Introduction to R Statistical Software

Z Kelly

Joint Legislative Audit and Review Committee

Washington State Legislature

4/27/2016

Overview

- Background on R
- What can R do for you?
- Examples
  - File formats
  - Data management
  - Summary statistics
  - Visualization
  - Data analysis
- Advanced functionality
  - Markdown authoring
  - Rpres – HTML5 presentations
  - Shiny
  - LaTeX
  - Github
- Many, many other features
  - Scripting, C++, python, batch processing, geocoding, google maps, Markdown + CSS
- Suggested references

What is R?

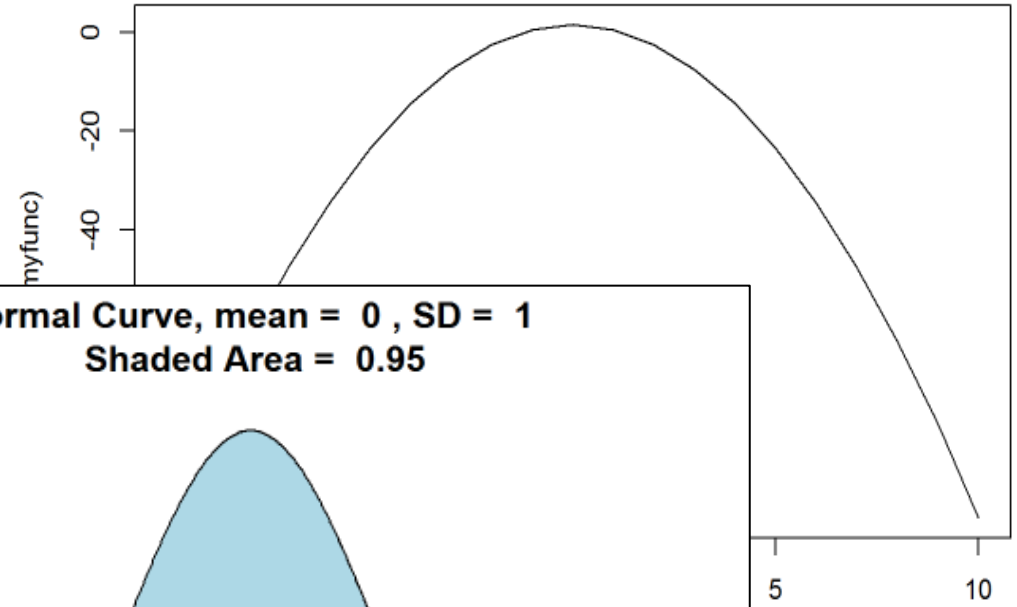
- The fastest growing language on stack exchange

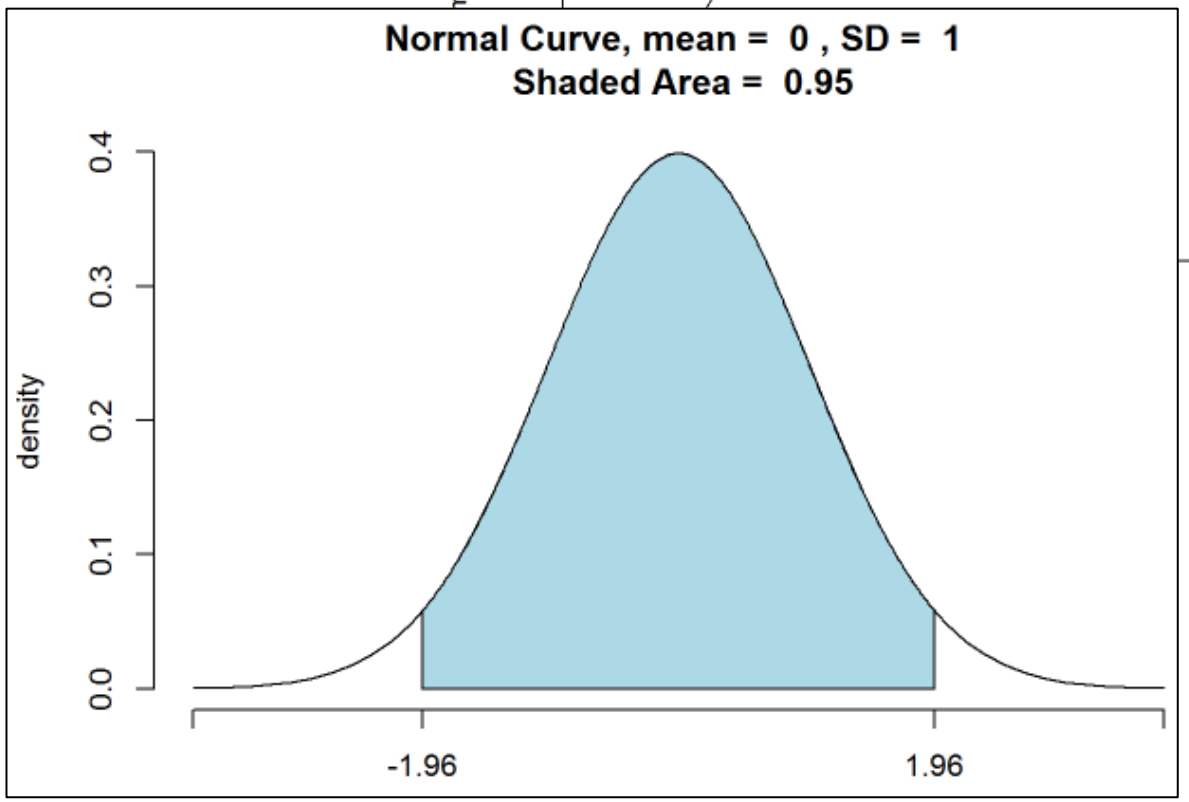
But is it __?

- A calculator
- A *graphing* calculator
- A statistical reference

```
>> 2 + 2

## [1] 4

>> 2 * sqrt(pi)

## [1] 3.545
```

```
>> myfunc <- deriv(~1.5 - sqrt(x^4), "x")
>> x <- -10:10
>> eval(myfunc)
```





R is a language

- A language and a development environment
  - A full suite of tools for data manipulation, storage, operations, analysis, graphical display, and programming
  - Intended for all aspects of statistical analysis, broadly defined
  - Based on S

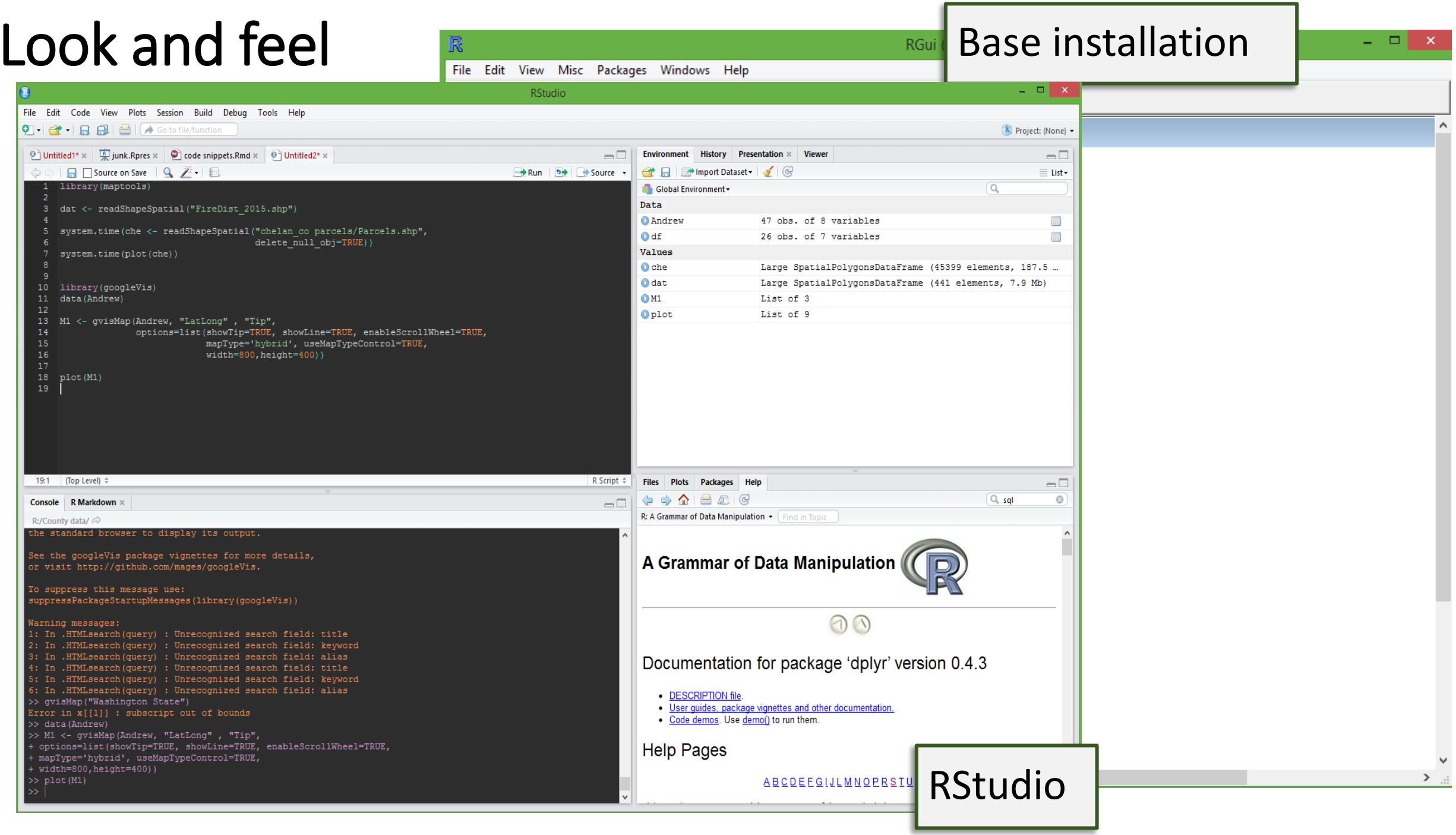
  - Base R is easily extended by user-contributed modules called *packages* and it is easy to integrate C, C++, Java, Python, Ruby, Pearl, and more

  - Distribution: GNU GPL

Considerations

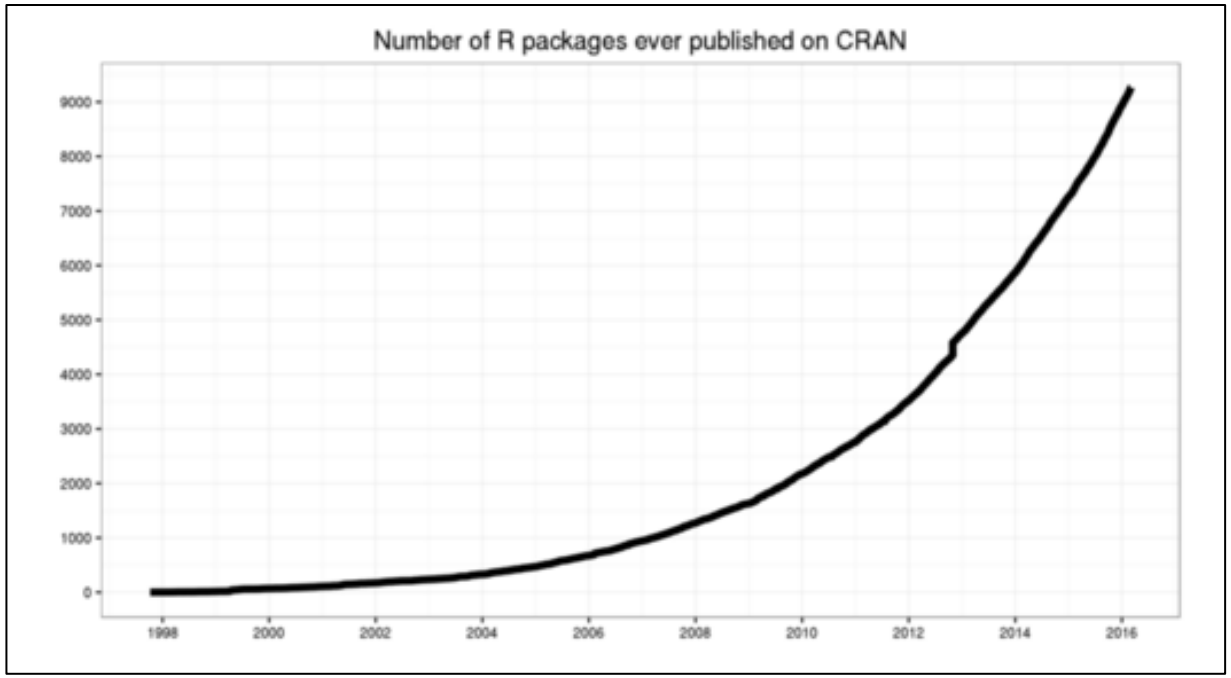
- Complete suite of data tools
- Efficient functionality
- Nice graphics
- Well documented
- Large, helpful community; new packages added almost daily
- Price
- Learning curve
- What you see is what you get

Look and feel



What are packages?

- User-written functionality
  - Hosted on multiple mirrors worldwide
- Browse at: cran.r-project.org



Bayesian	Bayesian Inference
ChemPhys	Chemometrics and Computational Physics
ClinicalTrials	Clinical Trial Design, Monitoring, and Analysis
Cluster	Cluster Analysis & Finite Mixture Models
DifferentialEquations	Differential Equations
Distributions	Probability Distributions
Econometrics	Econometrics

Task view

The packages in this view can be roughly structured i

Basic linear regression

- *Estimation and standard inference* : Ordinary l
  various methods such as `summary()` and `anova()`
- *Further inference and nested model compariso*
  instead of *F* tests) and plug-in of other covariar
  nonlinear hypotheses in `deltaMethod()` in car.
- *Robust standard errors* : HC and HAC covariar
- *Nonnested model comparisons* : Various tests f
  nested models is provided by nonnest2 (and sp
- *Diagnost checking* : The packages car and lmte

Microeconometrics

- *Generalized linear models (GLMs)* : Many star
  in particular logit and probit models for modeli

car: Companion to Applied Regression

Functions and Datasets to Accompany J. Fox and S. Weisberg, An R Companion to Applied Regression, Second Edition, Sage, 2011.

High standard for documentation

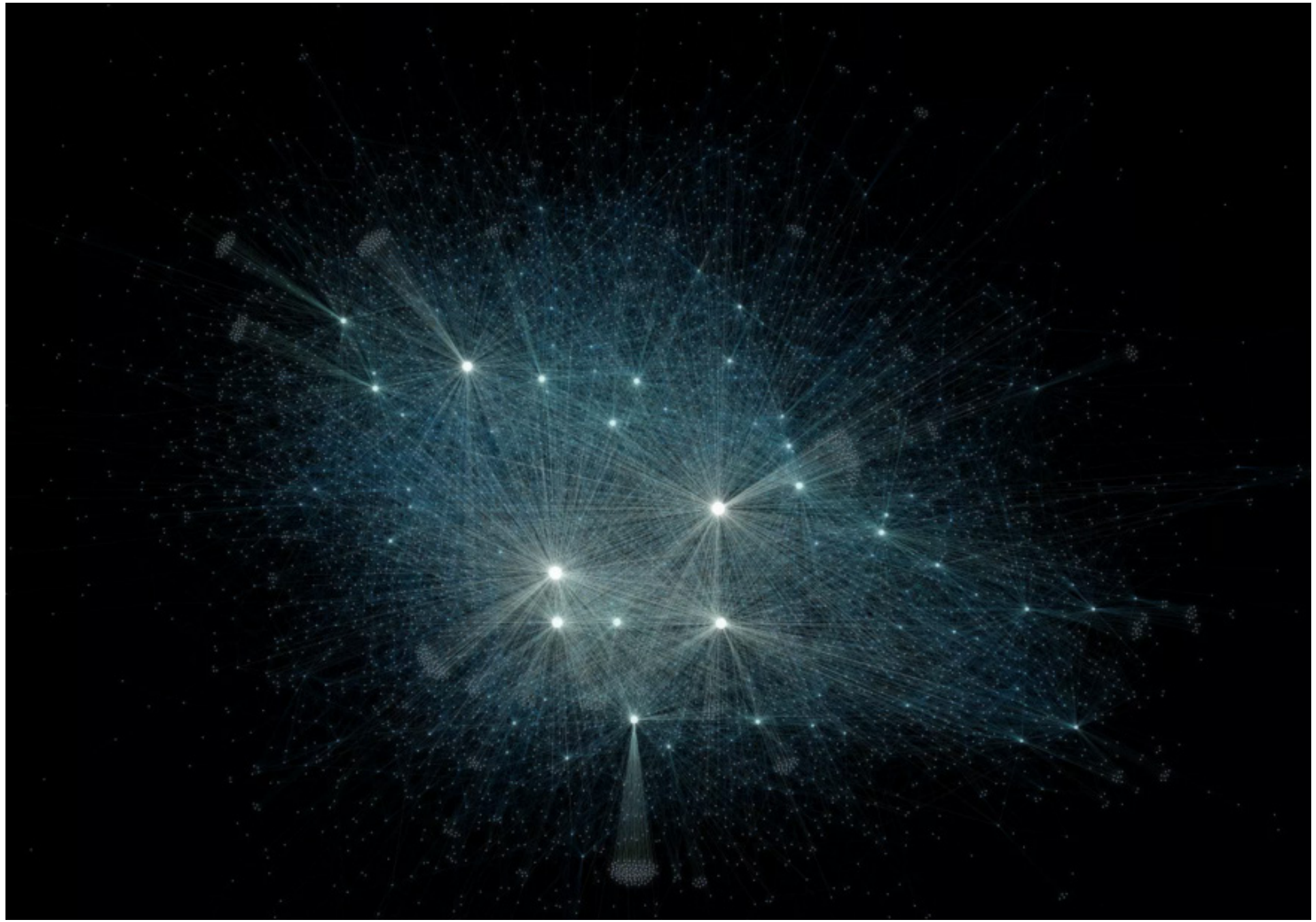
Version:	2.1-2
Depends:	R (≥ 3.2.0)
Imports:	MASS, mgcv, nnet, pbkrtest (≥ 0.4-4), quantreg, grDevices, utils, stats, graphics
Suggests:	alr4, boot, coxme, leaps, lme4, lmtest, Matrix, MatrixModels, nlme, rgl (≥ 0.93.960), sandwich, SparseM, survival, survey
Published:	2016-03-25
Author:	John Fox [aut, cre], Sanford Weisberg [aut], Daniel Adler [ctb], Douglas Bates [ctb], Gabriel Baud-Bovy [ctb], Steve Ellison [ctb], David Firth [ctb], Michael Friendly [ctb], Gregor Gorjanc [ctb], Spencer Graves [ctb], Richard Heiberger [ctb], Rafael Laboissiere [ctb], Georges Monette [ctb], Duncan Murdoch [ctb], Henric Nilsson [ctb], Derek Ogle [ctb], Brian Ripley [ctb], William Venables [ctb], David Winsemius [ctb], Achim Zeileis [ctb], R-Core [ctb]
Maintainer:	John Fox <jfox at mcmaster.ca>
License:	GPL-2 | GPL-3 [expanded from: GPL (≥ 2)]
URL:	https://r-forge.r-project.org/projects/car/, http://CRAN.R-project.org/package=car, http://socserv.socsci.mcmaster.ca/jfox/Books/Companion/index.html
NeedsCompilation:	no
Citation:	car citation info
Materials:	NEWS
In views:	Econometrics, Finance, Multivariate, SocialSciences
CRAN checks:	car results

Downloads:

Reference manual:	car.pdf
Vignettes:	Using car functions inside user functions
Package source:	car_2.1-2.tar.gz
Windows binaries:	r-devel: car_2.1-2.zip, r-release: car_2.1-2.zip, r-oldrel: car_2.0-25.zip
OS X Snow Leopard binaries:	r-release: car_2.1-2.tgz, r-oldrel: car_2.0-25.tgz
OS X Mavericks binaries:	r-release: car_2.1-2.tgz
Old sources:	car archive

Reverse dependencies:

Reverse depends:	AER, alr3, alr4, bartMachine, bgmm, candisc, CrypticIBDcheck, DAMisc, Deducer, DiagTest3Grp, DistatisR, DJL, epr, extRemes, genridge, granova, heplots, hysteresis, ibd, ITEMAN, lmSupport, mosaic, muma, mvinfluence, papeR, pequod, phia, quantification, Rcmdr, RcmdrMisc, RcmdrPlugin.SM, seeg, specificity, stepp, systemfit, VARSEDIG
Reverse imports:	afex, anacor, apaTables, apt, ART, ARTool, AutoModel, bayesLife, BCA, caret, Countr, drc, drsmooth, dynlm, easyanova, EffectLiteR, ez, FactoMineR, FSA, funcy, fxregime, gamclass, gcmr, medflex, miceadds, micEconCES, mixlm, NHPoisson, panelAR, plm, plsRglm, rasclass, RcmdrPlugin.BCA, referenceIntervals, rockchalk, rpubchem, RVAideMemoire, ryouready, sdcMicro, sjPlot, tadaatoolbox, translateSPSS2R, userfriendlyscience, VIM, VSE, zetadiv
Reverse suggests:	abd, agridat, betareg, bglm, BIFIEsurvey, BiodiversityR, bnormnlr, codingMatrices, compareGroups, Ecdat, Ecfun, effects, fastR, fscaret, gmm, gmnl, gmum.r, GoodmanKruskal, HH, HistData, lmtest, lsmeans, matlib, McSpatial, meboot, mistat, mlogit, Morpho, multcomp, NPC, pedometrics, perturb, pglm, polywog, pscl, R2MLwiN, Rchoice, RcmdrPlugin.EZR, RcmdrPlugin.IPSUR, RcmdrPlugin.pointG, rddtools, REdaS, sand, sandwich, SenSrivastava, sla, Sleuth2, Sleuth3, SMIR, strucchange, tableplot, vcdExtra, WRS2
Reverse enhances:	memisc



Getting data into R

- Base R can read many file formats
  - “read.csv”
  - “load” – for .Rdata format
  - “readLines” – generic
  - “url” – to open a connection

- Many contributed packages extend functionality
  - “readSpatialPolygons”
  - “readHTMLTable”
  - Database drivers – SQLite, MongoDB, etc…

Analysis

Describing data

Summarizing the dataset

```
data <- gss08
dim(gss08)
```

```
[1] 2023   12
```

```
names(gss08)
```

```
 [1] "sex"      "race"     "degree"   "relig"    "polparty" "cappun"
 [7] "tvhours"  "marijuan" "owngun"   "gunlaw"   "age"      "chldidel"
```

```
str(gss08)
```

```
'data.frame':    2023 obs. of  12 variables:
 $ sex     : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 2 1 1 1 1 ...
 $ race    : Factor w/ 3 levels "AfrAm","Other",..: 2 2 1 1 1 1 1 1 1 1
 $ degree  : Factor w/ 5 levels "Bachelor","Graduate",..: 3 2 3 3 3 1 3
```

Summarize variables

```
summary(gss08)
```

```
       degree                relig
Bachelor   : 355   Catholic   : 470
Graduate   : 194   Jewish     :  39
HighSchool:1003   None       : 332
JunColl    : 173   Other      : 133
NotHs      : 297   Protestant:1040
NA's       :   1   NA's       :   9

    tvhours          marijuan     owngun
 Min.   : 0    Legal   :496   No  :870
 1st Qu.: 1    NotLegal:751   Yes :464
 Median : 2    NA's    :776   NA's:689
 Mean   : 3
 3rd Qu.: 4
 Max.   :24
 NA's   :699
```

Data: General Social Survey 2008

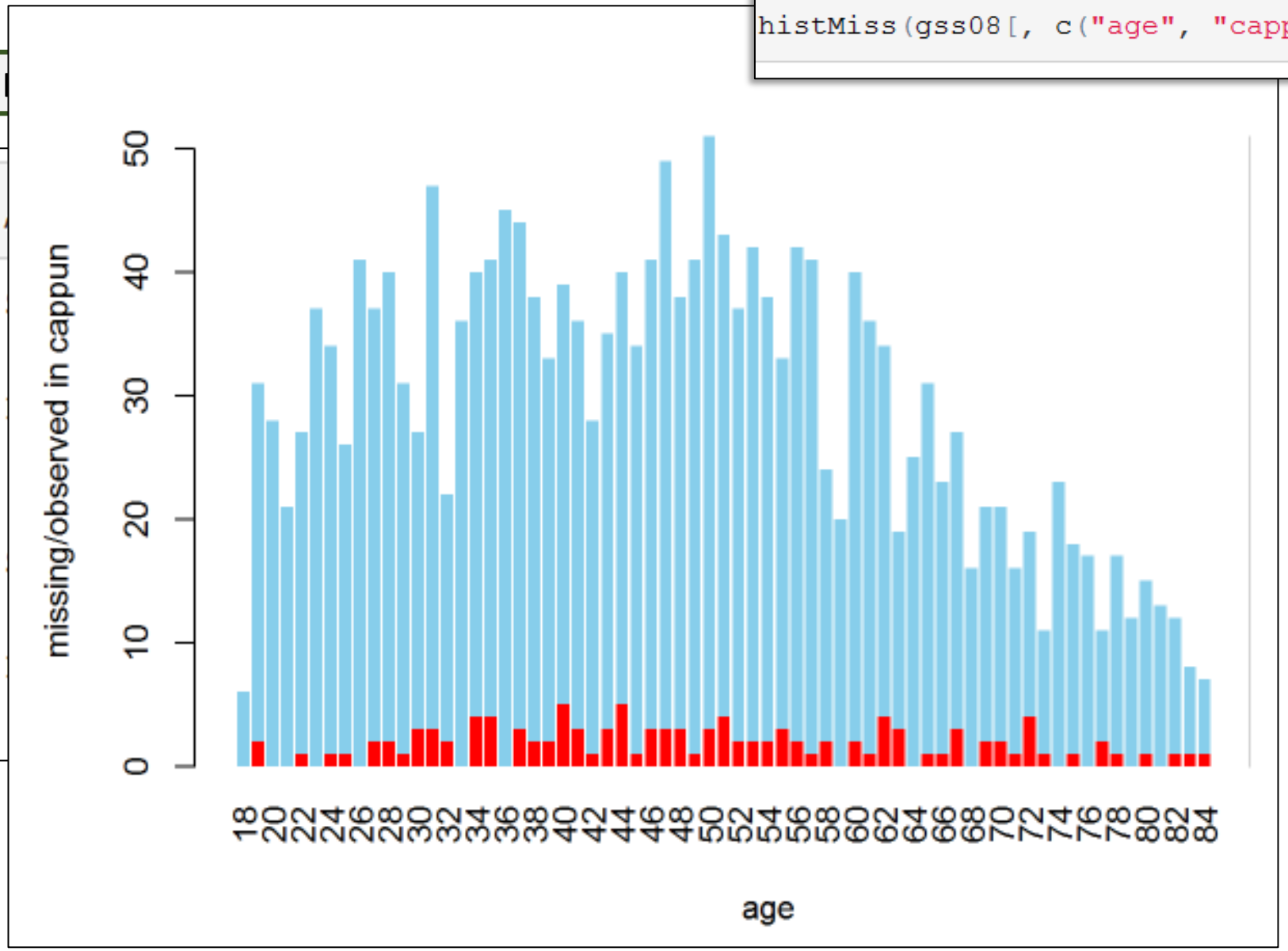
What's the pattern of missing responses?

One option: wr

```
apply(gss08, 2,
```

```
$sex

FALSE
 2023

$race

FALSE
 2023
```

```
library(VIM)
histMiss(gss08[, c("age", "cappun")], border = NA)
```



Tabulating data

Base R functions: “table” and “ftable”

```
>> ftable(table(gss08$sex, gss08$polparty,
 gss08$cappun, useNA = "ifany"))
```

```
##                       Favor Oppose  NA
##
## Female Democrat         203    200  19
##        Independent      217    123  41
##        Other              3      5   0
##        Republican       206     54  13
##        NA                 3      3   4
## Male   Democrat         169    118  12
##        Independent      240     99  26
##        Other             19      9   2
##        Republican       200     28   4
##        NA                 3      0   0
```

Many others: “htmlTable”, “xtable”, “stargazer”, “Hmisc”, “**printr**”

```
>> kable(x, caption = "A table produced
 by kable.")
```

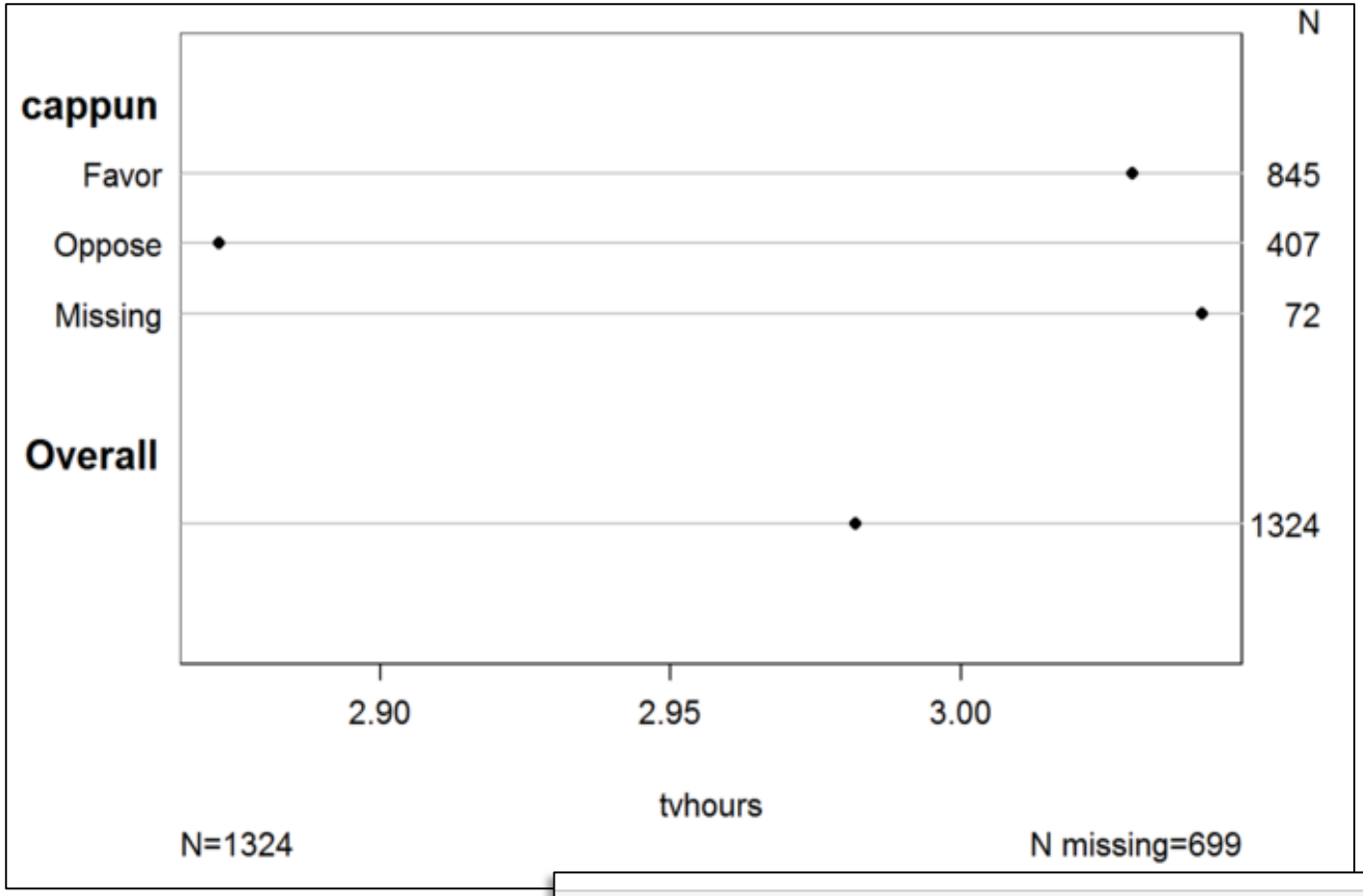
A table produced by kable.

	Female	Male
Democrat	422	299
Independent	381	365
Other	8	30
Republican	273	232

Descriptive plots

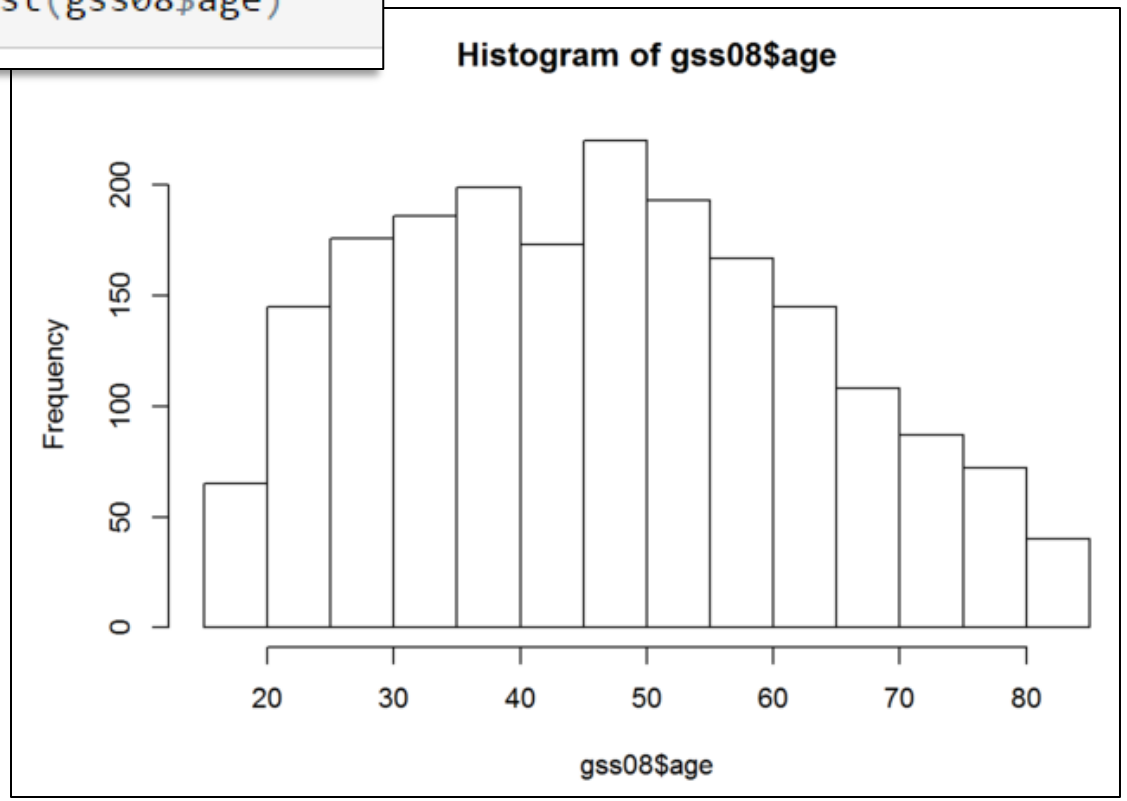
- Again, many options (beyond base R)

- Ex: from “Hmisc”
  - Calls several functions, nested together

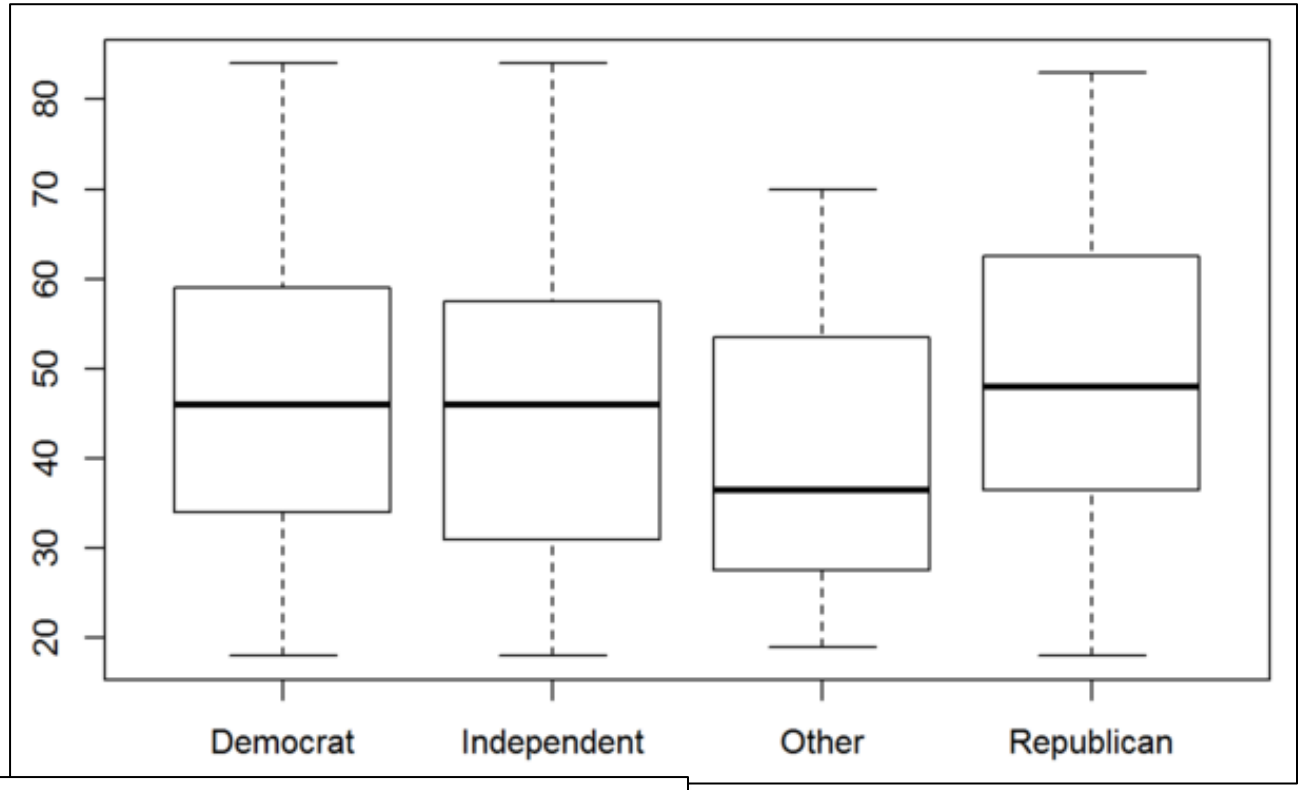



```
plot(summary(tvhours ~ cappun, data = gss08))
```

```r
# base R histogram
hist(gss08$age)
```



Base R is very efficient for making such plots

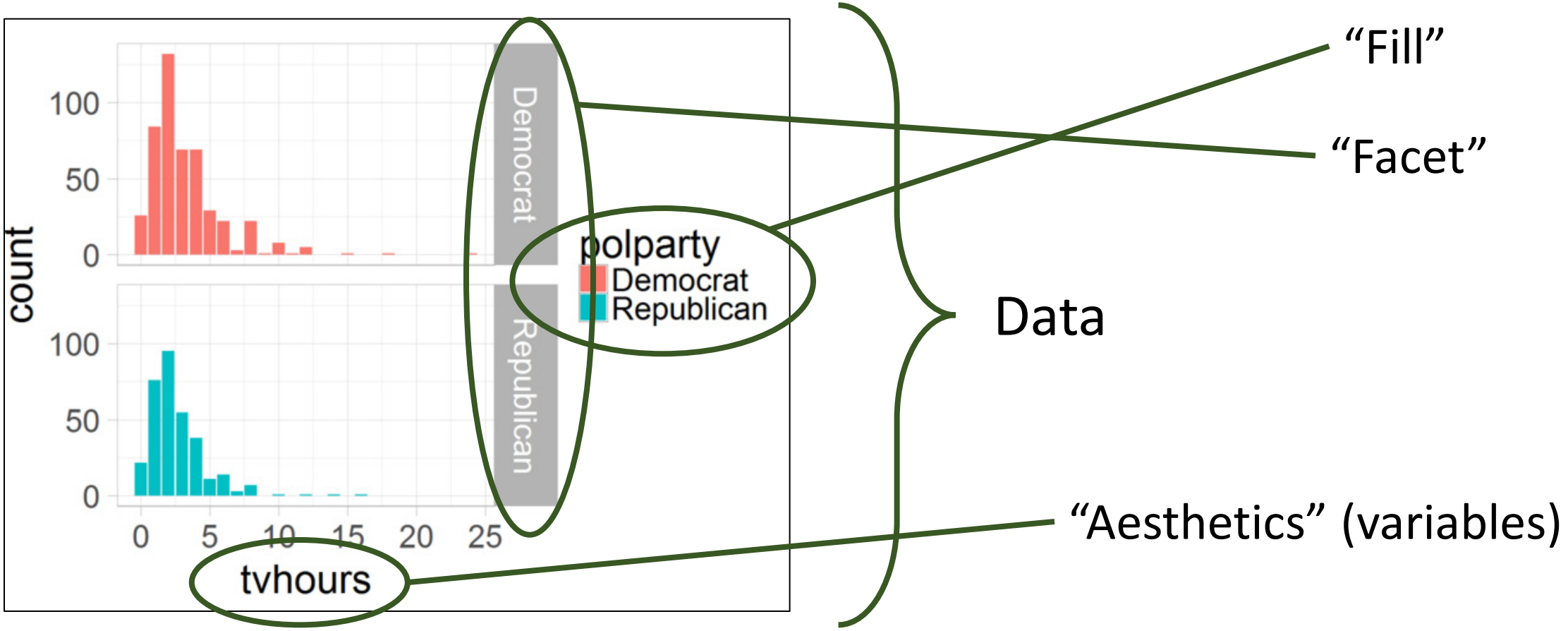



```r
boxplot(gss08$age ~ gss08$polparty)
```

But default plotting methods produce simple graphics

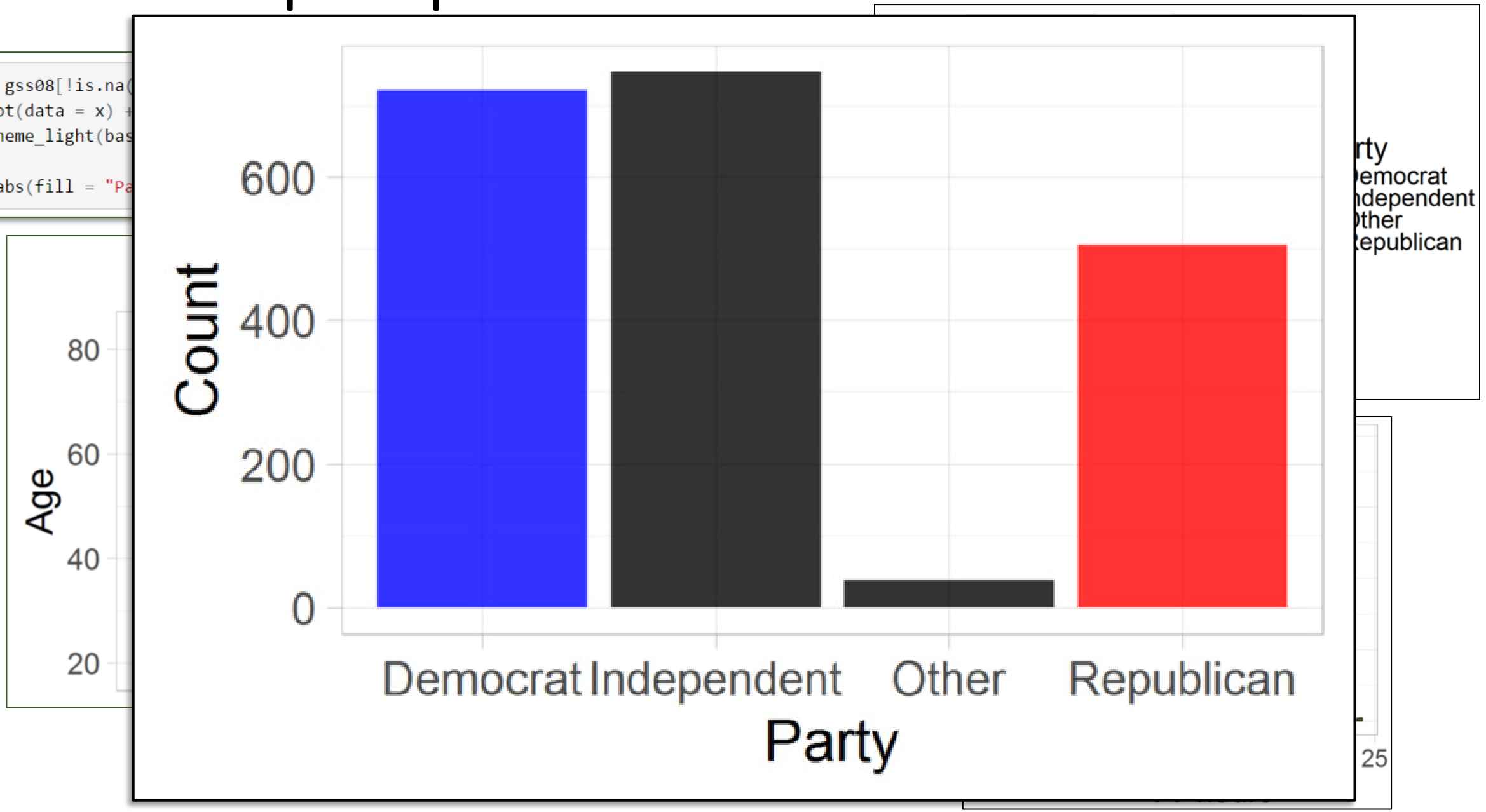
Descriptive plots, cont.

- The most popular options for high-quality graphics are “lattice” and “ggplot2”
  - Sidenote: ggplot2 and the “grammar of graphics” (very, very quickly)



More descriptive plots

```
x <- gss08[!is.na(
ggplot(data = x) +
   theme_light(bas

   labs(fill = "Pa
```



Statistical tests

- Far too many to list here
  - T-tests, X^2, rank-sum tests, Anova, unit-roots, etc...
- Examples:
  - Does opinion on legal marijuana by political party?
  - What about average hours of TV viewing?

```
t <- table(gss08$polparty, gss08$marijuan)
chisq.test(t)
```

```
 Pearson's Chi-squared test

data:  t
X-squared = 30, df = 3, p-value = 0.0000004
```

```
x <- dplyr::filter(gss08, polparty == "Democrat" | polparty == "Republican")
t.test(tvhours ~ polparty, data = x)
```

```
 Welch Two Sample t-test

data:  tvhours by polparty
t = 4, df = 800, p-value = 0.0005
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.262 0.923
sample estimates:
  mean in group Democrat mean in group Republican
                    3.21                      2.62
```

- If there's a test you'd like to perform, it's probably already implemented somewhere

Statistical Modeling

- Almost every type of modeling uses a common *formula* style

$$Y \sim Var_1 + Var_2 + \ldots$$

- Bivariate/multivariate models easily implemented
  - Use “summary()” or “print()” to see the results
  - Several packages exist for creating nicely formatted regression tables and plots

ANOVA

Example:

- First run ANOVA with TV hours as dependent variable
- Then, run Tukey's honest differences test

```
fit <- aov(tvhours ~ polparty + sex + degree, data = gss08)
summary(fit)
```

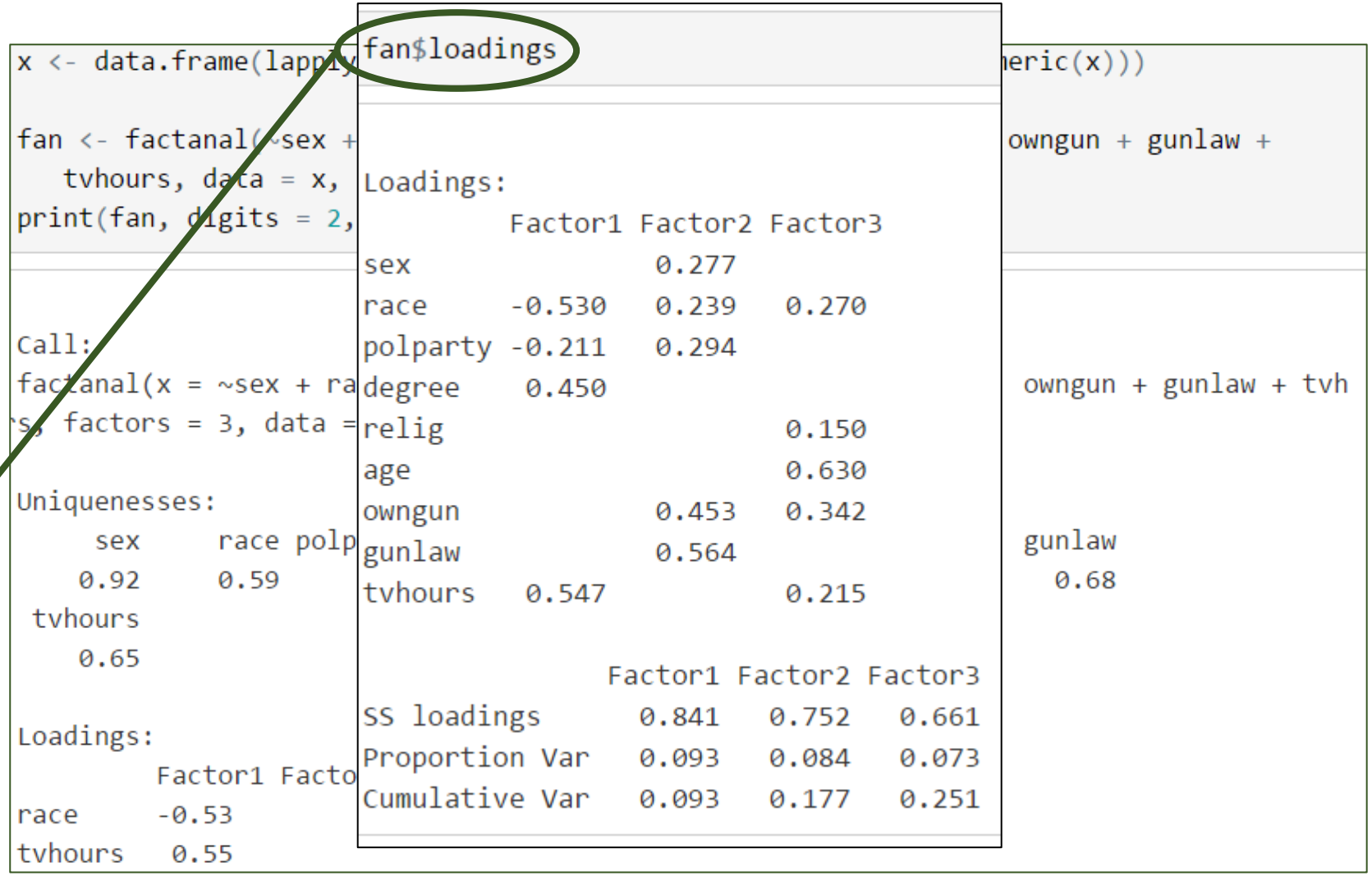
```
             Df Sum Sq Mean Sq F value Pr(>F)
polparty      3     99    32.9    5.08 0.0017 **
sex           1     12    11.5    1.78 0.1819
degree        4    766   191.4   29.58 <2e-16 ***
Residuals  1309   8471     6.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
705 observations deleted due to missingness
```

```
TukeyHSD(fit)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = tvhours ~ polparty + sex + degree, data = gss08)

$polparty
                            diff    lwr     upr p adj
Independent-Democrat    -0.161 -0.581  0.2597 0.759
Other-Democrat          -1.341 -2.739  0.0557 0.065
Republican-Democrat     -0.593 -1.064 -0.1213 0.007
Other-Independent       -1.181 -2.577  0.2149 0.130
Republican-Independent  -0.432 -0.899  0.0350 0.082
Republican-Other         0.749 -0.663  2.1608 0.522

$sex
               diff    lwr    upr p adj
Male-Female -0.186 -0.461 0.0891 0.185

$degree
                        diff    lwr   upr p adj
Graduate-Bachelor    -0.245 -1.029 0.539 0.913
HighSchool-Bachelor   0.831  0.295 1.368 0.000
JunColl-Bachelor      0.326 -0.466 1.119 0.794
NotHs-Bachelor        2.339  1.653 3.025 0.000
HighSchool-Graduate   1.077  0.390 1.763 0.000
JunColl-Graduate      0.572 -0.329 1.473 0.414
```

Factor Analysis/PCA

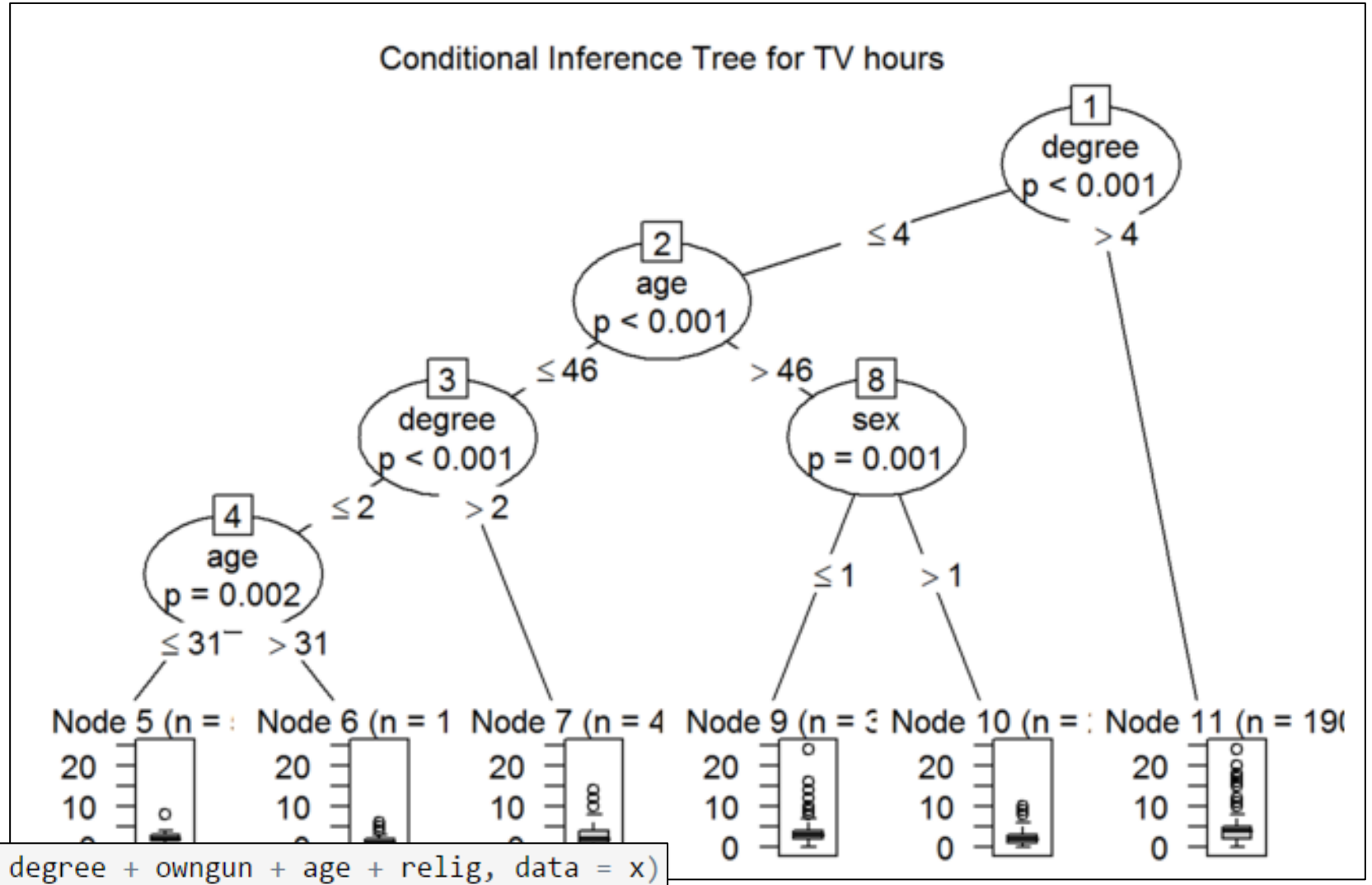
- Uses the standard formula call
  - Specify number of factors and rotation
  - Can plot from there
  - And the resulting object retains p-values, loadings, etc…



(There are also packages for Structural Equation Modeling)

Tree models

- Via “rpart”, “party”, “randomForest”
  1. Grow a tree
  2. Visualize results
  3. Prune




```
fit <- ctree(tvhours ~ sex + polparty + degree + owngun + age + relig, data = x)
plot(fit, main = "Conditional Inference Tree for TV hours")
```

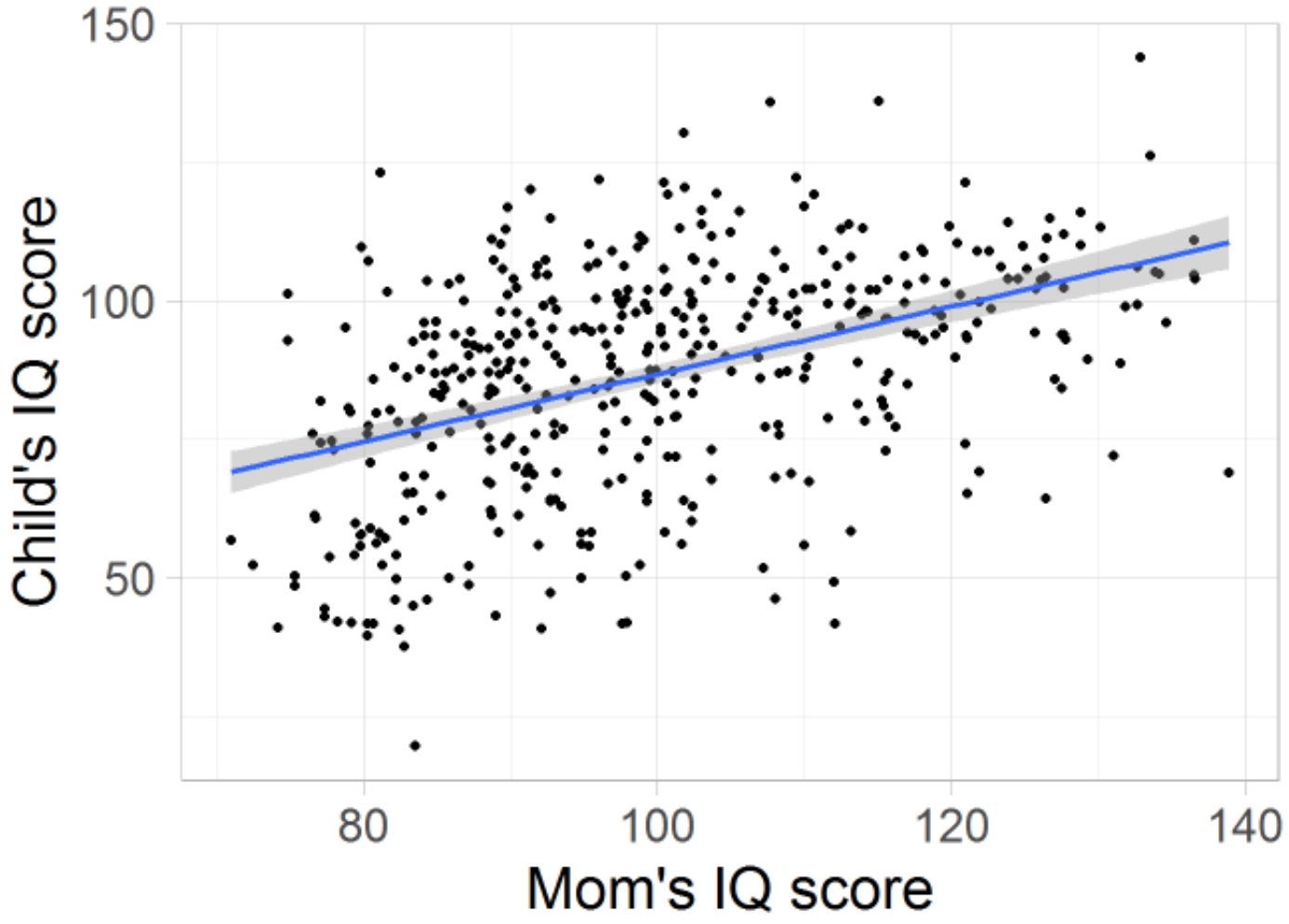
Regression

Data: child IQ measures

```
fit01 <- lm(kid_score ~ mom_iq, data = iq)
```

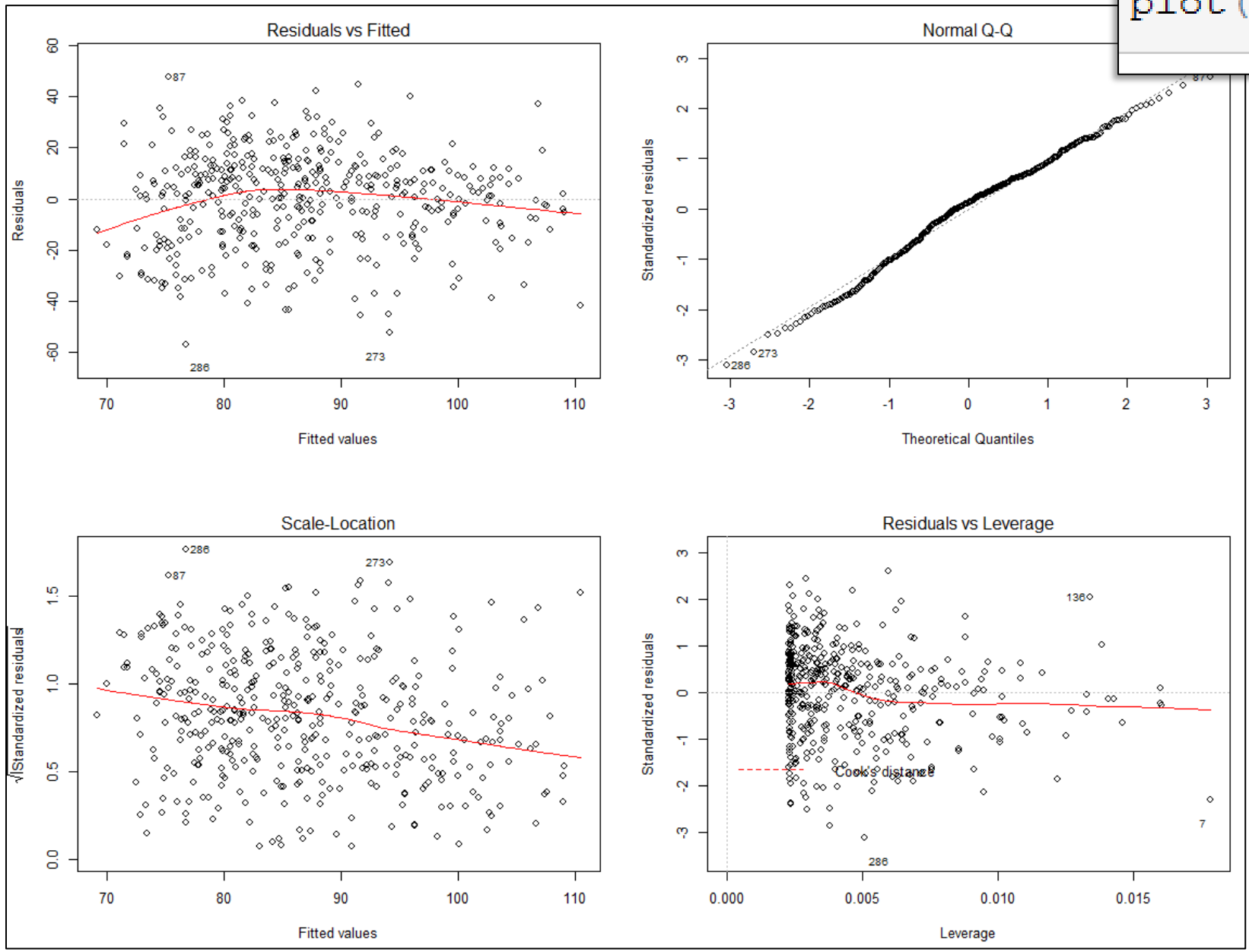
My Model

	Dependent variable:
	kid_score
	Child's Score
Mom's Score	0.610***
	(0.059)
Intercept	25.800***
	(5.920)
Observations	434
R^2	0.201
Adjusted R^2	0.199
Residual Std. Error	18.300 (df = 432)
F Statistic	109.000*** (df = 1; 432)
Note:	*p<0.1;* ***p<0.05;*** p<0.01

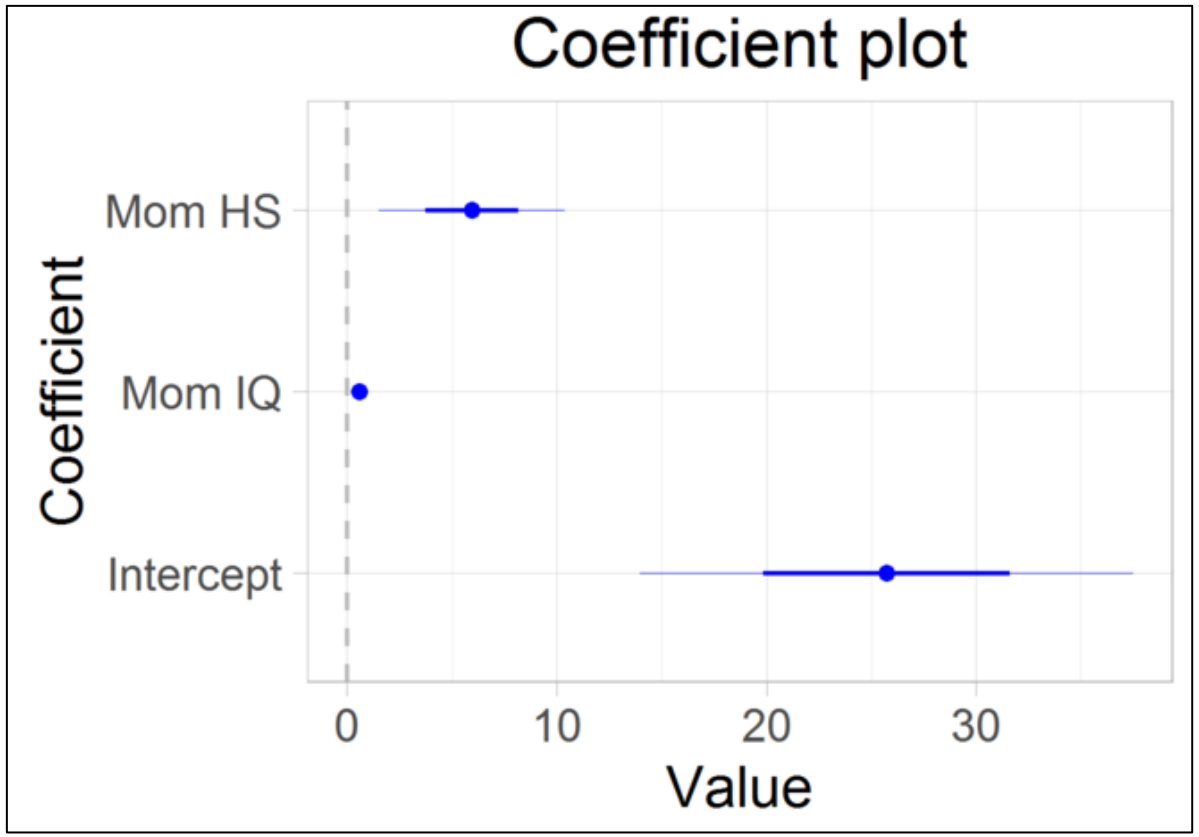


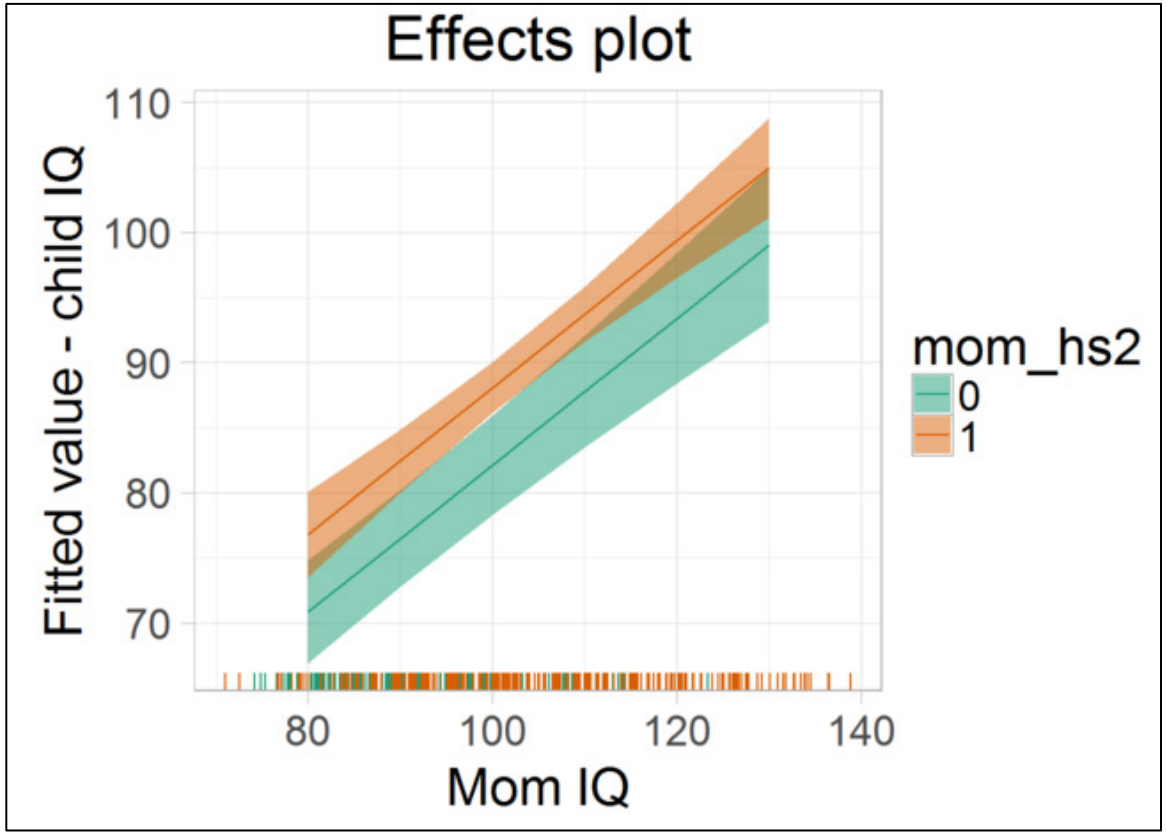
Model Diagnostics

```
plot(fit01)
```



Additional post-estimation plots





Logistic regression (and other GLMs)

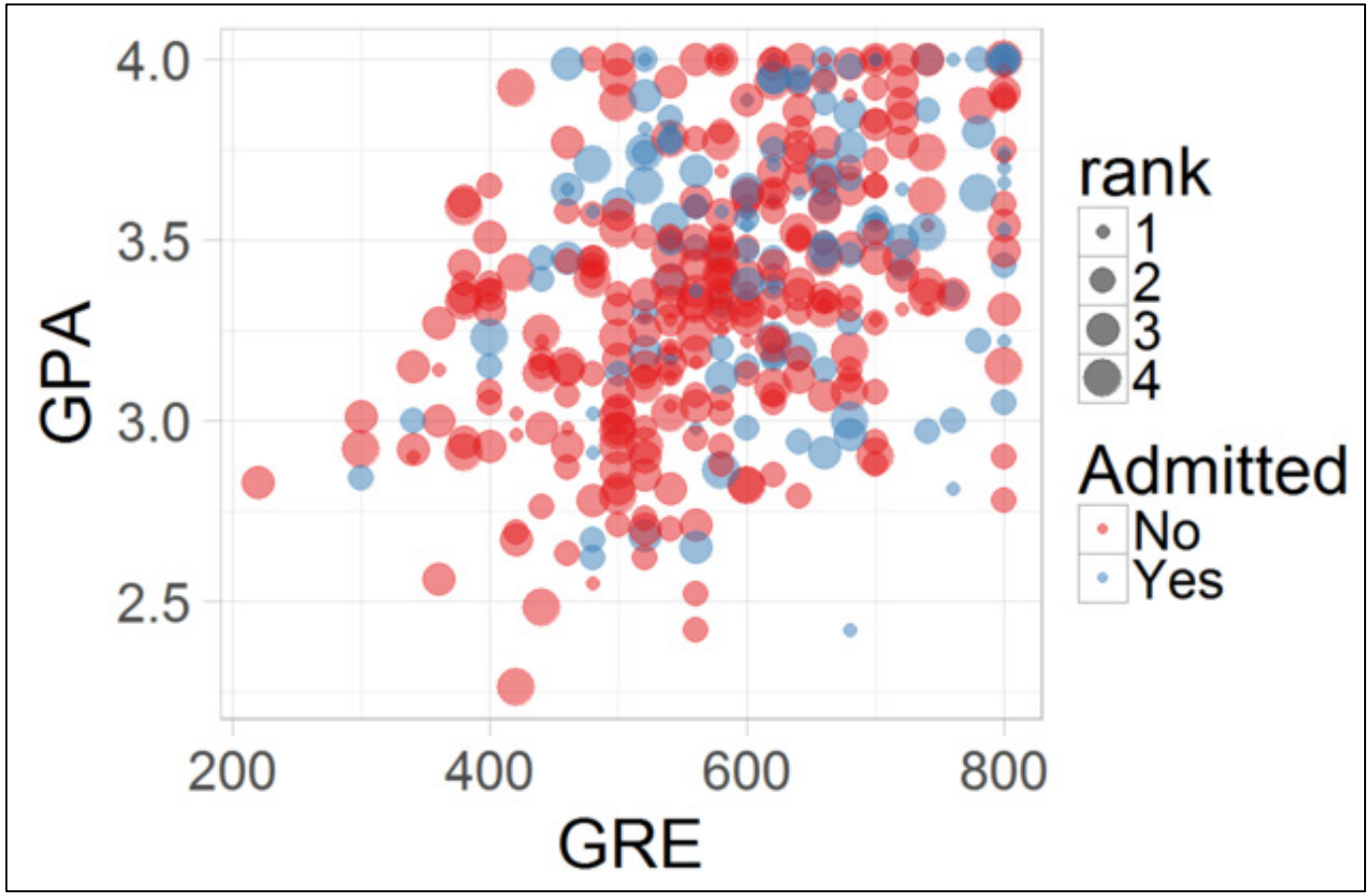
- …Are also easily implemented
- Example: graduate school admissions

```
head(mydata)
```

	admit	gre	gpa	rank
1	0	380	3.61	3
2	1	660	3.67	3
3	1	800	4.00	1
4	1	640	3.19	4
5	0	520	2.93	4
6	1	760	3.00	2

```
xtabs(~admit + rank
```

admit	rank 1	2	3	4
0	28	97	93	55
1	33	54	28	12



Data: http://www.ats.ucla.edu/stat/data/binary.csv

GLM, cont.

```
>> mylogit <- glm(admit ~ gre + gpa + ran
family = "binomial")
```

```
confint(mylogit)
```

```
                2.5 %    97.5 %
(Intercept) -6.271620 -1.79255
gre          0.000138  0.00444
gpa          0.160296  1.46414
rank2       -1.300889 -0.05675
rank3       -2.027671 -0.67037
rank4       -2.400027 -0.75354
```

```
exp(cbind(OR = coef(mylogit), confint(mylogit)))
```

```
                OR   2.5 % 97.5 %
(Intercept) 0.0185 0.00189  0.167
gre         1.0023 1.00014  1.004
gpa         2.2345 1.17386  4.324
rank2       0.5089 0.27229  0.945
rank3       0.2618 0.13164  0.512
rank4       0.2119 0.09072  0.471
```

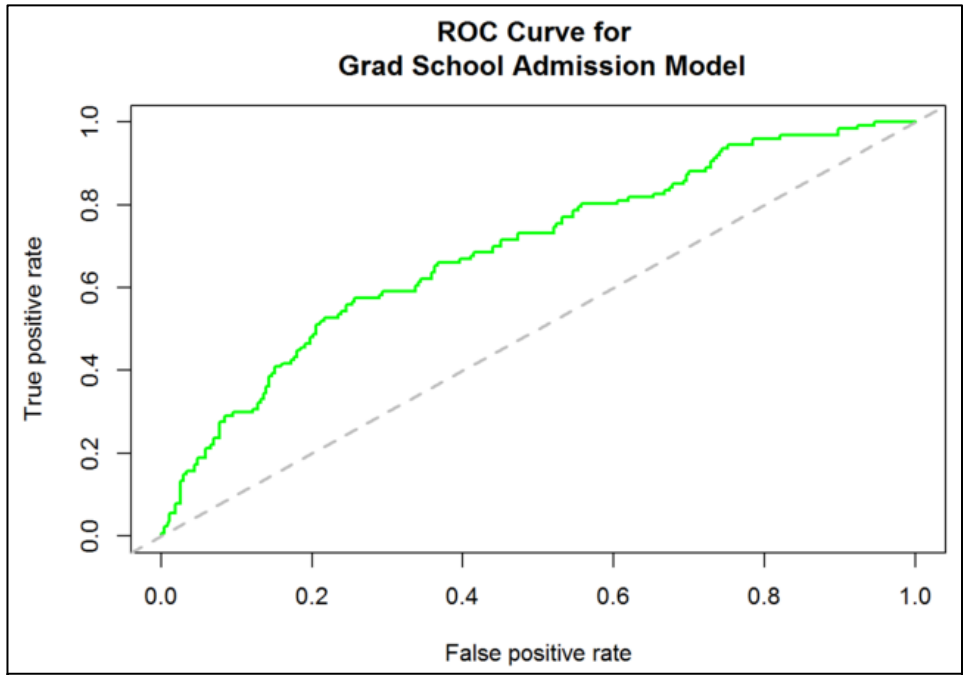
```
Call:
glm(formula = admit ~ gre + gpa + rank, family = "binomial",
    data = mydata)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.627  -0.866  -0.639   1.149   2.079

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.98998    1.13995   -3.50  0.00047 ***
gre          0.00226    0.00109    2.07  0.03847 *
gpa          0.80404    0.33182    2.42  0.01539 *
rank2       -0.67544    0.31649   -2.13  0.03283 *
rank3       -1.34020    0.34531   -3.88  0.00010 ***
rank4       -1.55146    0.41783   -3.71  0.00020 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.98  on 399  degrees of freedom
Residual deviance: 458.52  on 394  degrees of freedom
AIC: 470.5

Number of Fisher Scoring iterations: 4
```

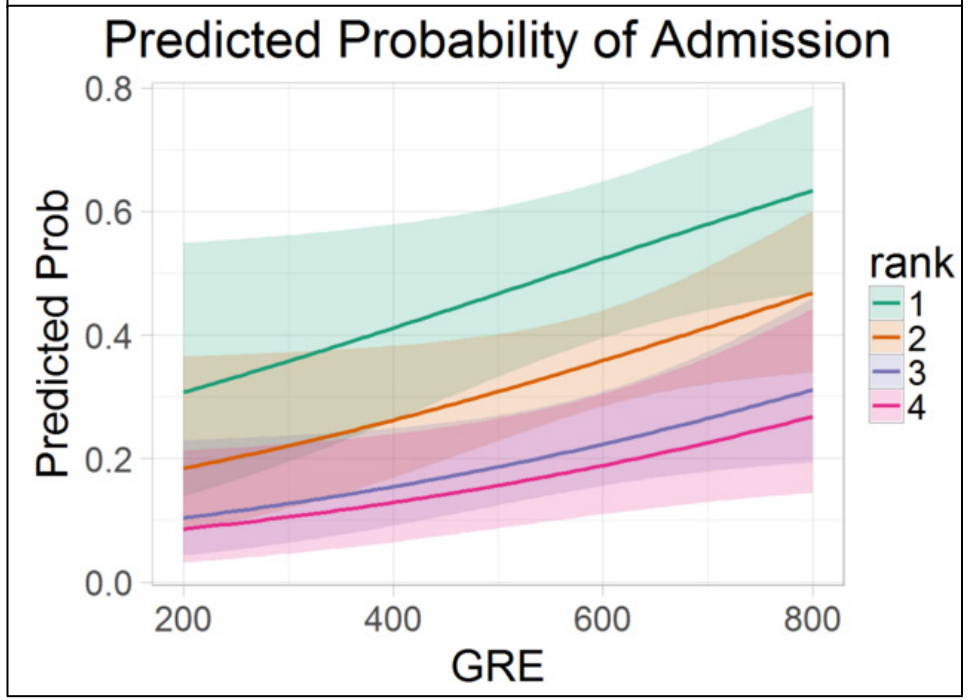
GLM, cont.

Grad School Admission Model

	Admitted
GRE	0.002**
	(0.001)
GPA	0.804**
	(0.332)
Rank=2	-0.675**
	(0.316)
Rank=3	-1.340***
	(0.345)
Rank=4	-1.550***
	(0.418)
Constant	-3.990***
	(1.140)
N	400
Log Likelihood	-229.000
AIC	471.000

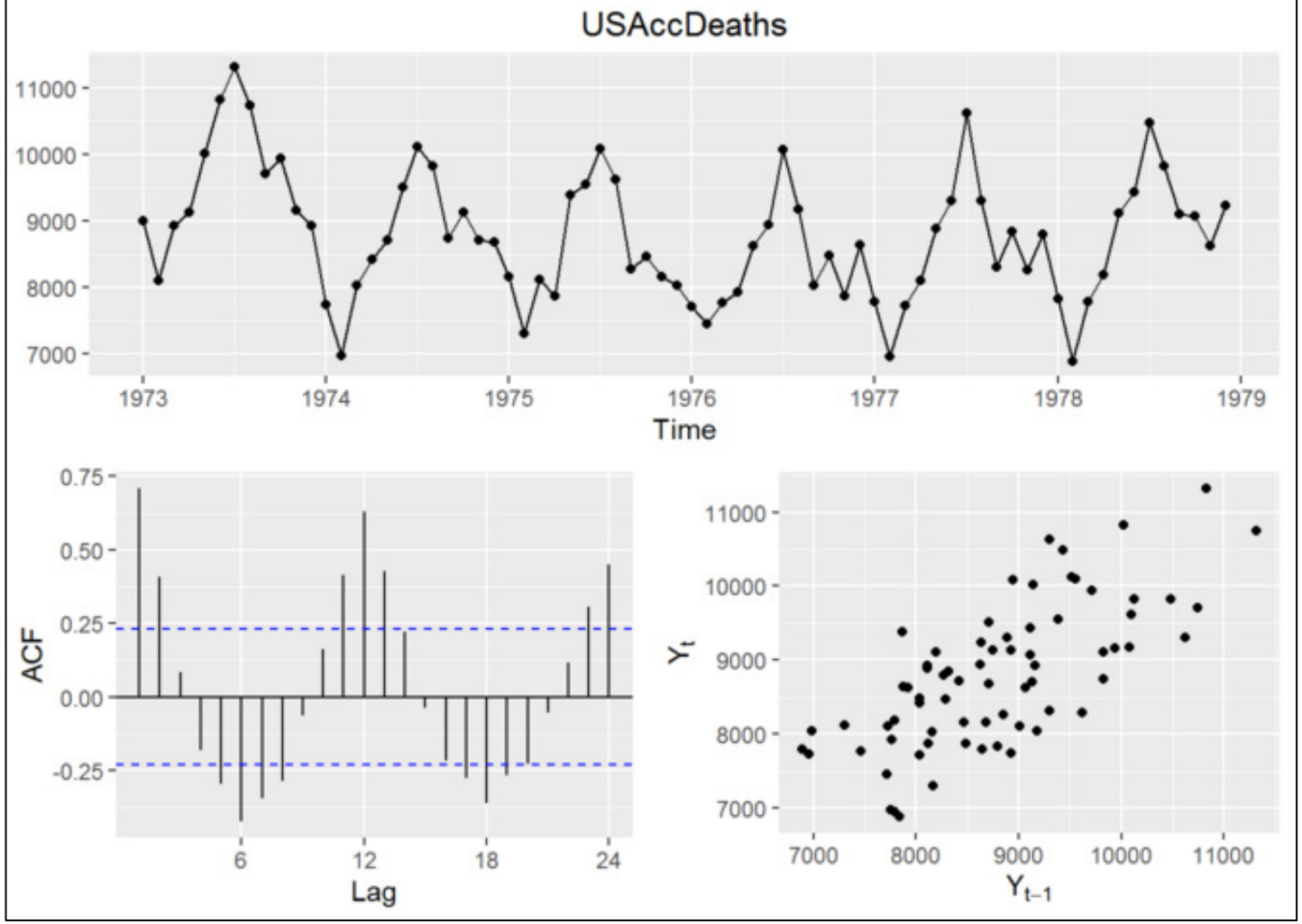
Note: ***p < .01;*** *p < .05;* p < .1
Rank = 1 omitted





Forecasting and prediction

```
>> ggtsdisplay(USAccDeaths, plot.type = "scatter")
```

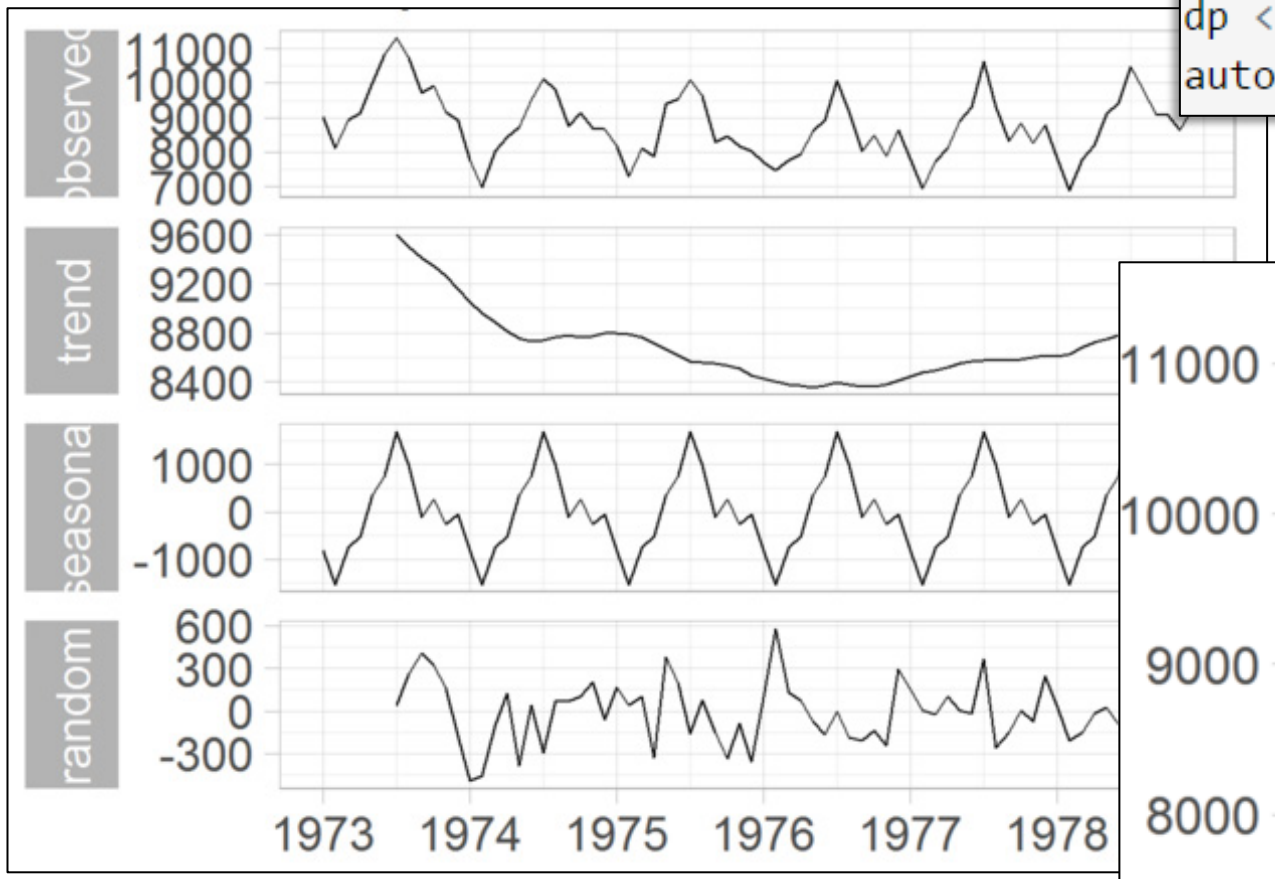


- Methods for time series as well as statistical models
- Seasonal decomposition, exponential smoothing
- Linear forecasts, MCMC, etc…

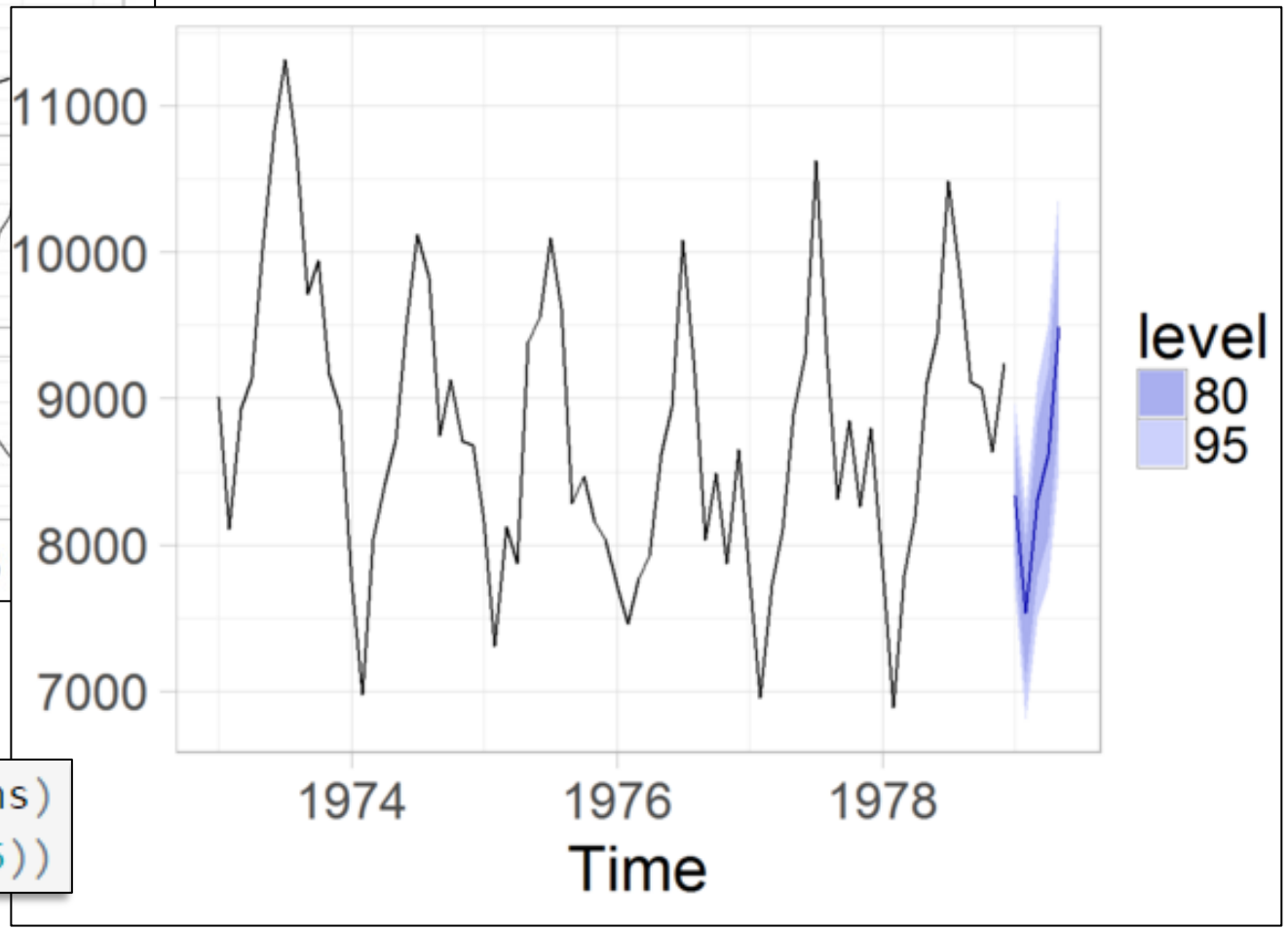
Data: US Accident deaths

Forecasting, prediction, time series

```
dp <- decompose(USAccDeaths)
autoplot(dp) + theme_light(base_size = 22)
```




```
fit <- auto.arima(USAccDeaths)
autoplot(forecast(fit, h = 5))
```

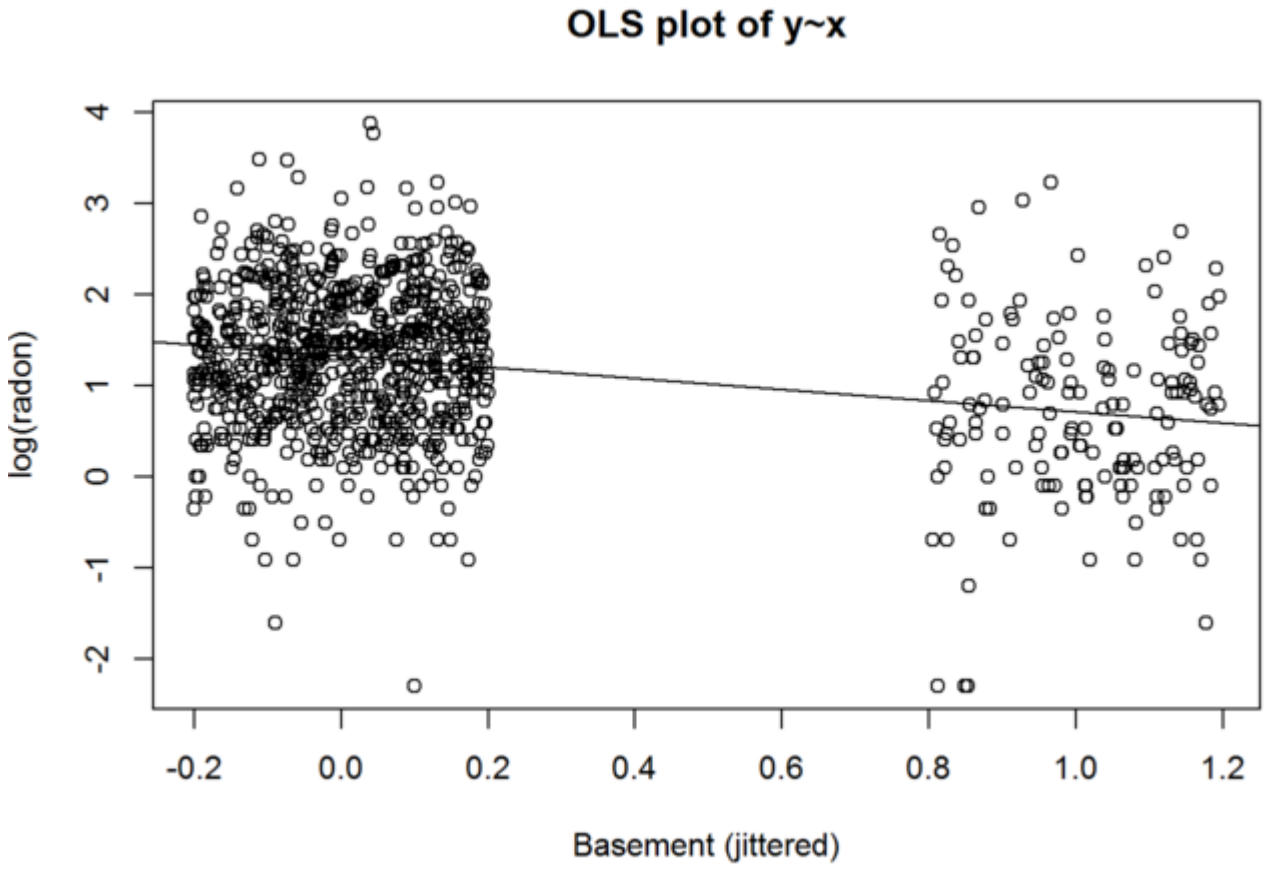


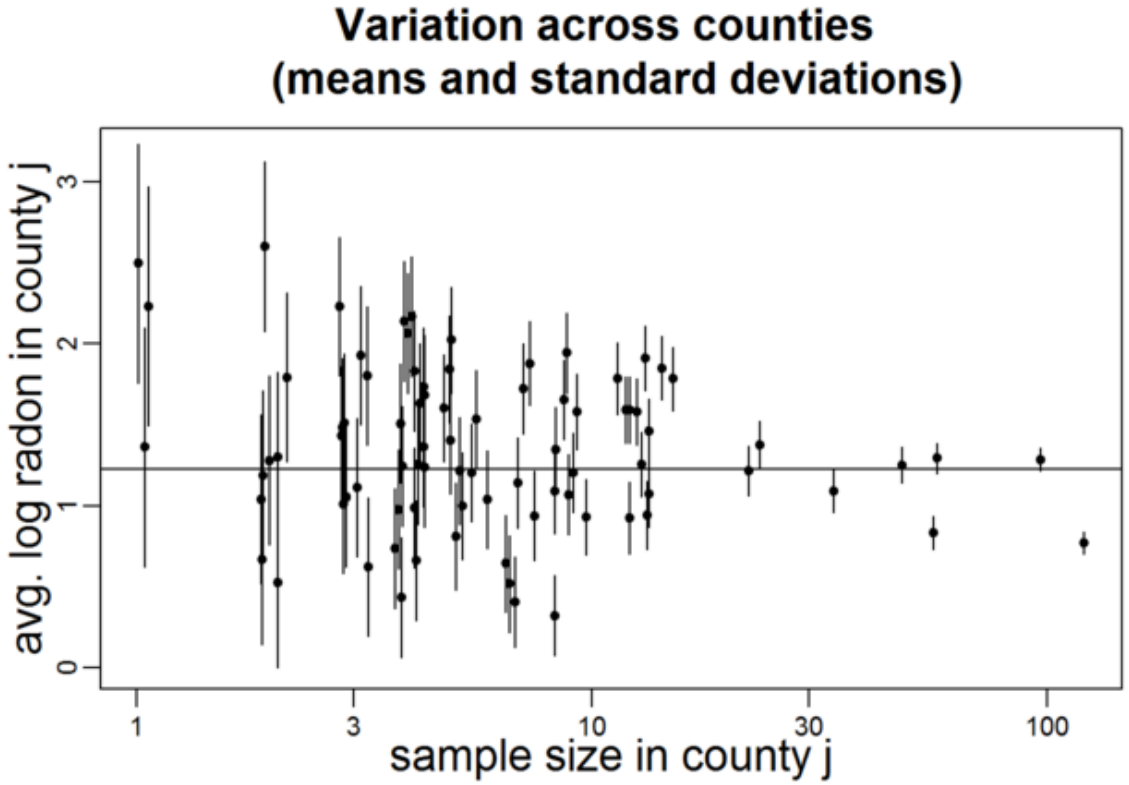
Hierarchical models (or nested, multilevel, etc…)

- MN radon data
  - Measurements are nested within 85 counties
  - Number of samples per county vary widely
  - Can model with varying slopes, intercepts, both

	log(radon)			
	OLS	linear mixed-effects		
	OLS	Varying intercepts	W/ group-level predictor	Varying slopes and intercepts
x	-0.613***	-0.693***	-0.668***	-0.671***
	(0.073)	(0.070)	(0.069)	(0.084)
u.full			0.720***	0.808***
			(0.092)	(0.091)
x:u.full				-0.420*
				(0.227)
Constant	1.330***	1.460***	1.470***	1.470***
	(0.030)	(0.052)	(0.038)	(0.035)
N	919	919	919	919
R^2	0.072			
Adjusted R^2	0.071			
Log Likelihood		-1,086.000	-1,067.000	-1,063.000
Residual Std. Error	0.823 (df = 917)			
F Statistic	70.900*** (df = 1; 917)			
AIC		2,179.000	2,144.000	2,143.000
BIC		2,199.000	2,168.000	2,181.000
Note:	***p < .01;*** *p < .05;* p < .1			

HLM example, cont.





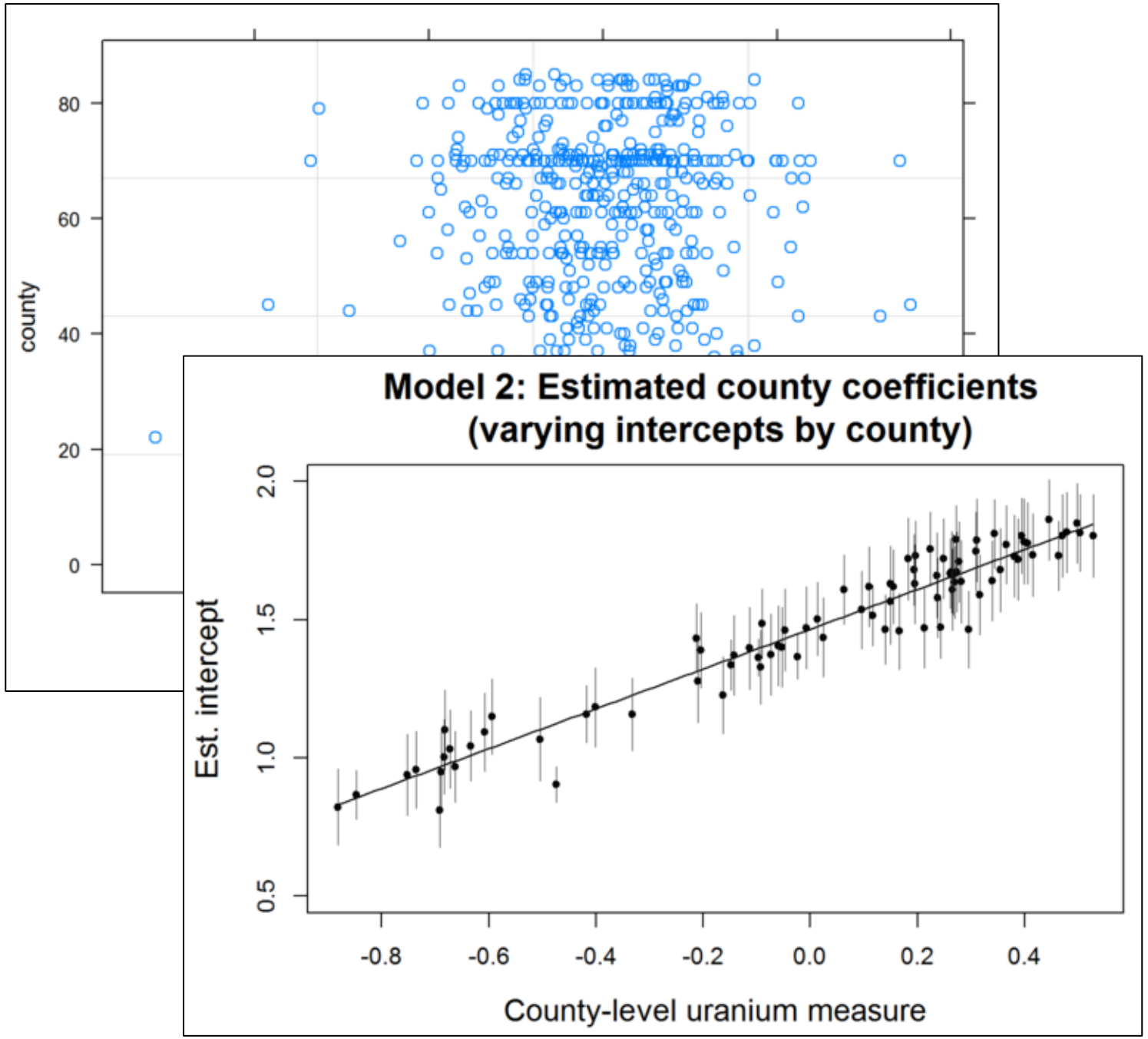
Perhaps try allowing county intercepts to vary

HLM example, cont.

- After running model, extract random effects from the fitted object

```
ranef(M2)

$county
   (Intercept)
1    -0.020642
2     0.011246
3     0.012422
4     0.111128
5     0.008236
```

- And examine the county-level residuals
- Or plot the individual county intercepts



Network analysis

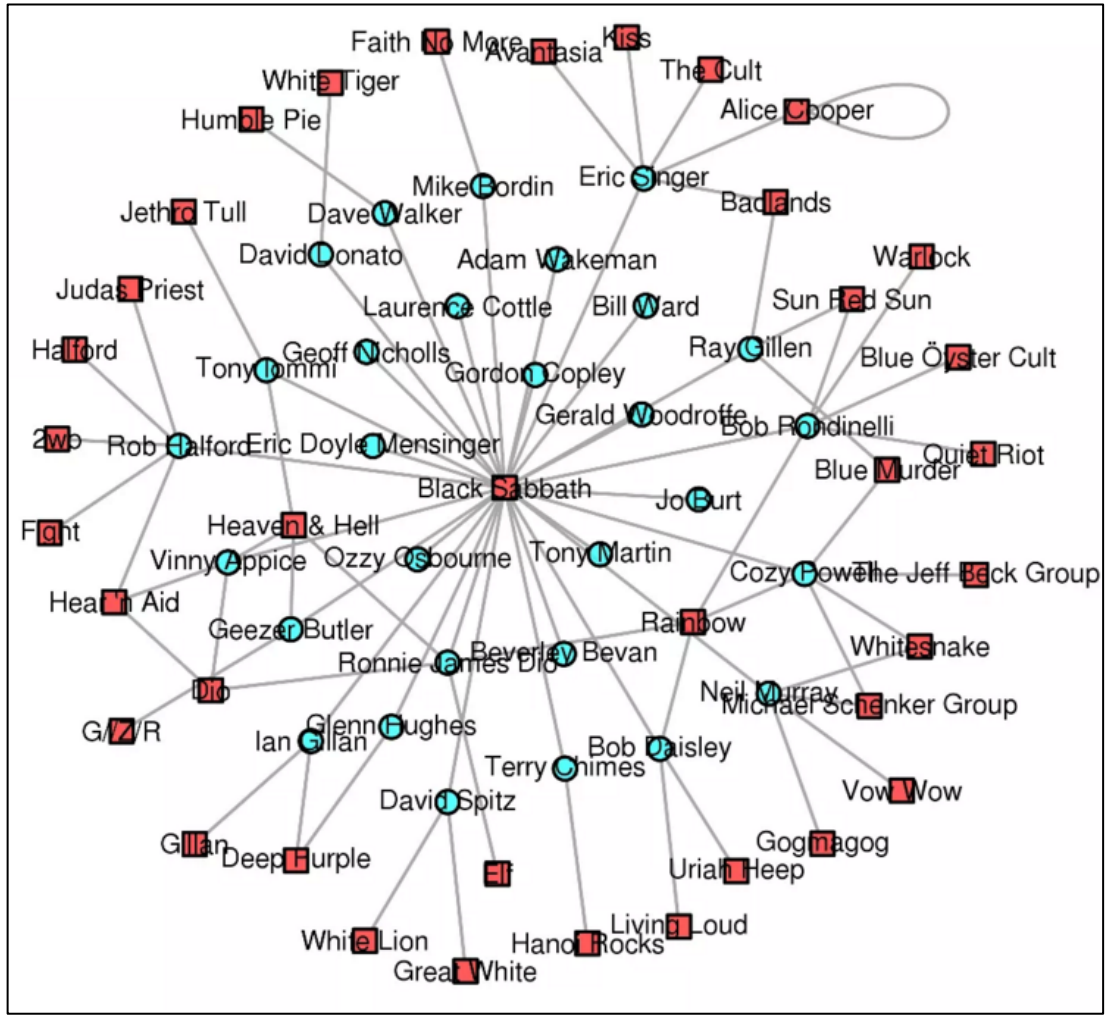
R has an expanding library of network analysis tools

- Descriptive statistics
- Exponential Random Graph Modeling for inferential models
- Visualizations

Co-occurrence network of country names in statements by US senators

(Katherine Ognyanova, kateto.net/countries)

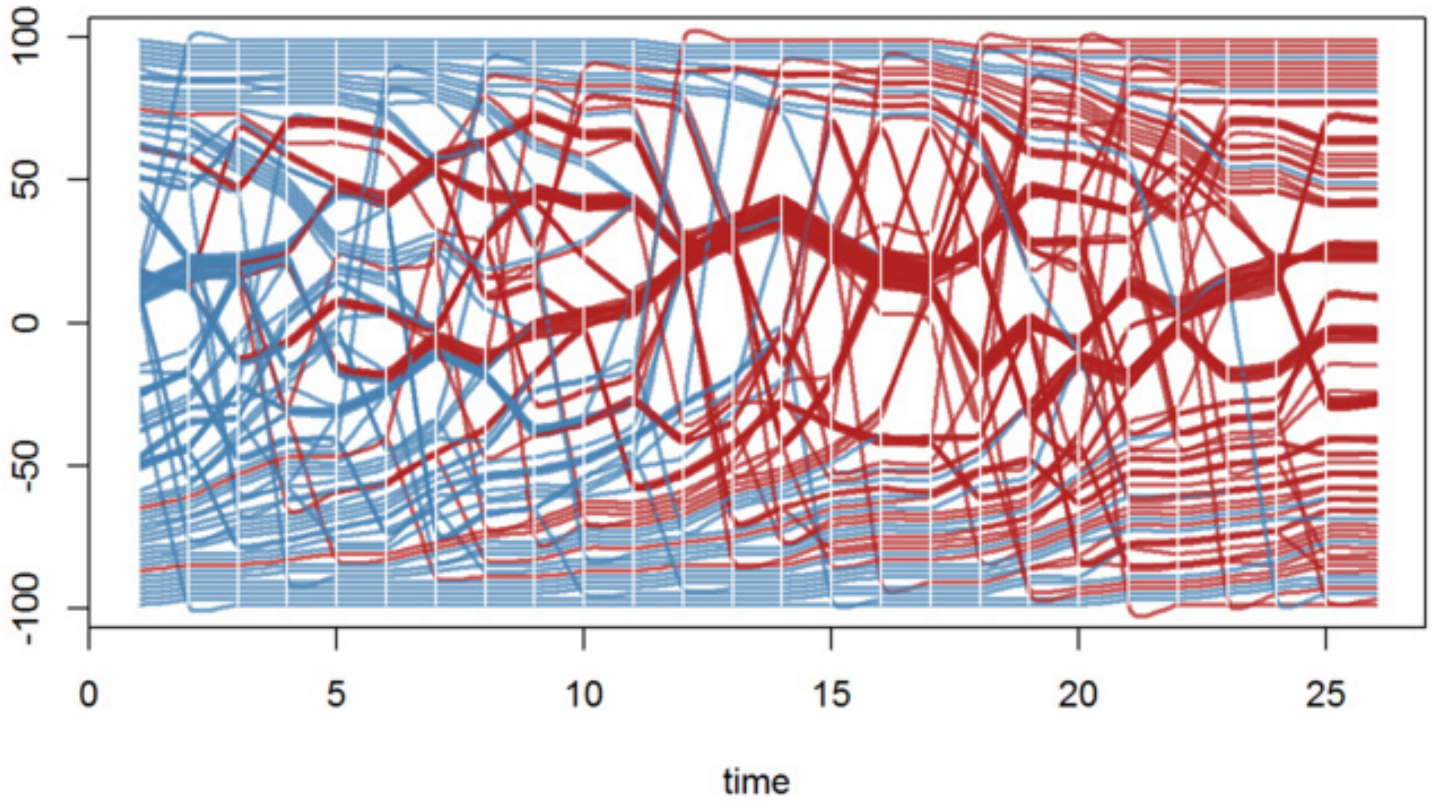
Black Sabbath ego network

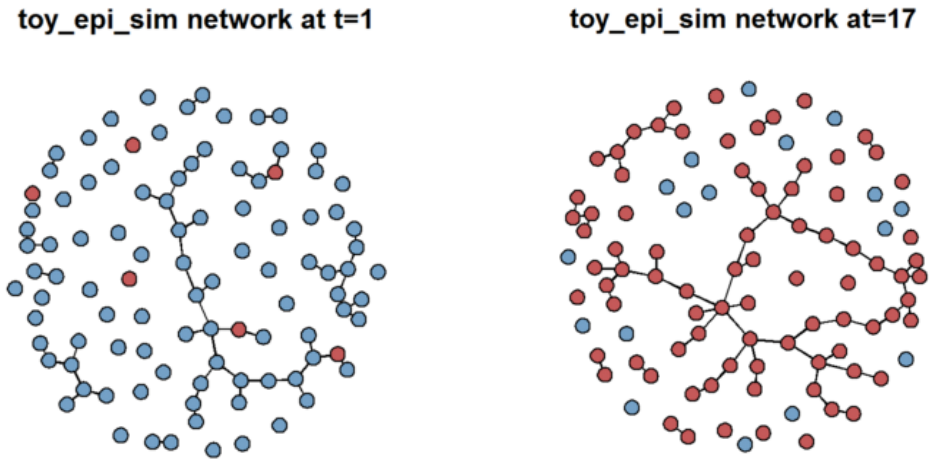


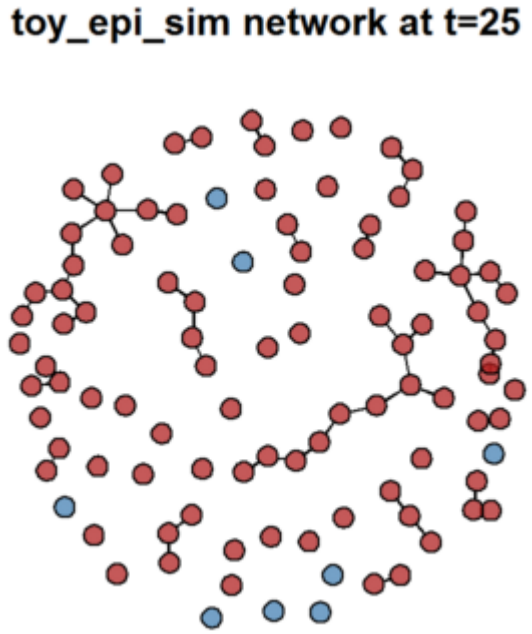
http://www.r-bloggers.com/patterns-in-the-ivy-the-small-world-of-metal/

SNA example: simulated disease spread

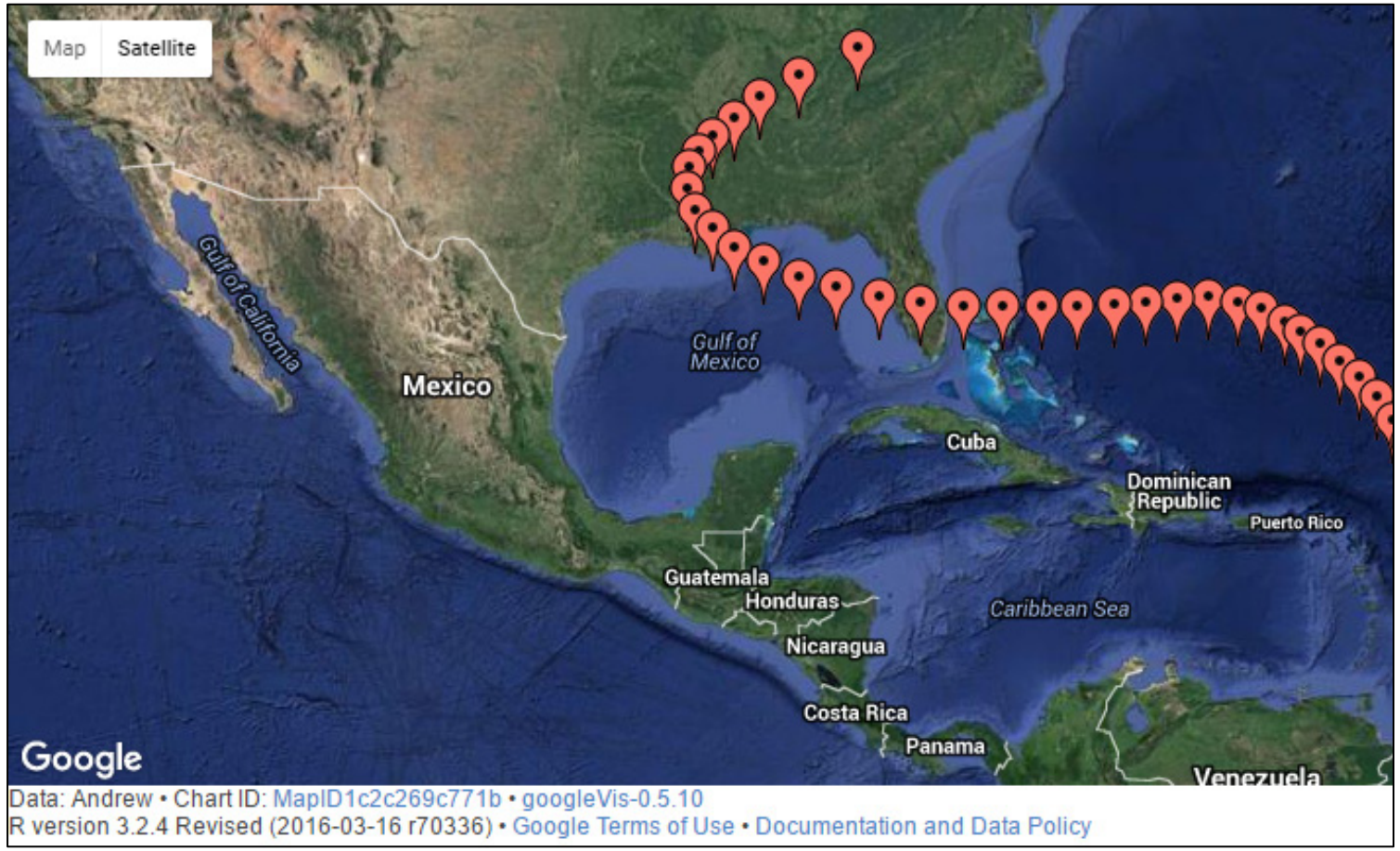
- Simulation
- Temporal network
  - Red = infected
  - Blue = not yet infected





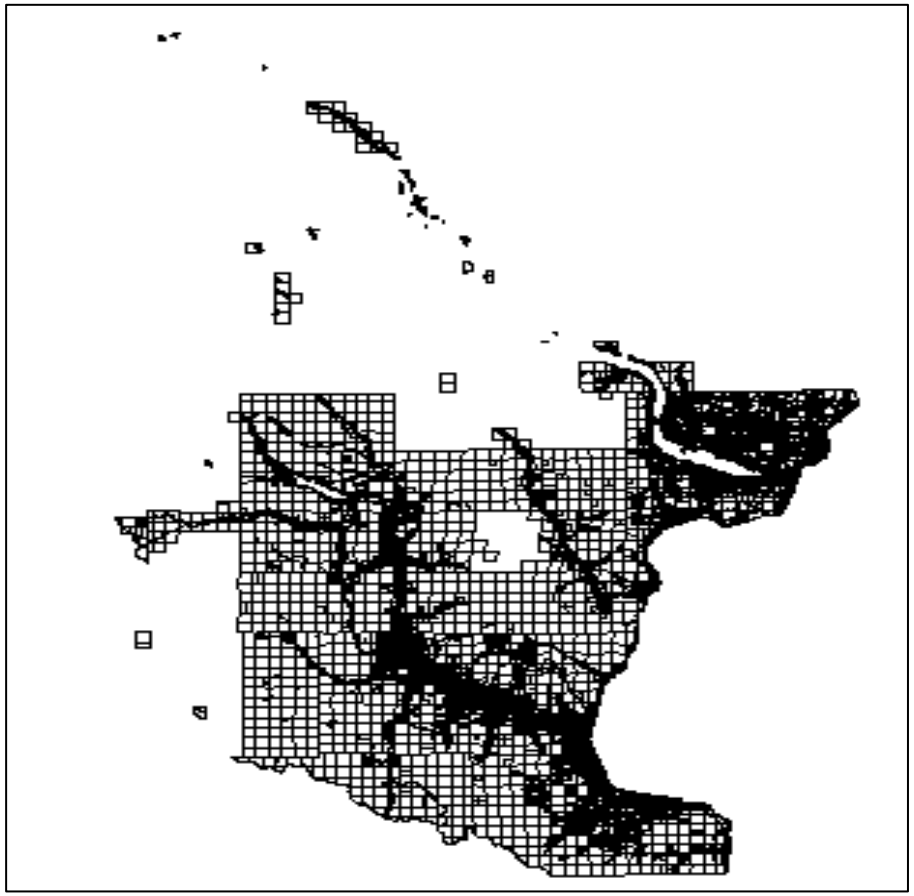


Mapping



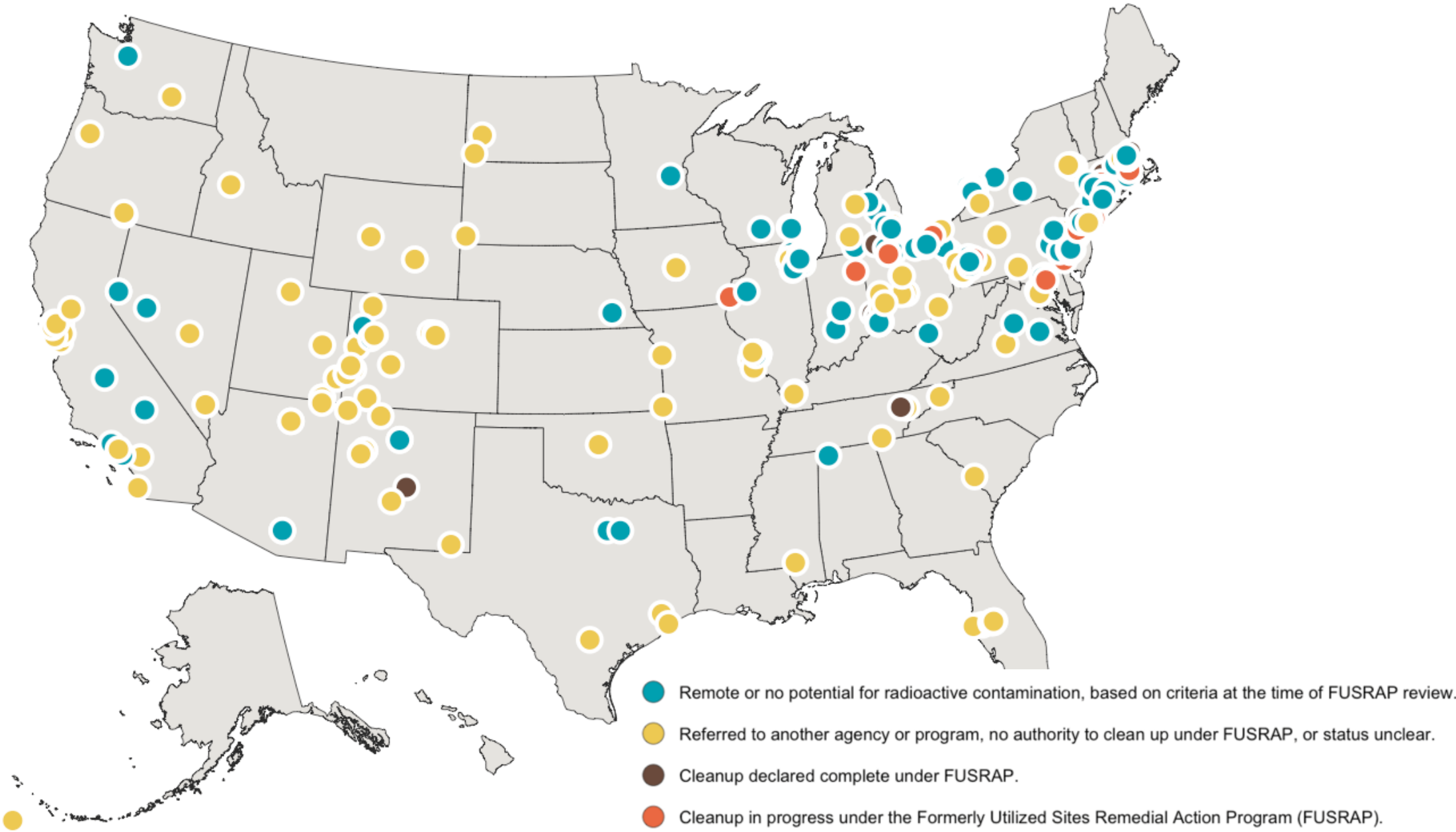
Google style

Base R



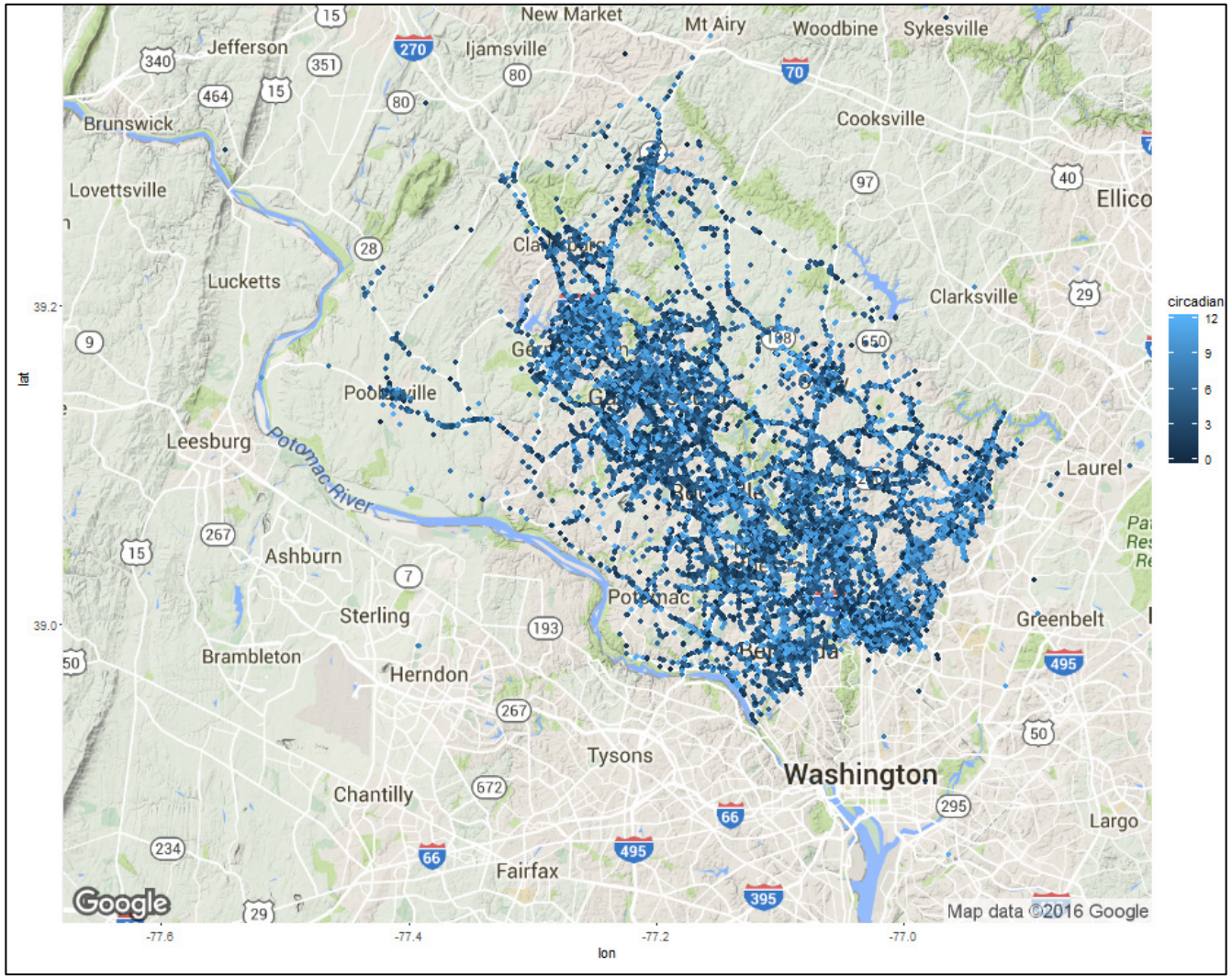
Waste Lands: America's Forgotten Nuclear Legacy

Washington Post



Integrate with data systems

- Download 800k records of Montgomery Co traffic violations
- Stored in a relational database
- Connect to database within R, query using SQL syntax
- Visualize with a Google Maps-style plot

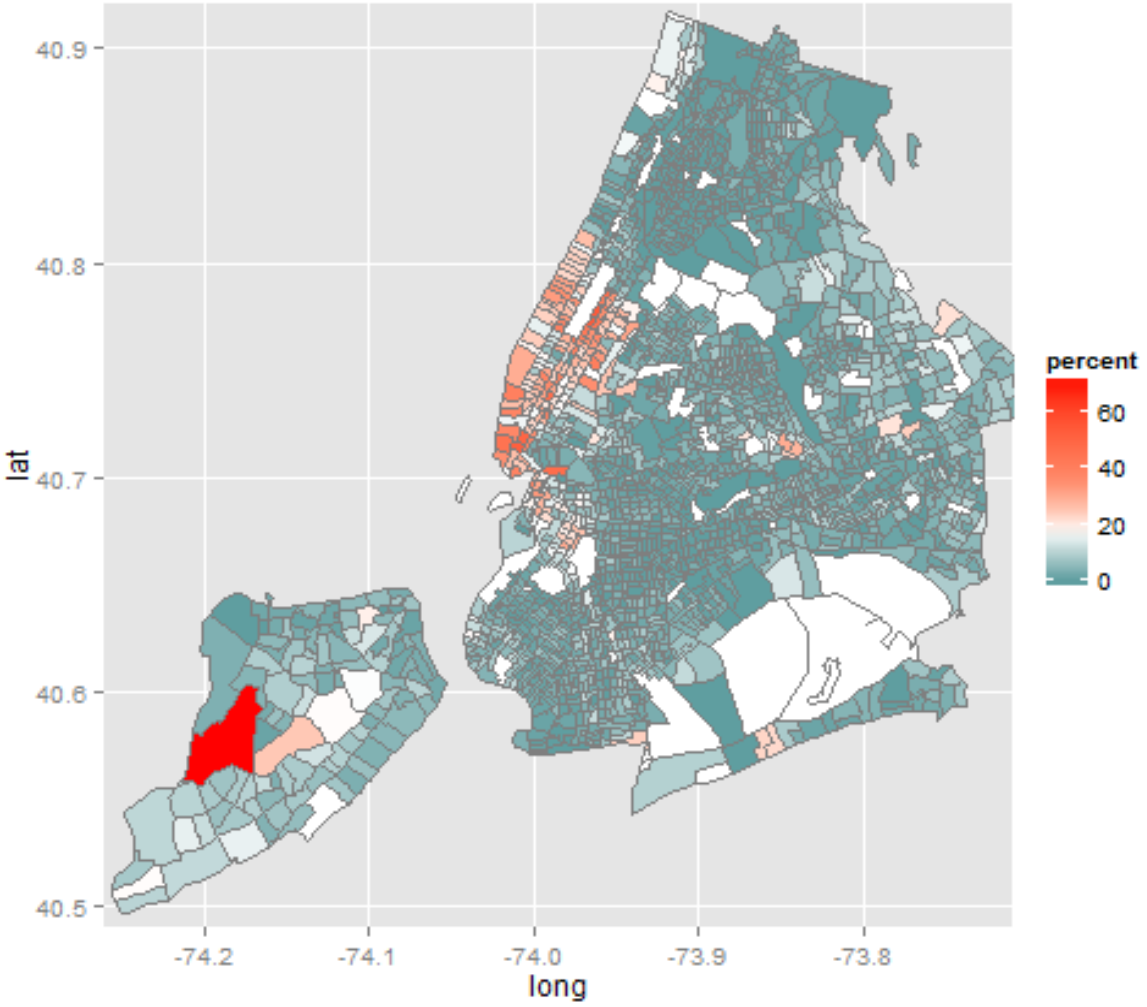


Traffic violations (shade = time of day)

Source: Srini Kumar, Director of Data Science at Microsoft

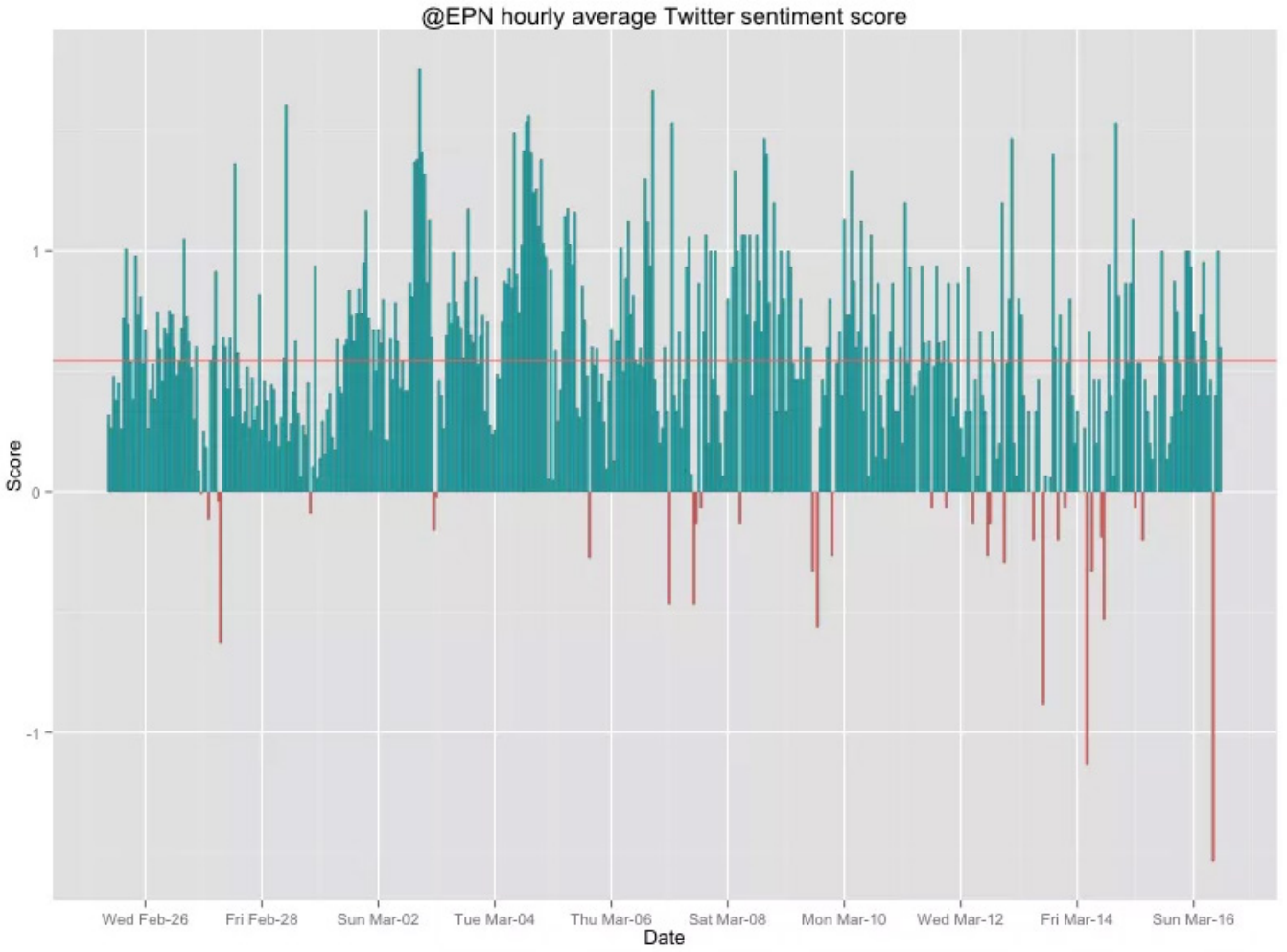
Access APIs

Data: *American Community Survey*, % income over $200k



Source: zevross.com/blog

Data: *Twitter*

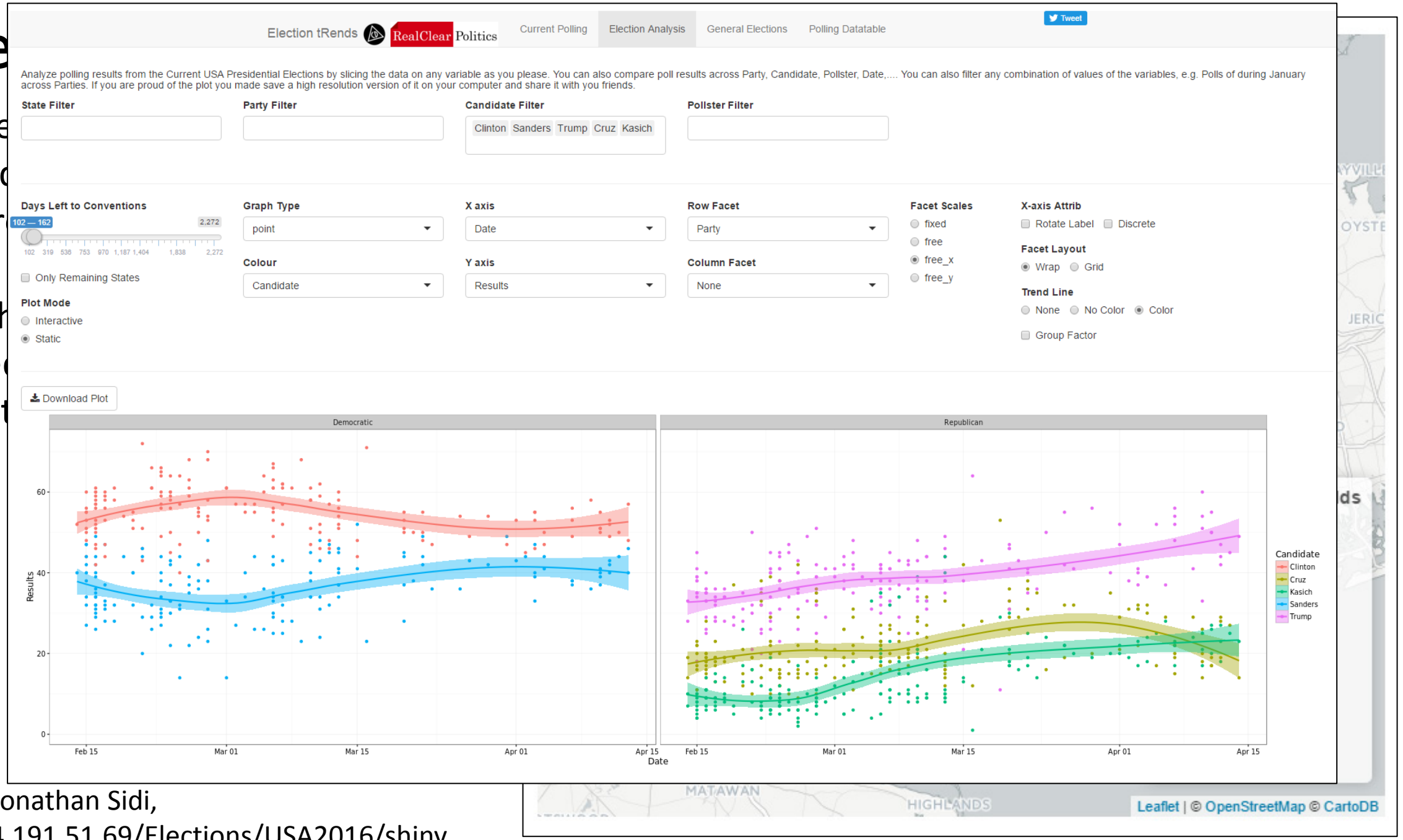


Source: Joe Gonzalez, http://www.r-bloggers.com/how-popular-is-the-president-of-mexico-on-twitter/

Inte

- "Le
wic
dir

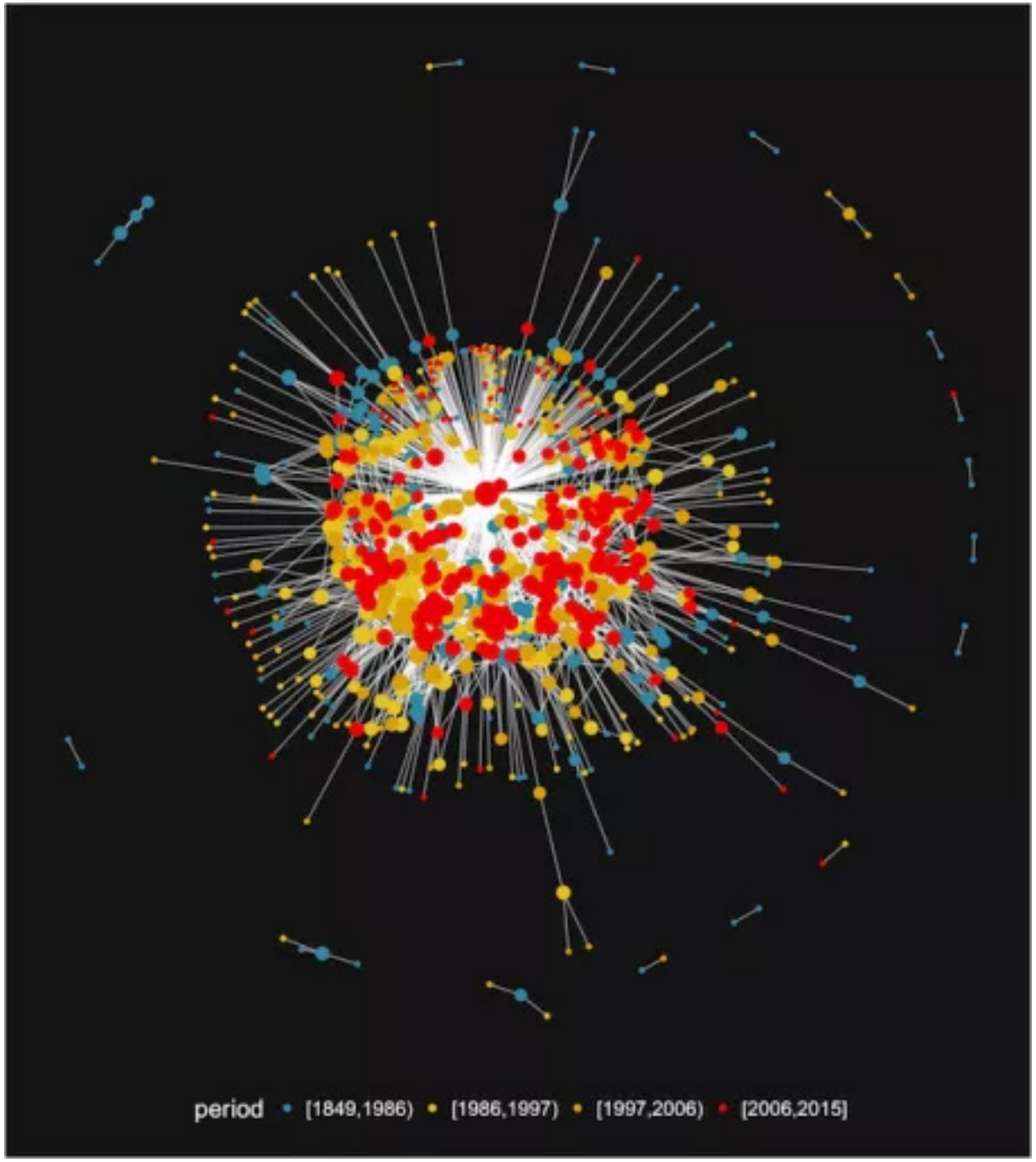
- "Sh
fle
dat



Source: Jonathan Sidi, http://54.191.51.69/Elections/USA2016/shiny

More visualization

- Mashup of text data with a network visualization
  - Text: Icelandic legal code
  - Nodes: 1,513 documents
  - Links: cross-references
  - Size: number of links
  - Color: time period



Source: Françoisn, http://www.r-bloggers.com/ggnetwork-network-geometries-for-ggplot2/

Putting it together

- R is a language
  - And a large analytical toolbox

- Best-use leverages multiple strengths
  - Use an IDE;
  - Create readable, reproducible code chunks;
  - Use markdown to author reports and presentations
    - (almost) all of the code for this presentation was created with markdown -> HTML output

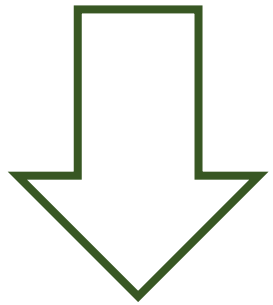
```
# detach("package:xtable", unload=TRUE)
```

A formula looks like this:

$$
\huge{Y \sim Var_1 + Var_2 + \dotso}
$$

Factor analysis

```{r factor_analysis, compress=TRUE}
# First I need to convert the factor variables to numeric
x <- data.frame(lapply(gss08[,names(gss08)], function(x) a
```



A formula looks like this:

$$Y \sim Var_1 + Var_2 + \dots$$

Factor analysis

```
>> # First I need to convert the factor variables to numeric
>> x <- data.frame(lapply(gss08[, names(gss08)], function(x)
```

HTML5 presentations

Preview

```
1  ```{r setup, include=FALSE}
2  library(knitr)
3  opts_chunk$set(cache=TRUE, echo=FALSE, messages=FALSE, warning=FALSE, tid
4  ```
5
6  r-pres example
7  PNIAF
8  ========================================================
9  author: Zane Kelly
10 date: 4/27/2016
11
12 First Slide
13 ========================================================
14
15 - Author presenations directly in RStudio
16 - And preview on the fly
17
18
19 Embed plots, control appearance
20 ========================================================
21
22 left: 70%
23 ```{r}
24 ggplot(diamonds) +
25    geom_point(aes(x = carat, y = price, color = cut)) +
26    theme_bw(base_size = 18) +
27    scale_color_brewer(palette = "Set1") +
28    ggtitle("Diamonds - cut, carat, price") +
29    xlab("Carat") +
30    ylab("Price") +
31    labs(color = "Cut")
32 ```
33
34 ***
35
36    Desc: A colorful plot of diamond prices by carats
37
```

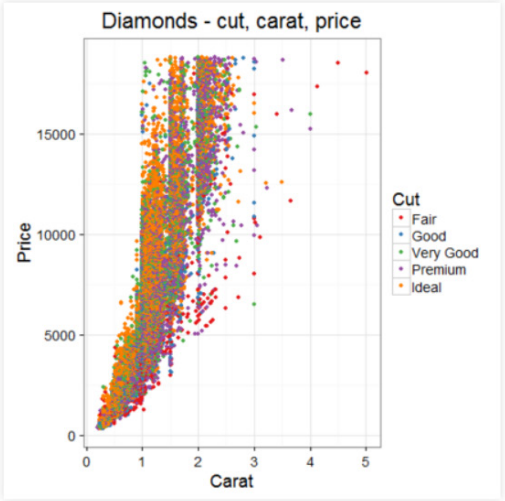
PNIAF

First Slide

- Author presenations directly in RStudio
- And preview on the fly

Embed plots, control appearance

left: 70%



Desc: A colorful plot of diamond prices by carats

Questions?

Contact Info

Zane Kelly

Research Analyst

Joint Legislative Audit and Review Committee

Washington State Legislature

360-786-5193

Zane.Kelly@leg.wa.gov

Suggested R References

- The R Project & CRAN
- r-bloggers
- Rstudio & Shiny
- rseek.org (custom Google search)
- The R Book, Michael J. Crawley (older edition can be found online)
- The R Cookbook
- ggplot2
- The R Journal
- Quick-R
- Advanced R
- Revolution analytics (now MSFT, their getting started links are here)