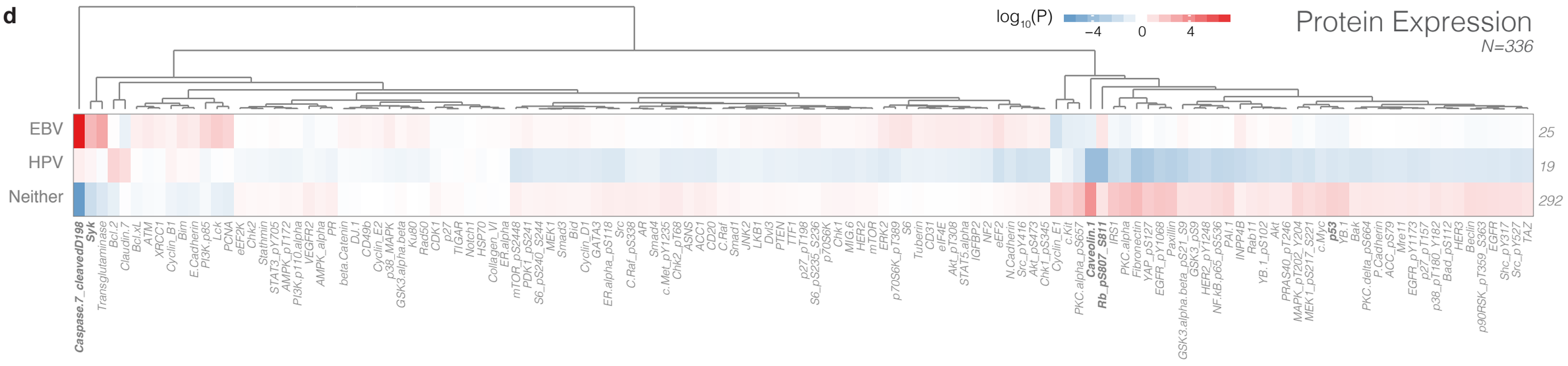
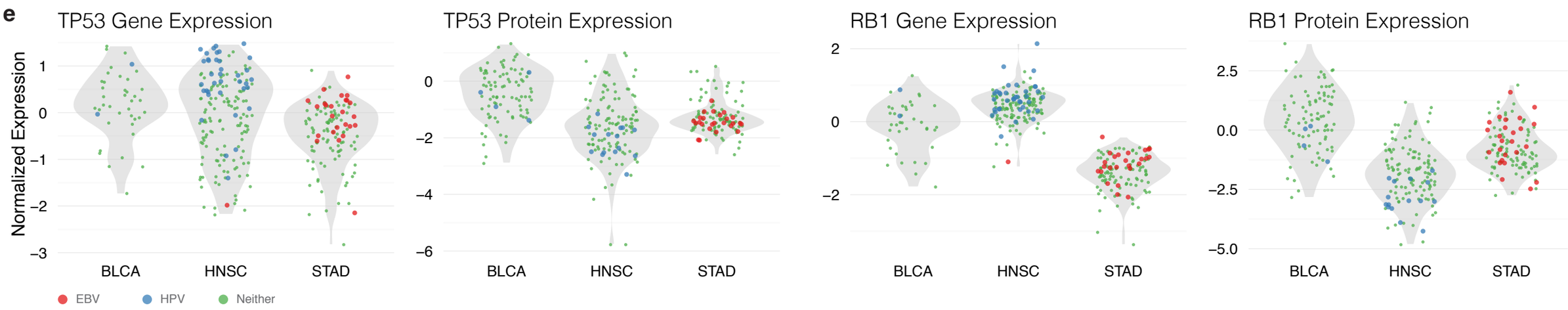
Data Visualization with ggplot2

Matt Wyczalkowski
February 8, 2018

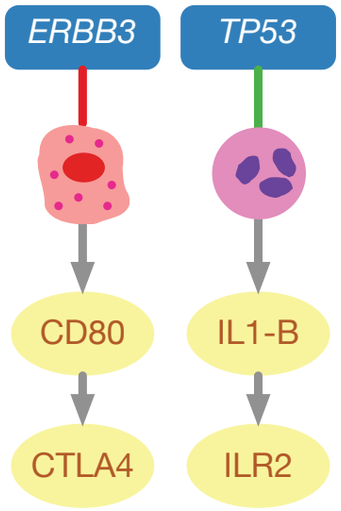
d


log10(P)
−4
0
4
Protein Expression
N=336
EBV
HPV
Neither
25
19
292
Caspase.7_cleavedD198
Syk
Transglutaminase
Bcl.2
Claudin.7
Bcl.xL
ATM
XRCC1
Cyclin_B1
Bim
E.Cadherin
PI3K.p85
Lck
PCNA
eEF2K
Chk2
Stathmin
STAT3_pY705
AMPK_pT172
PI3K.p110.alpha
VEGFR2
AMPK_alpha
PR
beta.Catenin
DJ.1
CD49b
Cyclin_E2
p38_MAPK
GSK3.alpha.beta
Ku80
Rad50
CDK1
p27
TIGAR
Notch1
HSP70
Collagen_VI
ER.alpha
mTOR_pS2448
PDK1_pS241
S6_pS240_S244
MEK1
Smad3
Bid
Cyclin_D1
GATA3
ER.alpha_pS118
Src
C.Raf_pS338
AR
Smad4
c.Met_pY1235
Chk2_pT68
ASNS
ACC1
CD20
C.Raf
Smad1
JNK2
LKB1
Dvl3
PTEN
TTF1
p27_pT198
S6_pS235_S236
p70S6K
Chk1
MIG.6
HER2
mTOR
ERK2
p70S6K_pT389
S6
Tuberin
CD31
eIF4E
Akt_pT308
STAT5.alpha
IGFBP2
NF2
eEF2
N.Cadherin
Src_pY416
Akt_pS473
Chk1_pS345
Cyclin_E1
c.Kit
PKC.alpha_pS657
Caveolin.1
Rb_pS807_S811
IRS1
PKC.alpha
Fibronectin
YAP_pS127
EGFR_pY1068
Paxillin
GSK3.alpha.beta_pS21_S9
GSK3_pS9
HER2_pY1248
NF.kB.p65_pS536
PAI.1
INPP4B
Rab11
YB.1_pS102
Akt
PRAS40_pT246
MAPK_pT202_Y204
MEK1_pS217_S221
c.Myc
p53
YB.1
Bak
PKC.delta_pS664
P.Cadherin
ACC_pS79
Mre11
EGFR_pY1173
p27_pT157
p38_pT180_Y182
Bad_pS112
HER3
Beclin
p90RSK_pT359_S363
EGFR
Shc_pY317
Src_pY527
TAZ

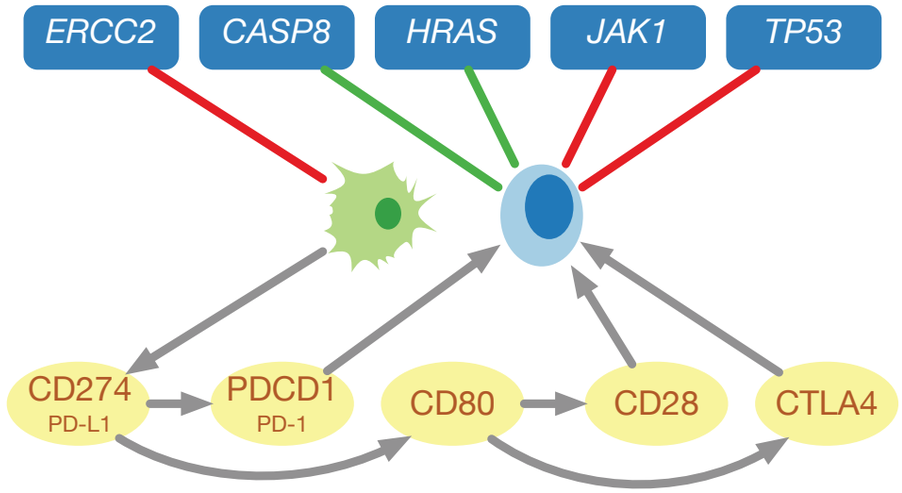
e


TP53 Gene Expression
Normalized Expression
1
0
−1
−2
−3
BLCA
HNSC
STAD
EBV
HPV
Neither
TP53 Protein Expression
0
−2
−4
−6
BLCA
HNSC
STAD
RB1 Gene Expression
2
0
−2
BLCA
HNSC
STAD
RB1 Protein Expression
2.5
0.0
−2.5
−5.0
BLCA
HNSC
STAD

C1: Wound healing



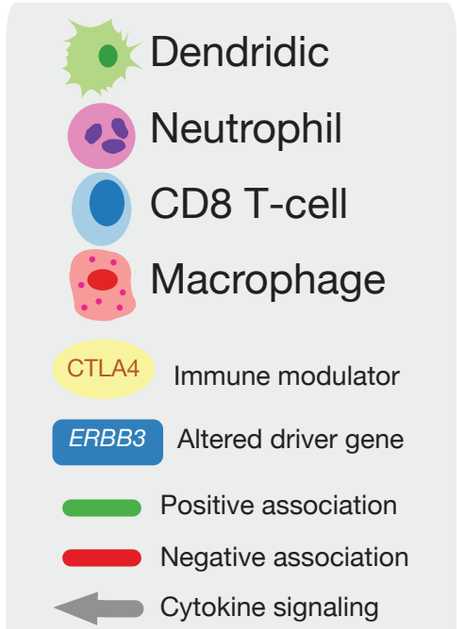
C2: IFN-γ dominant

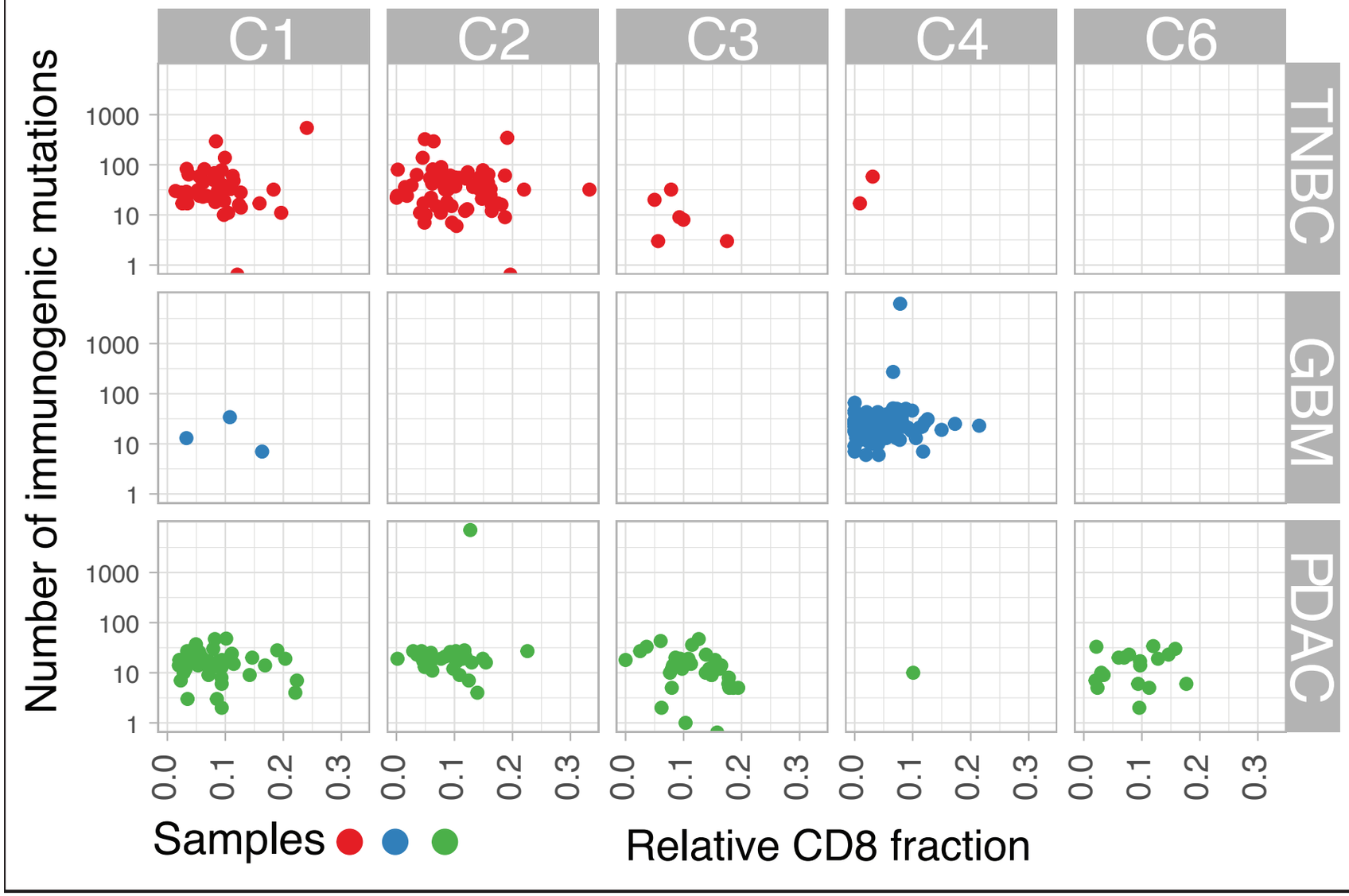


C6

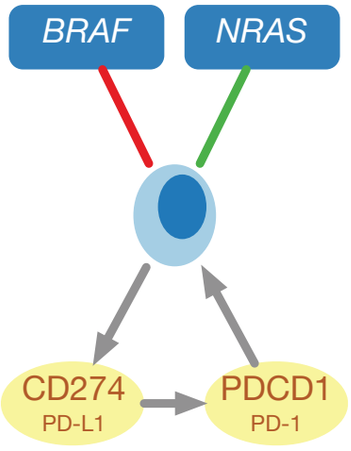
Insufficient Data

TGF-β dominant

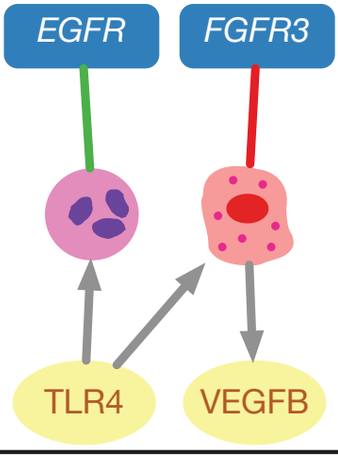


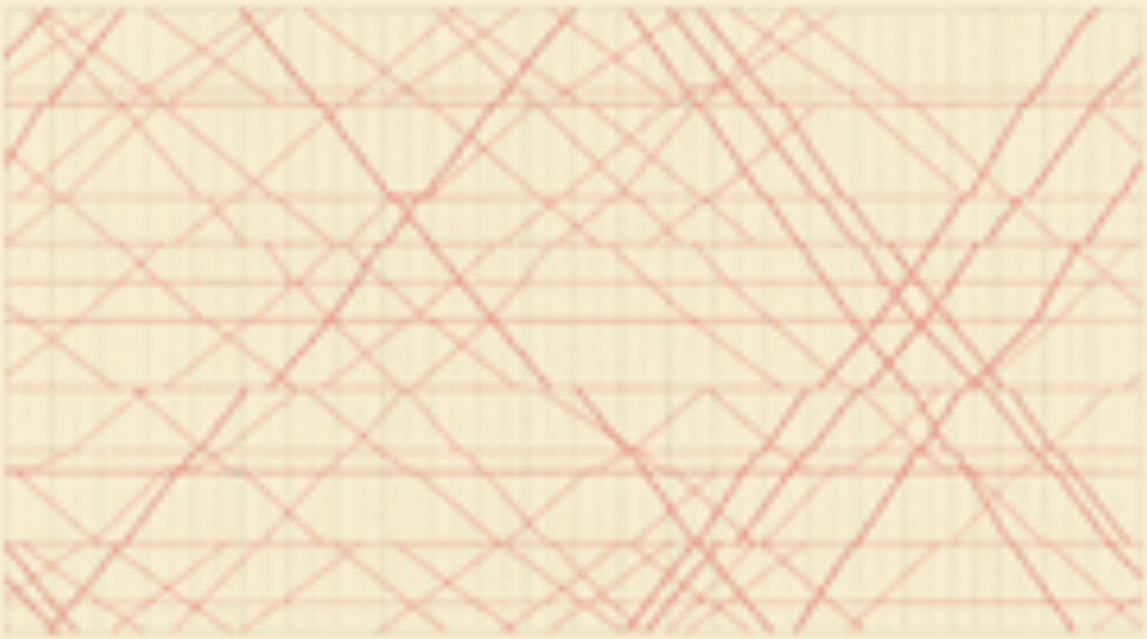


C3: Inflammatory



C4: Lymphocyte depleted





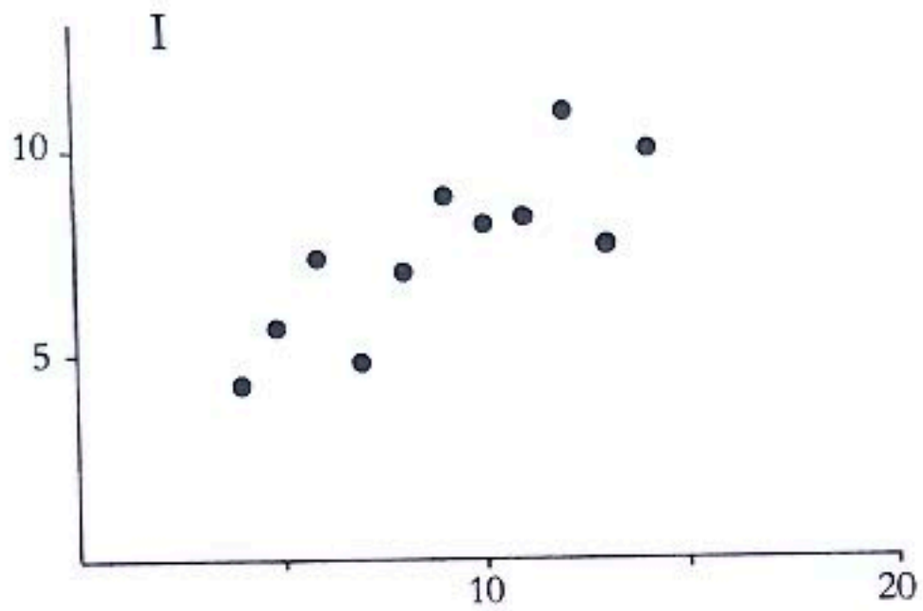
SECOND EDITION

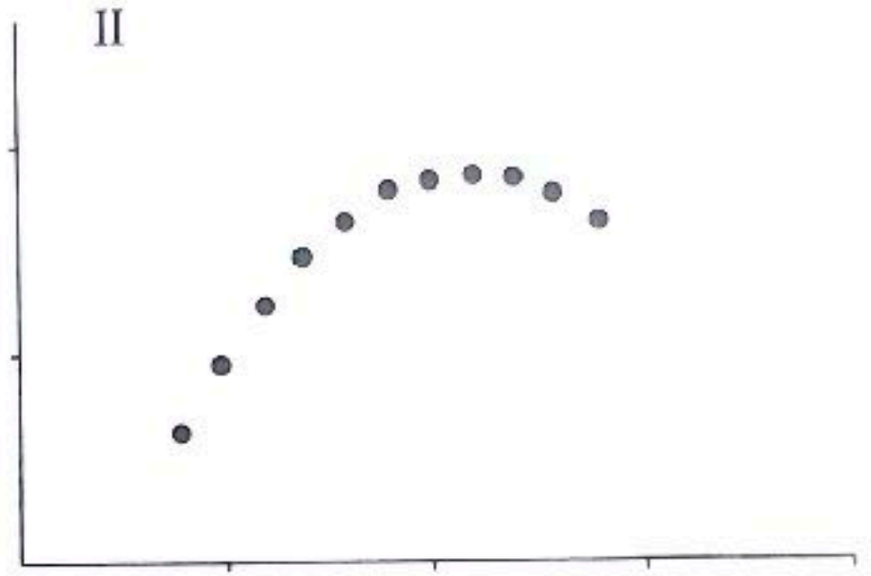
The Visual Display of Quantitative Information

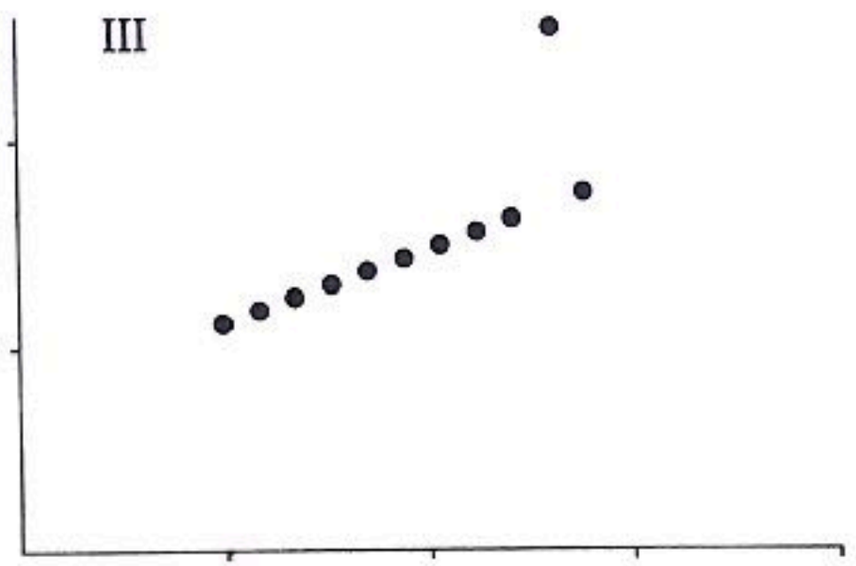
EDWARD R. TUFTE

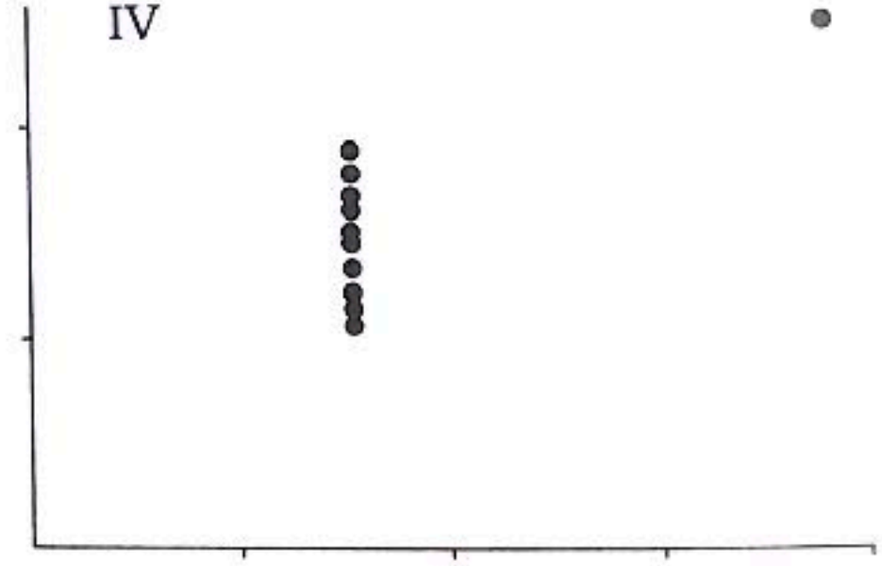
I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

N = 11
mean of X's = 9.0
mean of Y's = 7.5
equation of regression line: $Y = 3 + 0.5X$
standard error of estimate of slope = 0.118
t = 4.24
sum of squares $X - \overline{X} = 110.0$
regression sum of squares = 27.50
residual sum of squares of Y = 13.75
correlation coefficient = .82
$r^2 = .67$


I
10
5
10
20


II

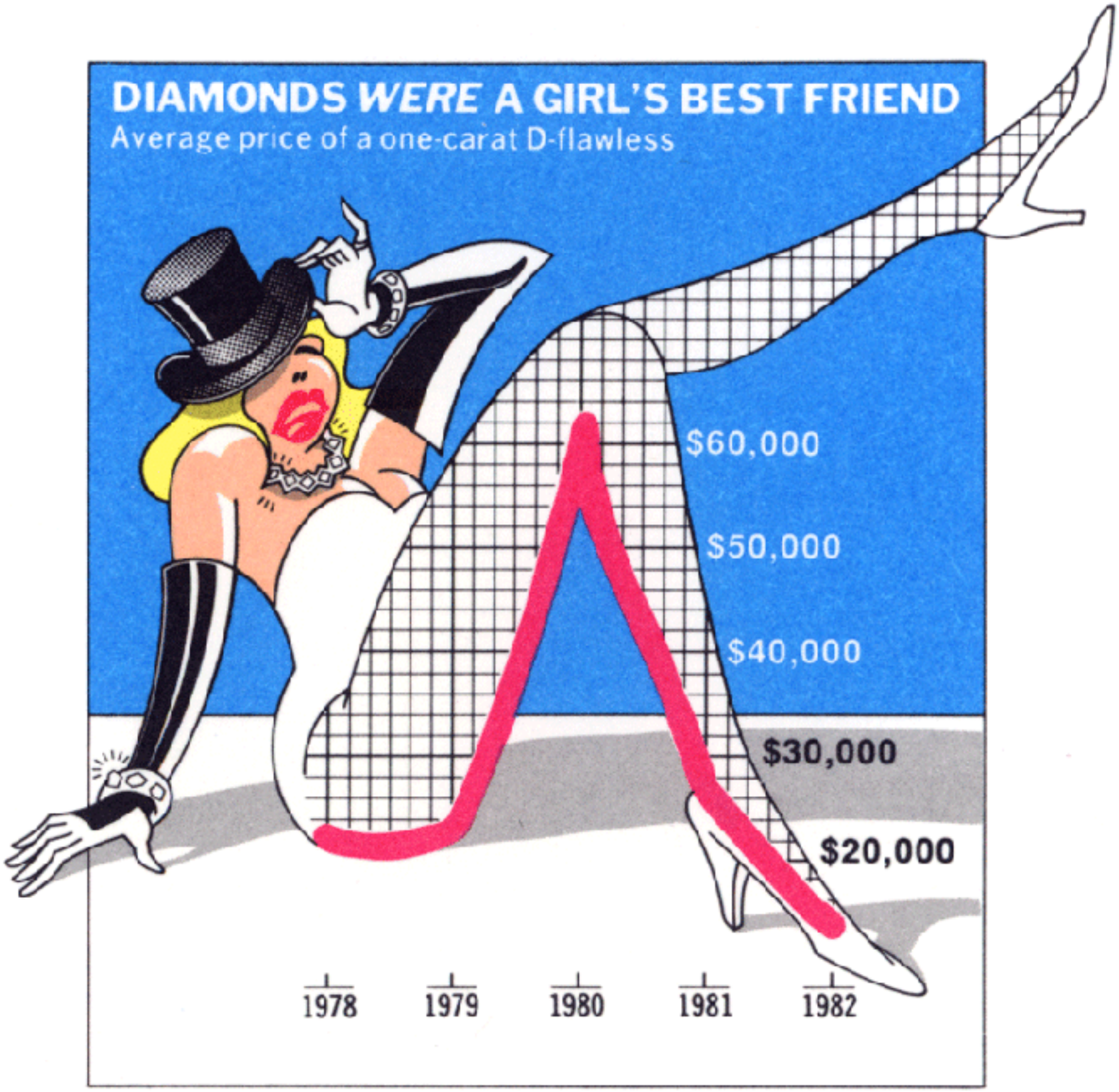

III


IV

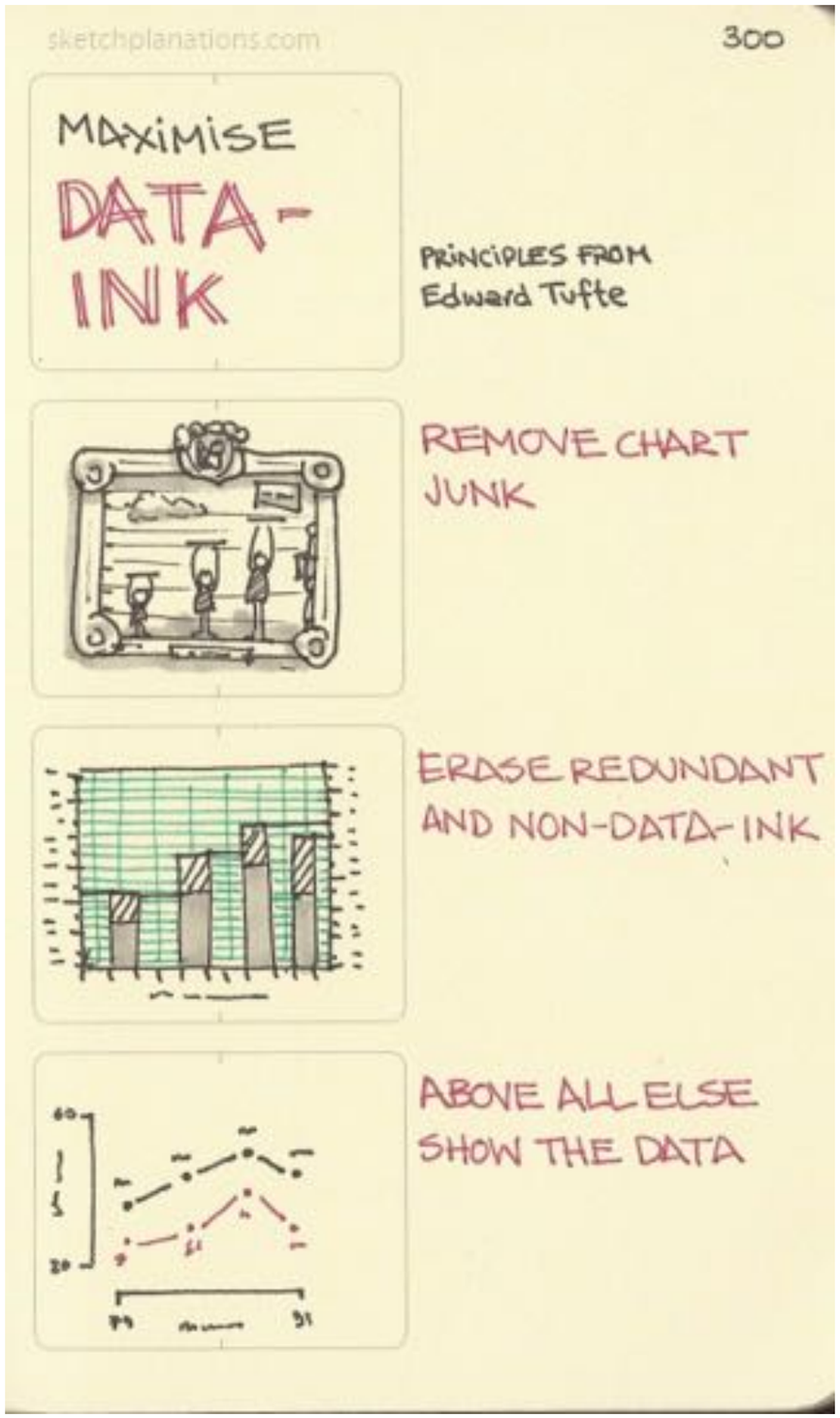
Graphics Reveal Data

- Graphical displays should…
  - Show the data
  - Avoid distortion
  - Induce viewer to think about substance
  - Make large datasets coherent
  - Encourage eye to compare different pieces of data

Chartjunk


DIAMONDS *WERE* A GIRL'S BEST FRIEND
Average price of a one-carat D-flawless
$60,000
$50,000
$40,000
$30,000
$20,000
1978
1979
1980
1981
1982

Maximize Data-ink Ratio



Grammar of Graphics

- A rational way of thinking about and making data graphics
- All plots are composed of:
    - Data
    - Layers / Geoms
    - Scales
    - Coord
    - Facet
    - Theme
- ggplot2 is an R-language implementation of Grammar of Graphics principles

Data

- The data you want to visualize, and *aesthetic mappings* describing how variables in data map to attributes you can see

- Much of the work with ggplot2 is creating the data frame - the rest is easy!

Layers and Geoms

- Layers are made of geometric elements (geoms)
    - geom_point
    - geom_bar
    - geom_violin
- Layers also have statistical transformations (e.g. binning or counting observations)

Scales

- Scales map values in data to values of aesthetics
    - Color
    - Size
    - Shape
- Define contents of legend

Coord

- Coordinate system describes how data coordinates are mapped to the plane of the graphic.

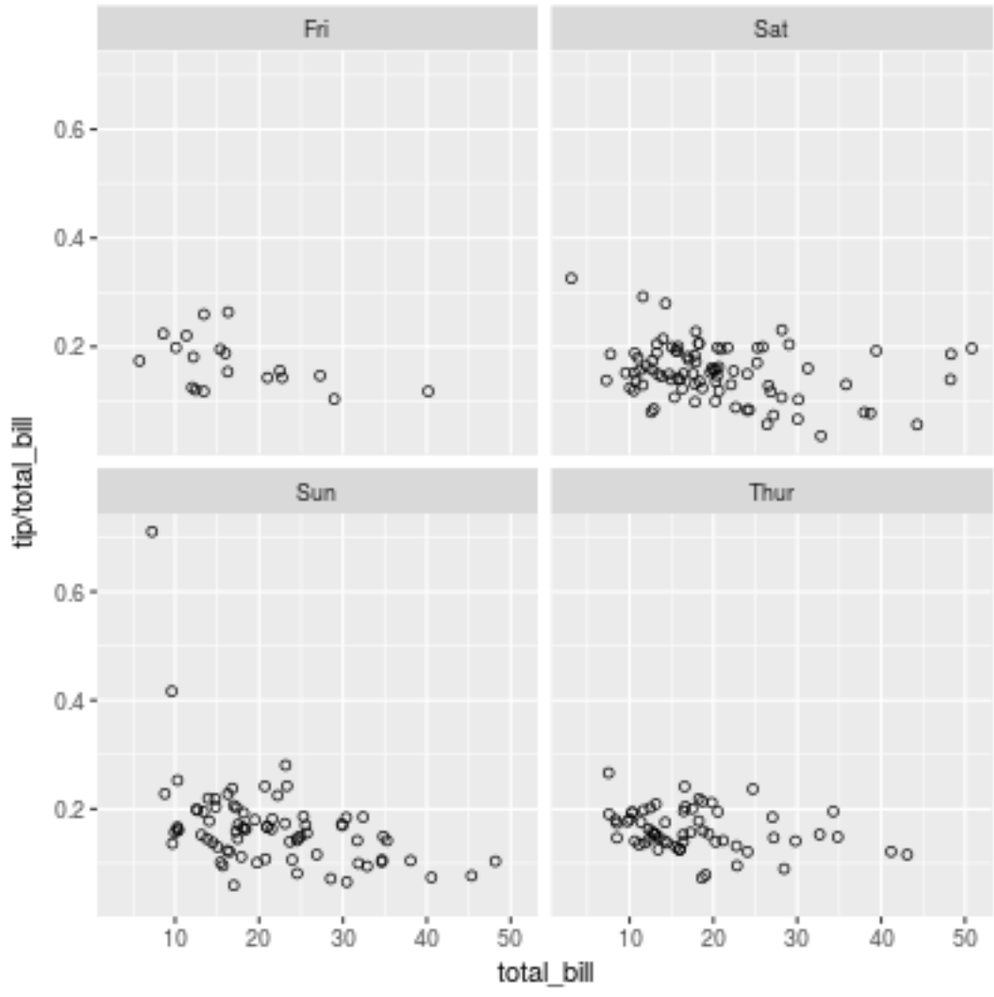
- Cartesian most common

- Make pie charts by changing to polar coordinates

Facets

- Breaks up data into subsets of small multiples

- Create an array of plots



Themes

- Controls details of display
- Size of fonts
- Grid lines
- Background color

Example

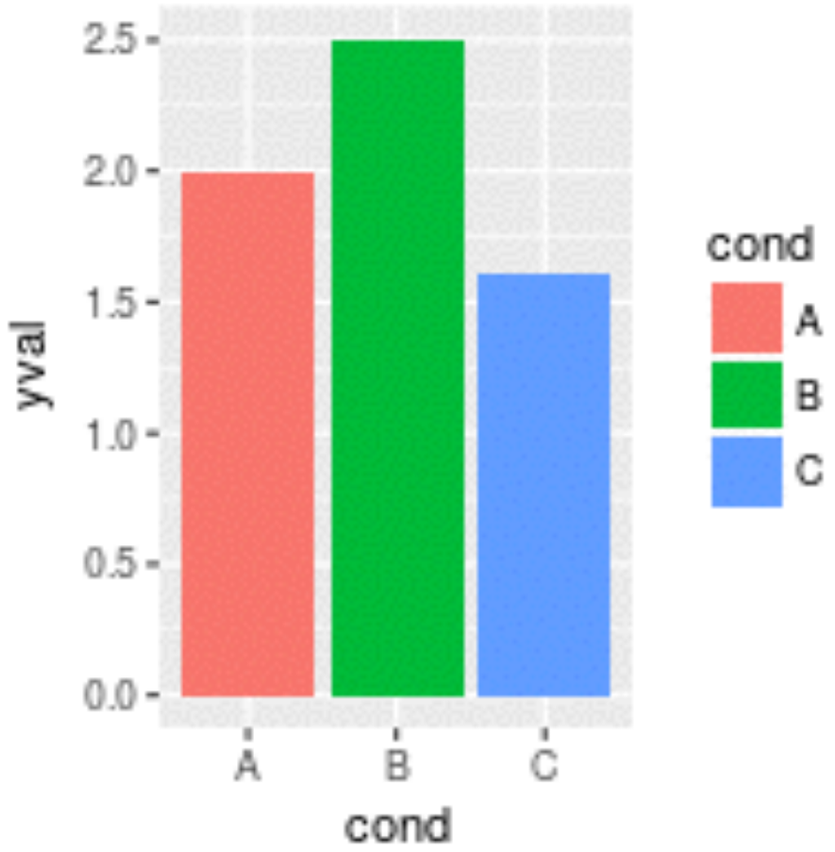
```
p <- ggplot(data=BRFSS)

p <- p + geom_point(aes(x=height, y=weight),
color="blue")

p
```

Data: Categorical vs. Quantitative

- Basic distinction in types of data
- Categorical
    - Take on discrete values, e.g., names
- Quantitative (or continuous)
    - Take on continuous values, e.g., height
- A figure will show relationship between two or more variables
    - Type of figure to use is influenced by type of data

Colors

- ggplot2 has default color scheme
  - Relatively nice, but has drawbacks
- Controlled by setting scales

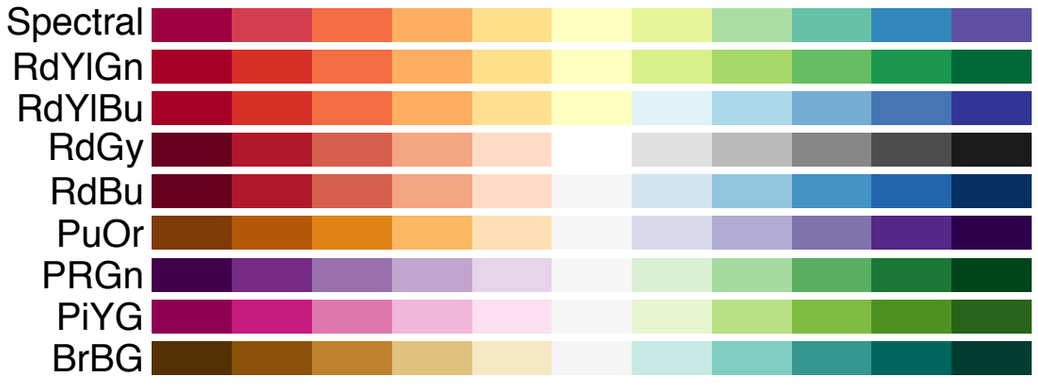


RColorBrewer

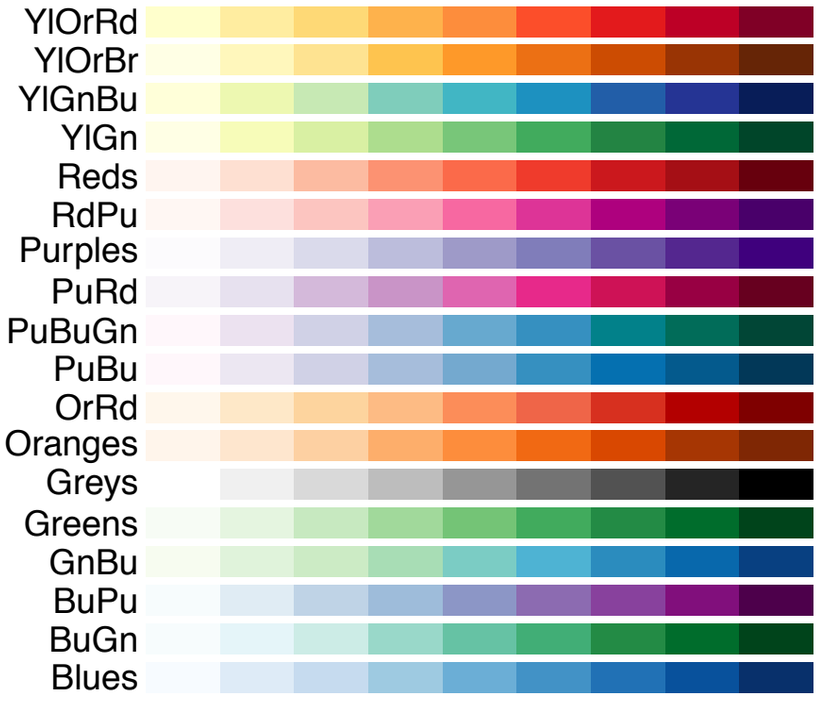
- Attractive set of alternative color palettes
- Color-blind friendly, looks distinguishable in monochrome
- My go-to colors for illustrations and figures!

Qualitative

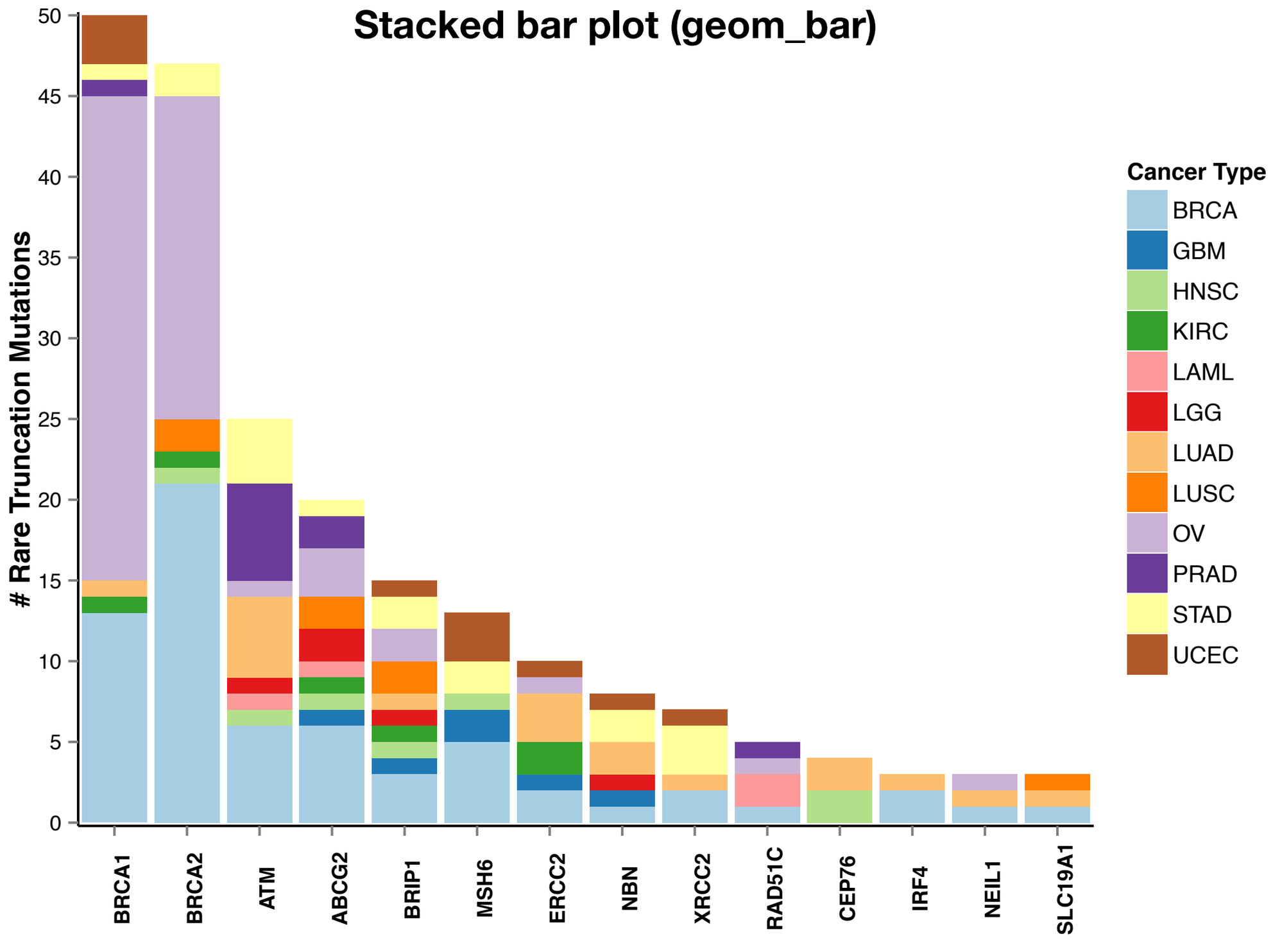
Diverging



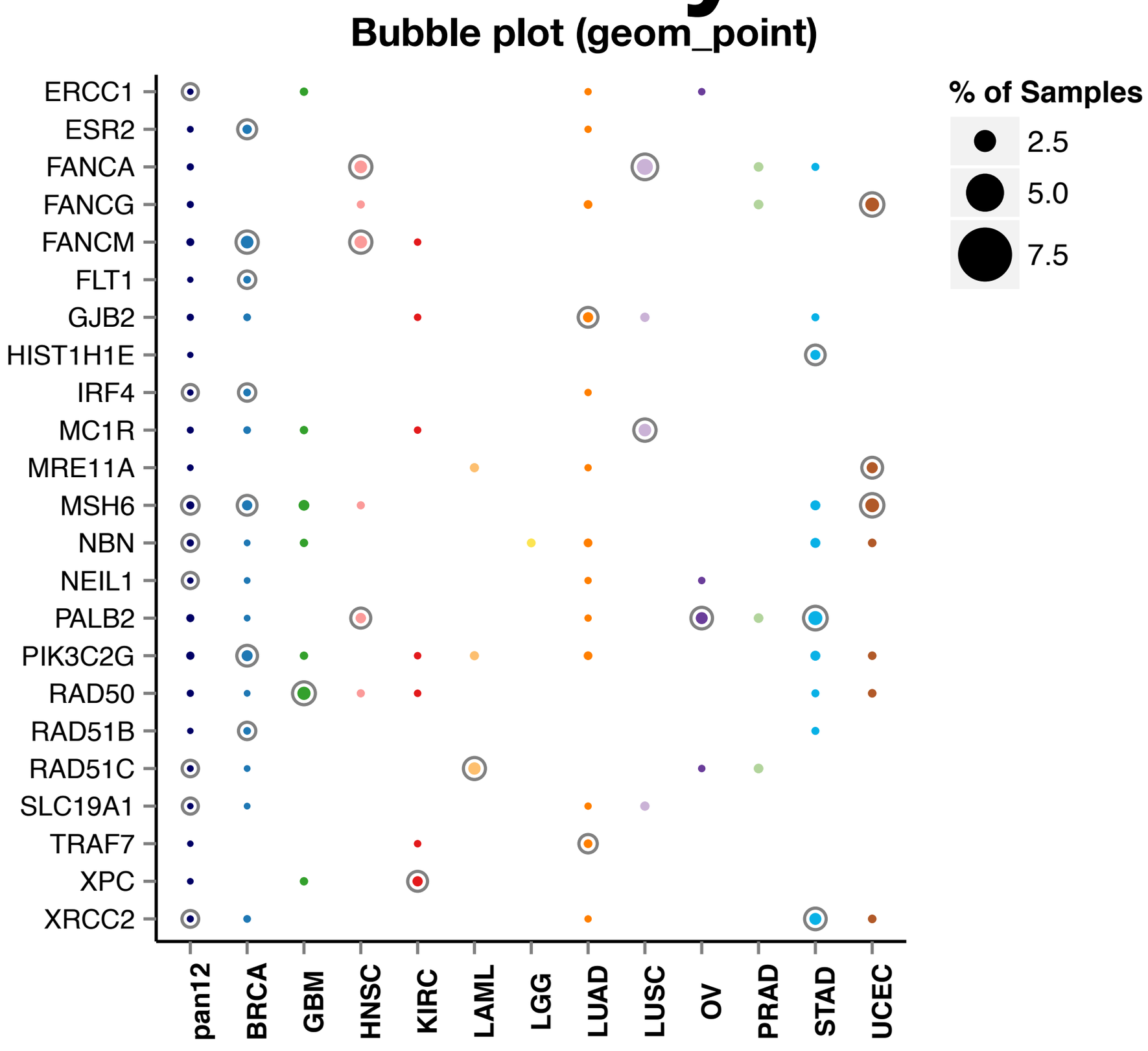
Sequential



Case Study: Before



Case Study: After



BRFSS Dataset

- 2016 BRFSS Survey Data and Documentation
    - Behavioral Risk Factor Surveillance System
    - https://www.cdc.gov/brfss/annual_data/annual_2016.html
- 486303 observations of 275 variables
    - 1.0 Gb in size
- Will subsample and select specific columns

Subsampled Dataset: BRFSS.48K.csv

- 48,630 observations across 10 variables
  - state
  - employed
  - income
  - seatbelt
  - diabetes
  - height
  - weight
  - bmi
  - age
  - sex

Getting more help

- ggplot2_book.pdf
  - For sale on amazon
  - Available for free https://github.com/hadley/ggplot2-book.git
  - Included in class notes
- R Cookbook: http://www.cookbook-r.com/Graphs/
- Google