLE (2)

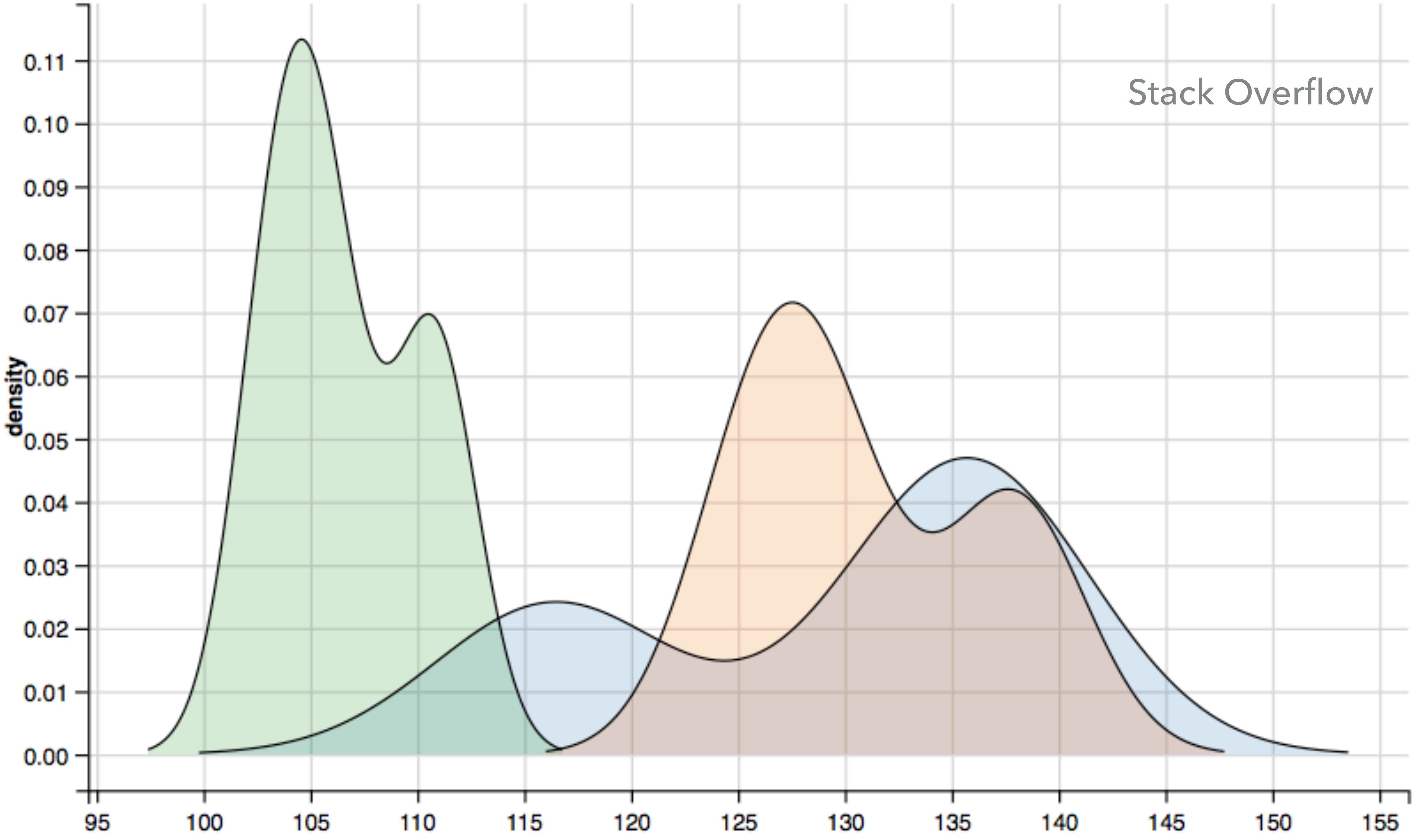

training
republicans
information
endorsements
financial
endorsement
question
definitely
education
friends
recognition
knocking
balancing
democrat
unopposed
emotional
experience
taking
relationships
issue
politics
win
candidates
made
republican
since
democratic
working
incumbent
conservative
overcoming
raised
think
different
candidate
known
women
community
raise
serving
management
support
needs
quit
male
races
even
will
hard
first
state
well
make
council
worked
anything
gender
issues
cant
pay
asking
actually
public
school
donors
active
lost
second
majority
right
amount
job
district
people
really
member
although
doors
kids
great
never
field
knock
ended
year
everyone
won
much
every
strong
small
party
lack
money
take
spent
social
votes
old
didnt
currently
day
done
fear
able
lots
exactly
house
local
get
full
know
may
find
always
talk
ballot
finding
losing
like
another
time
age
ive
voters
far
single
children
work
side
lot
etc
felt
use
labor
two
vote
democrats
making
calls
last
got
family
chosen
negative
told
come
life
one
wanted
income
contact
face
circle
reelection
getting
came
fact
dont
early
long
years
however
inner
sure
across
less
new
better
enough
election
needed
advice
others
town
city
door
can
area
seat
say
woman
manager
appear
winning
leaders
political
board
good
personal
found
fundraising
swing
now
level
many
emerge
helped
anyone
need
voter
took
current
fulltime
change
recent
elected
just
network
seriously
raising
opponent
county
officials
name
dirty
including
meetings
thought
help
funds
throughout
hours
beard
challenges
trying
campaigning
balance
previous
term
maintain
knew
advantage
primary
establishment
effective
volunteers
worklife
background
helping
outside
discrimination
progressive
leadership
incumbents
professional
struggle

HOW SHOULD WE VISUALIZE A GIVEN STORY?

- First, start with your data
  - Univariate analyses
  - Bivariate analyses
  - Multivariate analyses
  - Qualitative/other

UNIVARIATE ANALYSES

- Histograms, frequency plots, density plots, bar charts



Stack Overflow

UNIVARIATE ANALYSES

- Bar charts should be reserved for categorical data only

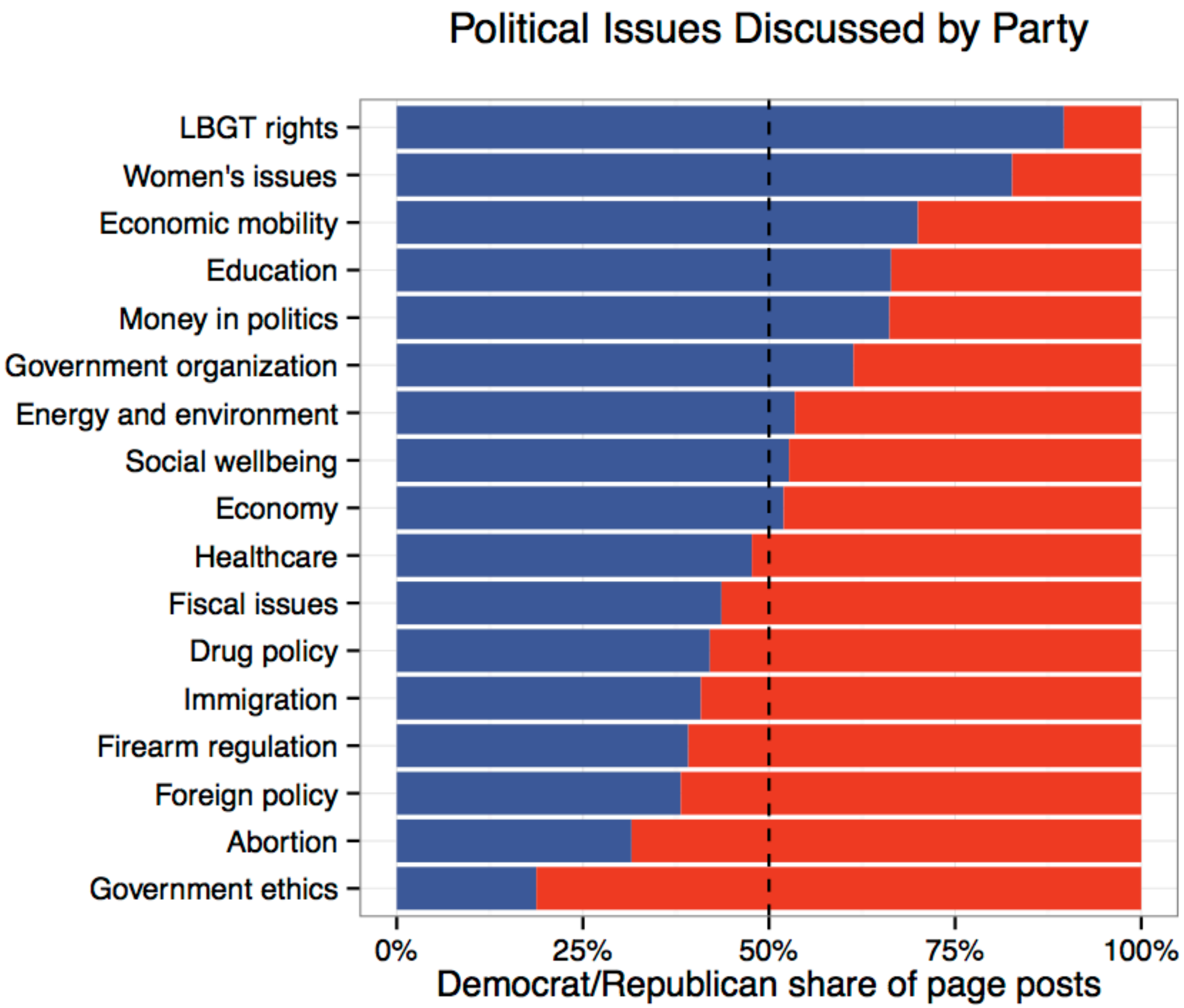
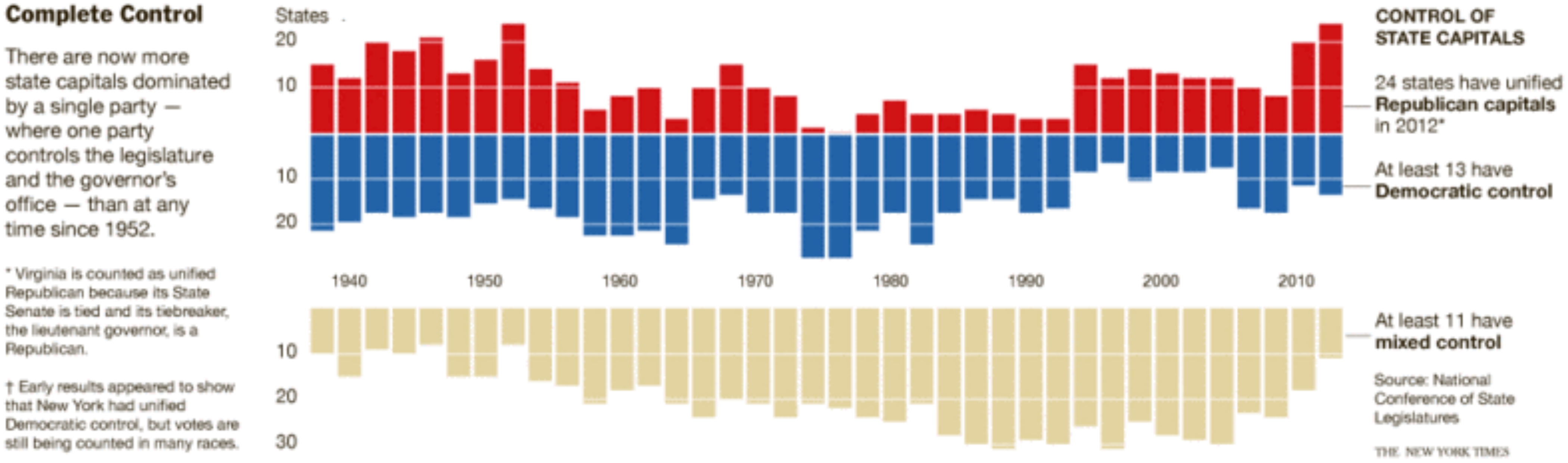


Chart from Solomon Messing's Wordpress

BIVARIATE ANALYSES

- Stacked bar, box plot, coefficient plot, scatterplot, line or multi-line plot, etc.



Complete Control

There are now more state capitals dominated by a single party — where one party controls the legislature and the governor's office — than at any time since 1952.

* Virginia is counted as unified Republican because its State Senate is tied and its tiebreaker, the lieutenant governor, is a Republican.

† Early results appeared to show that New York had unified Democratic control, but votes are still being counted in many races.

CONTROL OF STATE CAPITALS

24 states have unified **Republican capitals** in 2012*

At least 13 have **Democratic control**

At least 11 have **mixed control**

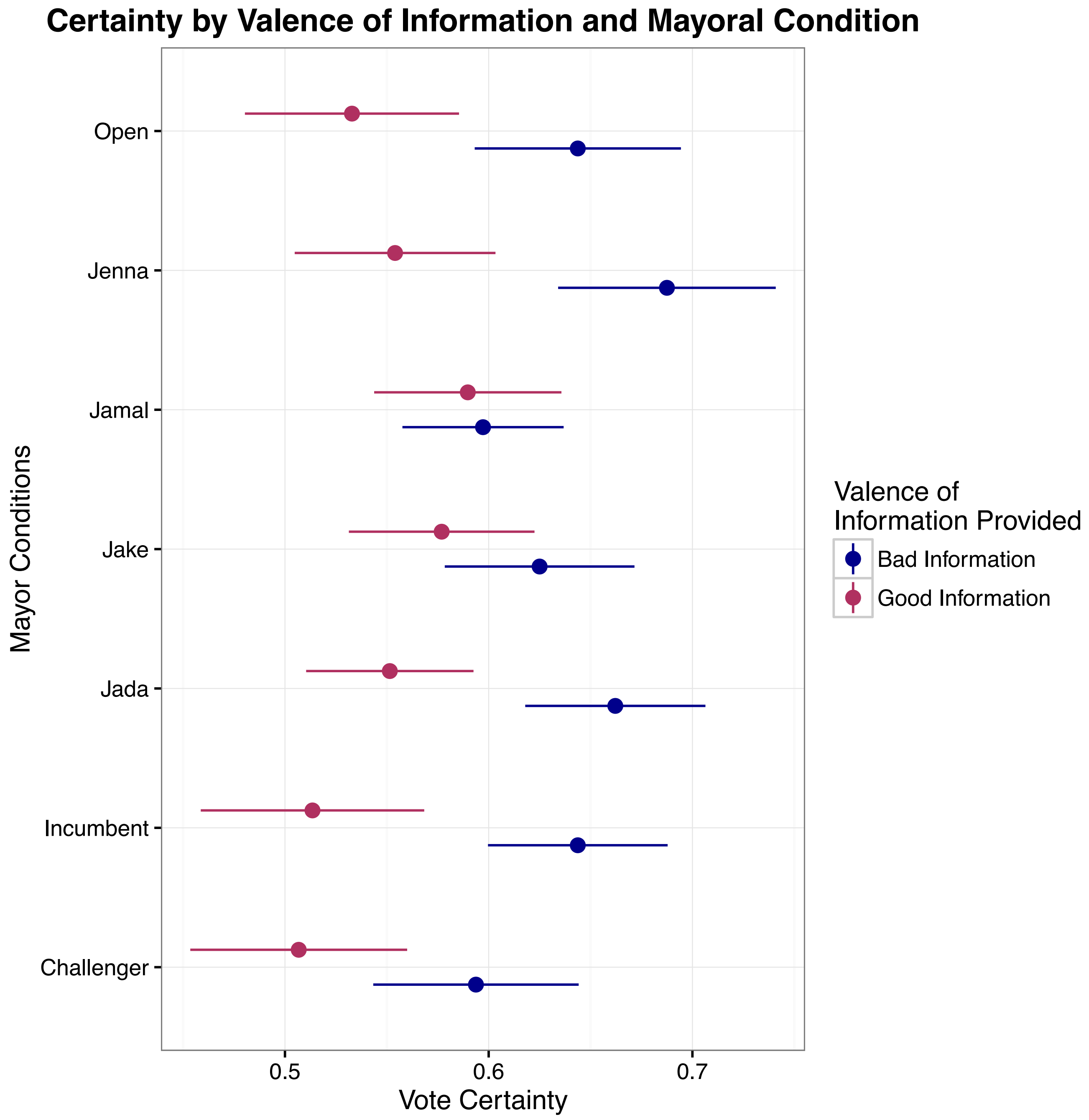
Source: National Conference of State Legislatures

THE NEW YORK TIMES

NY Times

BIVARIATE ANALYSES

- When your X variable is categorical or ordinal, you'll need a stacked bar, box plot, or coefficient plot



BIVARIATE ANALYSES

- Analyses with ordinal Y variables are often best presented as mean-centered stacked bars

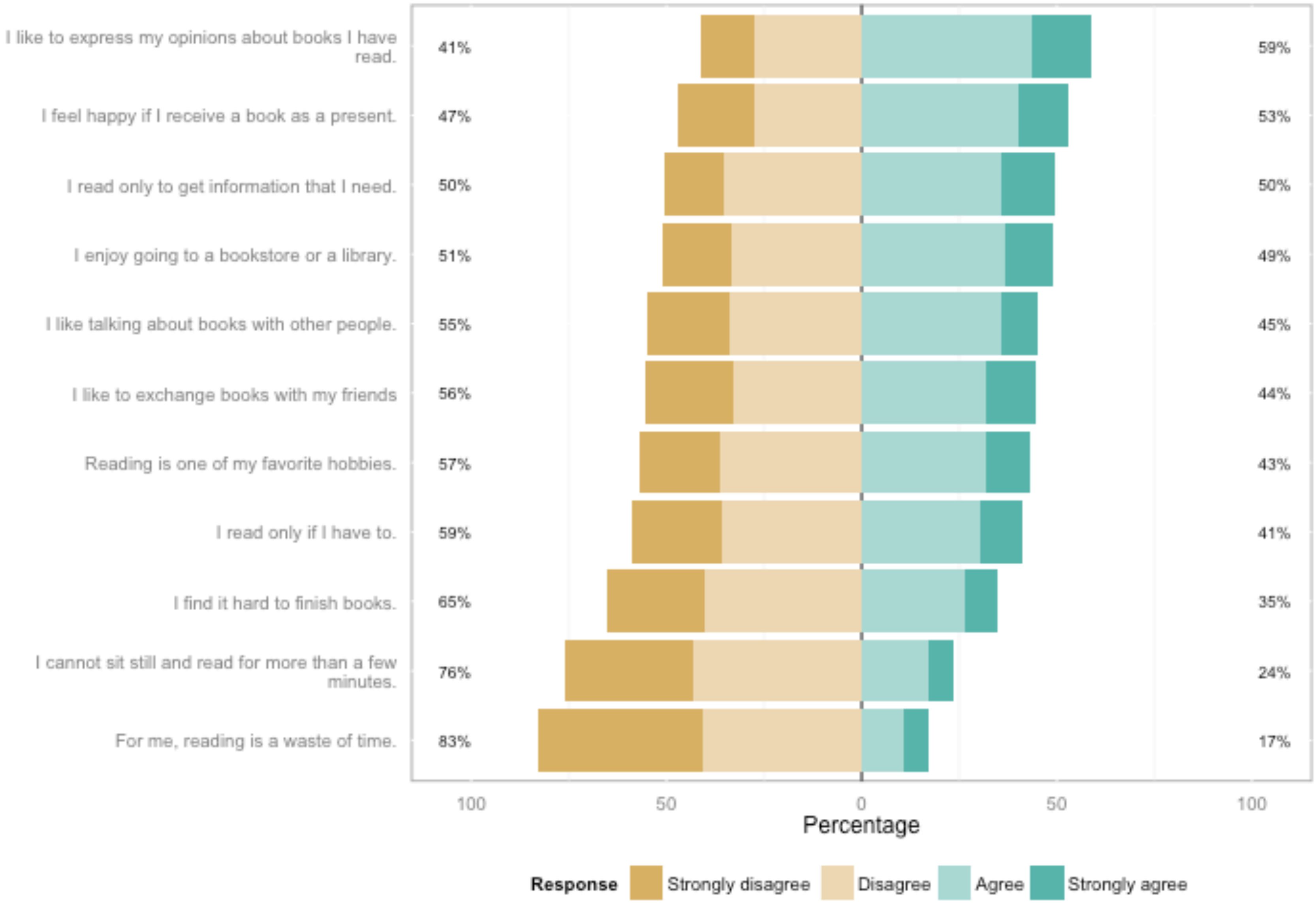


Chart from Jason Bryer

BIVARIATE ANALYSES

- Analyses with discrete X variables (for example, by year, month, or day of week) variables allow you some additional options

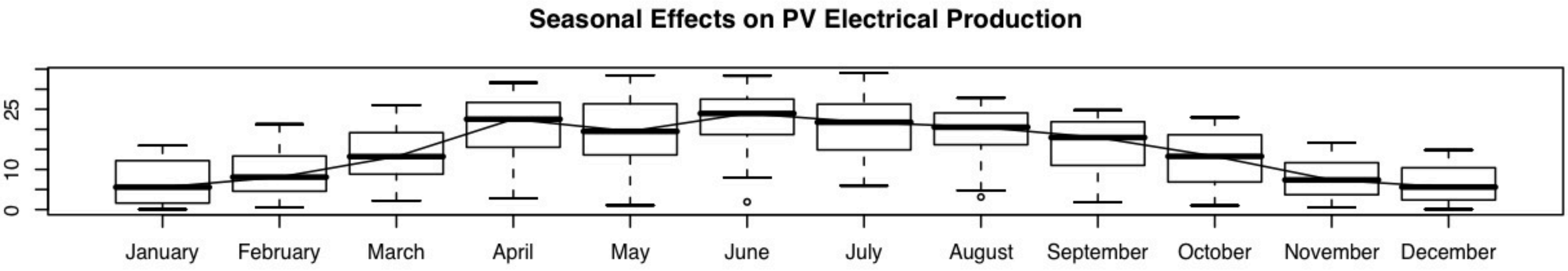
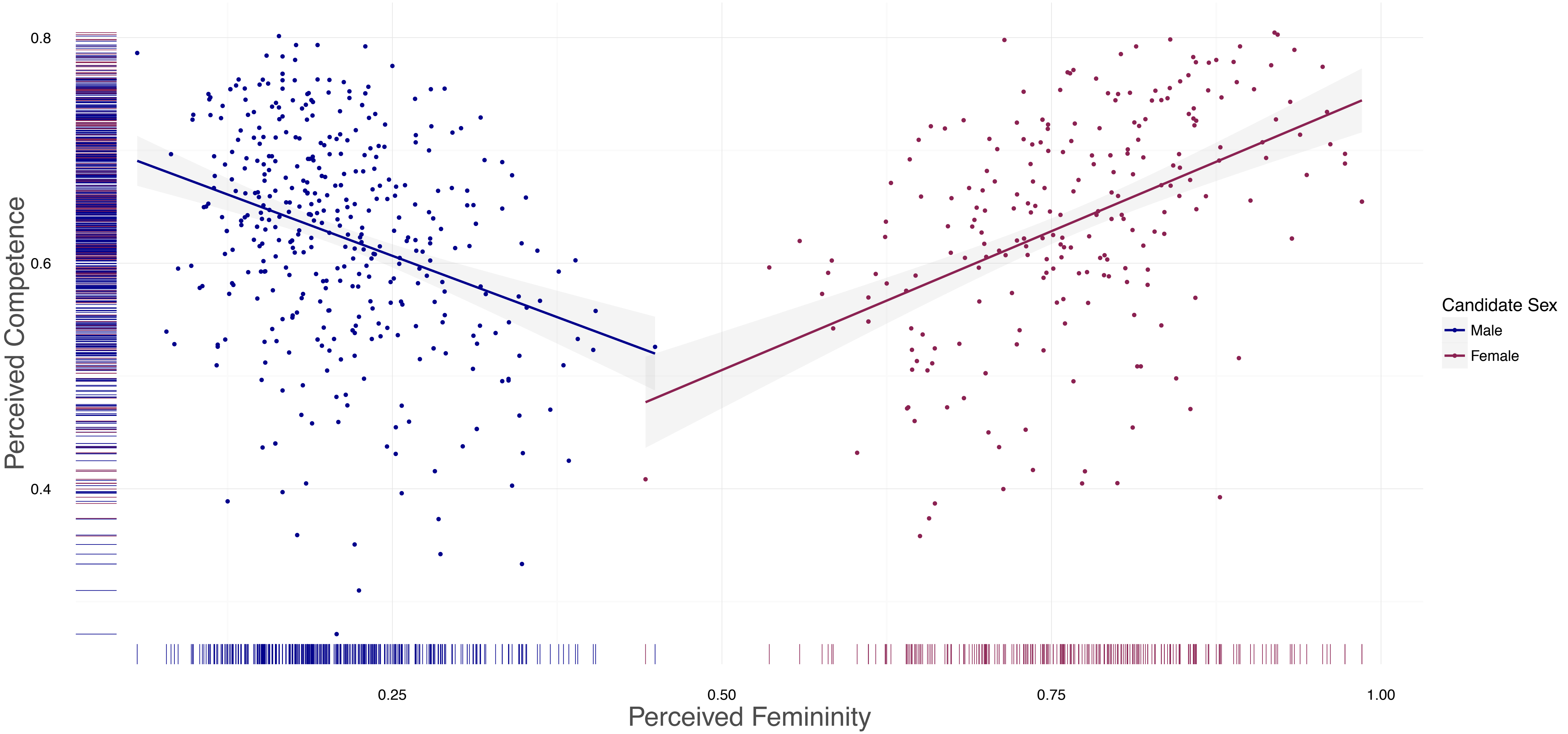


Chart from the
Personality Project

BIVARIATE ANALYSES

- With continuous X variables, the scatterplot is your best friend



MULTIVARIATE ANALYSES

- Multivariate analyses are more complicated to display—if you can plot it as a coefficient plot, or "facet" it into subsets, that's often best (otherwise, consider reverting to a regression table)

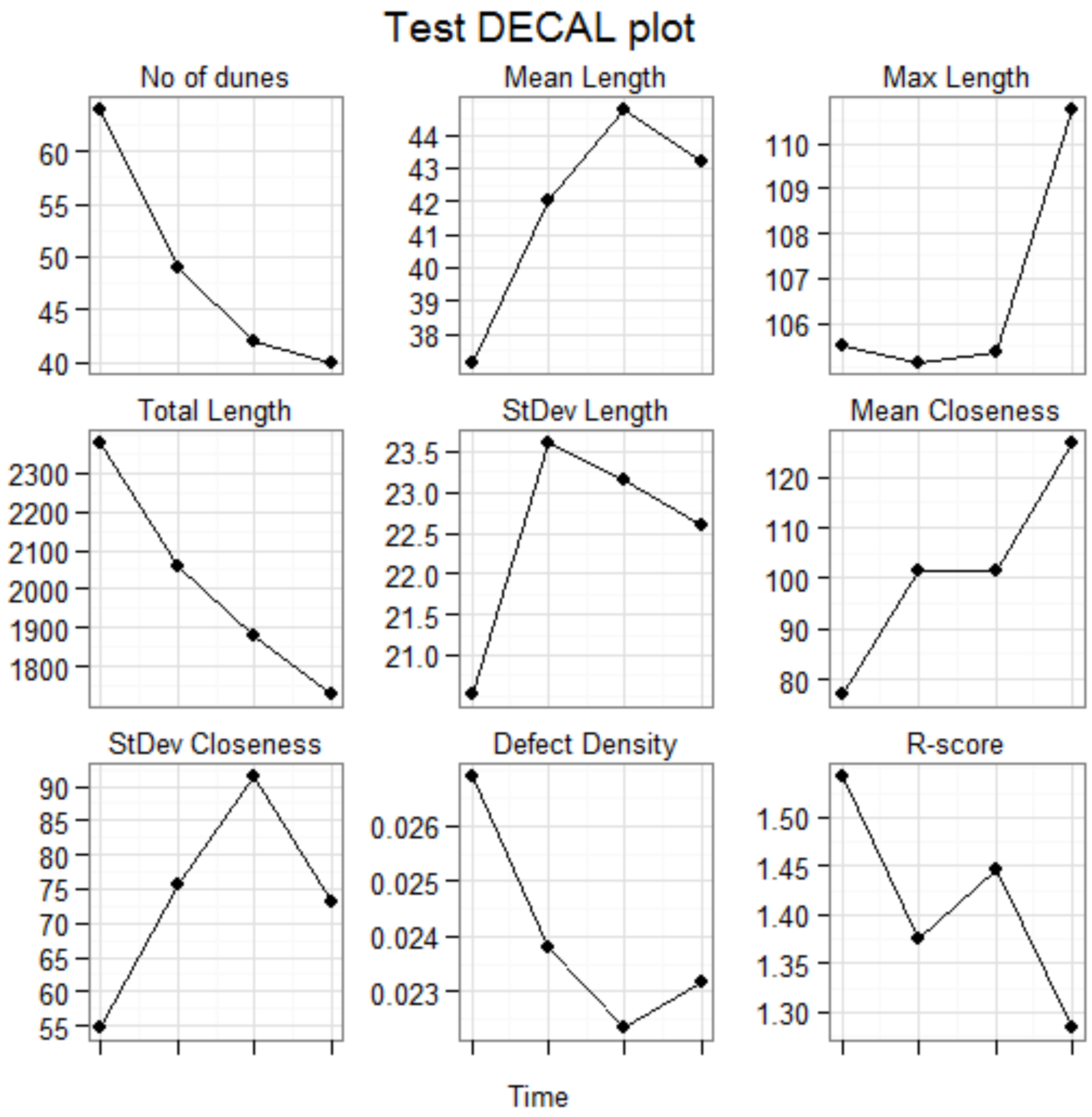
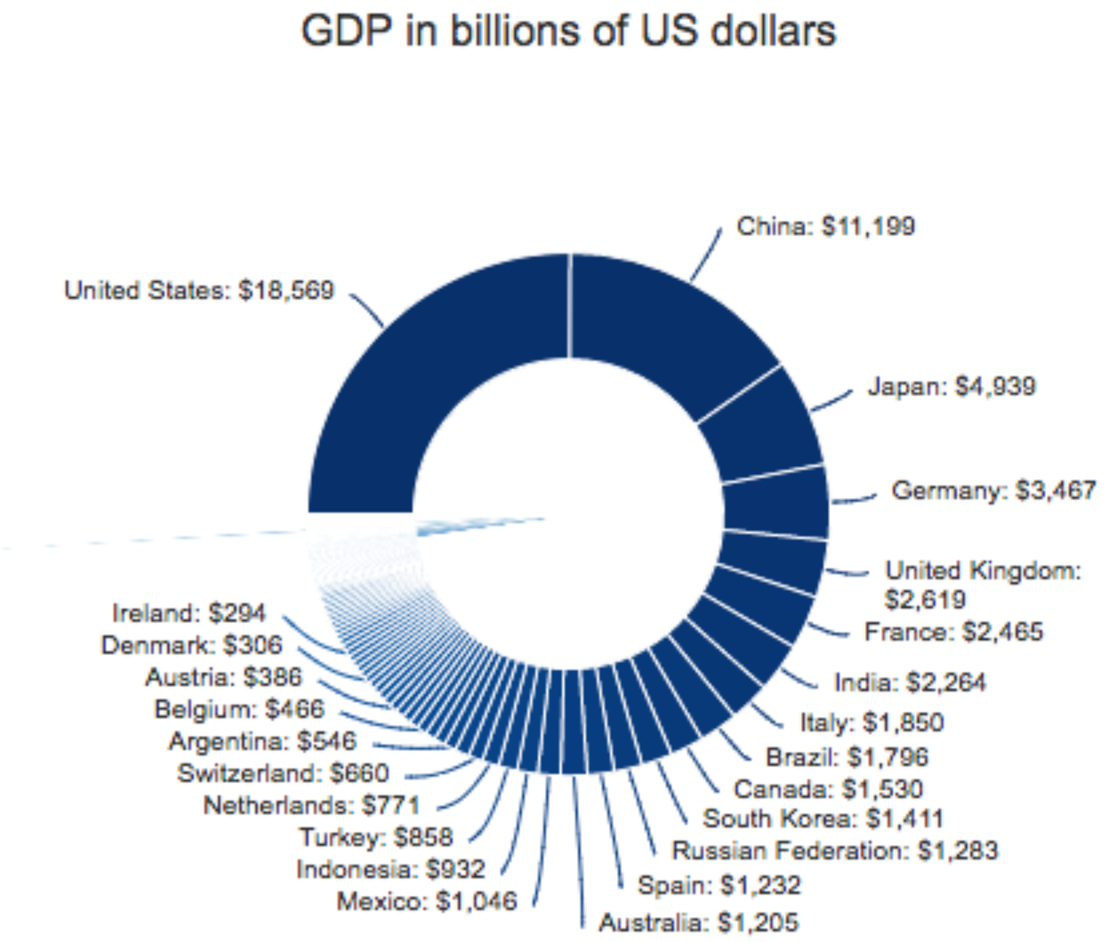


Chart from RT Wilson

OTHER KINDS OF DATA – TEXT

- The best way to display text data will often vary based on your audience; word clouds are good for large-audience presentations, but they also distort features of the data
- Donut plots are one alternative

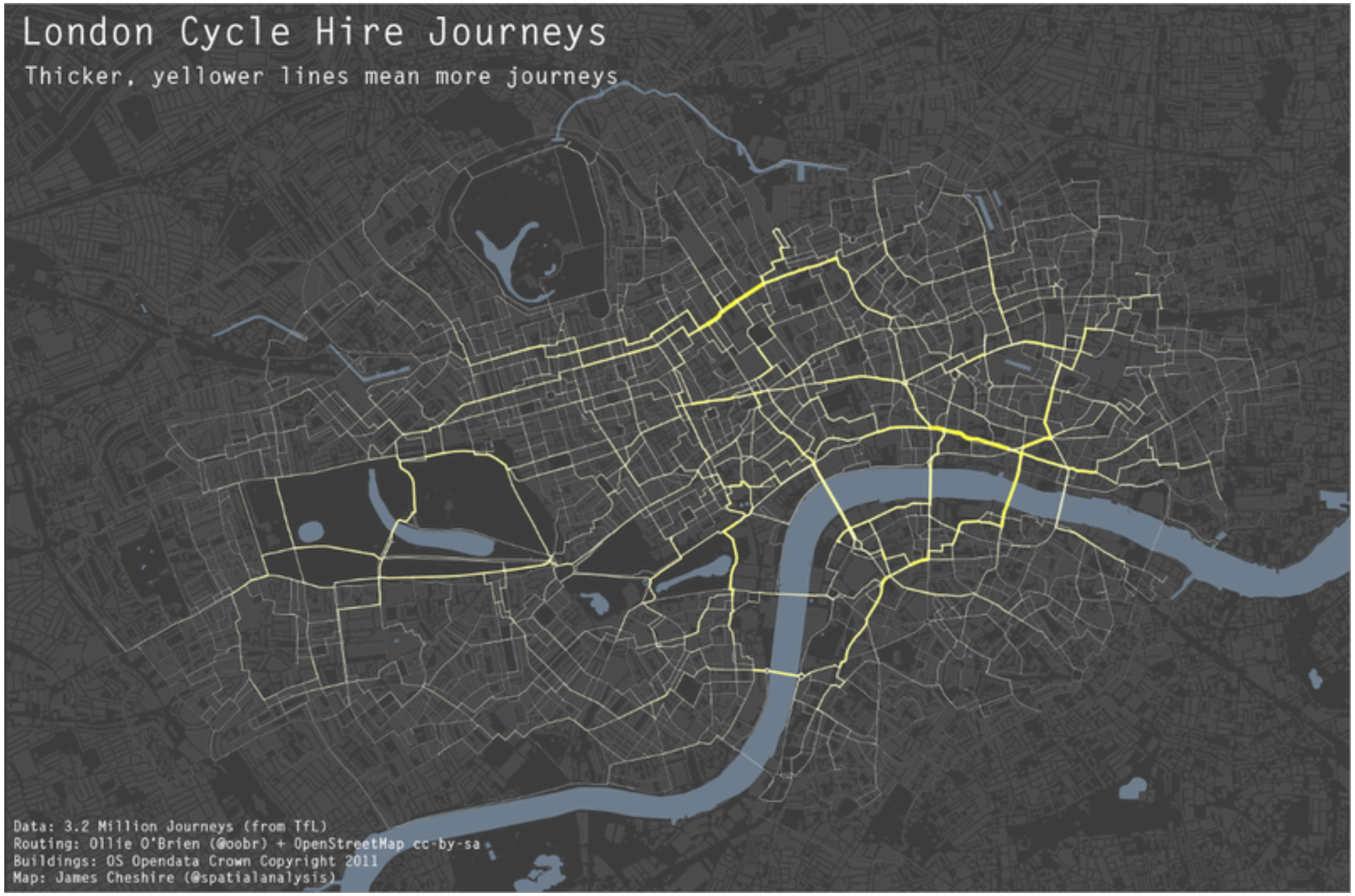


R-bloggers.com

(Thanks David for sharing!

OTHER KINDS OF DATA – MAPS

- For spatial data, simpler is almost always better
- Do not introduce non-geographic features into your geographic plots (e.g., bubbles/ cartograms, etc.)


London Cycle Hire Journeys
Thicker, yellower lines mean more journeys
Data: 3.2 Million Journeys (from TfL)
Routing: Ollie O'Brien (@oobr) + OpenStreetMap cc-by-sa
Buildings: OS Opendata Crown Copyright 2011
Map: James Cheshire (@spatialanalysis)

UCL Centre for Advanced Spatial Analysis

PLOTTING WITH GGPLOT (AND BEYOND)

PLOTTING WITH GGPLOT (AND BEYOND)

- Visuals have three key components in ggplot:

PLOTTING WITH GGPLOT (AND BEYOND)

- Visuals have three key components in ggplot:

    1. **Data** (the dataframe or matrix you’re using)

PLOTTING WITH GGPLOT (AND BEYOND)

- Visuals have three key components in ggplot:

  1. **Data** (the dataframe or matrix you're using)
  2. **Objects** (basically, the type of plot–line, bar, etc.)

PLOTTING WITH GGPLOT (AND BEYOND)

- Visuals have three key components in ggplot:

  1. **Data** (the dataframe or matrix you're using)
  2. **Objects** (basically, the type of plot–line, bar, etc.)
     - *Raw* data (think scatterplot points)

PLOTTING WITH GGPLOT (AND BEYOND)

- Visuals have three key components in ggplot:

  1. **Data** (the dataframe or matrix you're using)
  2. **Objects** (basically, the type of plot–line, bar, etc.)
     - *Raw* data (think scatterplot points)
     - *Summary* data (think regression lines)

PLOTTING WITH GGPLOT (AND BEYOND)

- Visuals have three key components in ggplot:
  1. **Data** (the dataframe or matrix you're using)
  2. **Objects** (basically, the type of plot–line, bar, etc.)
     - *Raw* data (think scatterplot points)
     - *Summary* data (think regression lines)
  3. **Aesthetics** (color, size, shape, etc.)

PLOTTING WITH GGPLOT (AND BEYOND)

PLOTTING WITH GGPLOT (AND BEYOND)

- Decisions about **objects** and **aesthetics** depend on your answers to two questions:

PLOTTING WITH GGPLOT (AND BEYOND)

- Decisions about **objects** and **aesthetics** depend on your answers to two questions:
    - What is your core argument? (I.e., why do you even need a graph in the first place?)

PLOTTING WITH GGPLOT (AND BEYOND)

- Decisions about **objects** and **aesthetics** depend on your answers to two questions:
    - What is your core argument? (I.e., why do you even need a graph in the first place?)
    - Who is your audience? (A journal article, a conference presentation, a TED talk?)

PLOTTING WITH GGPLOT (AND BEYOND)

- Decisions about **objects** and **aesthetics** depend on your answers to two questions:
    - What is your core argument? (I.e., why do you even need a graph in the first place?)
    - Who is your audience? (A journal article, a conference presentation, a TED talk?)
        - The former probably want to see more of the raw data (that is, you're using a chart so you can show more *complexity* and more *raw data* than a table of figures would allow), the latter probably want to see the most digestible version of the story