36-315: Statistical Graphics and Visualization

Handout 8
Date: February 10, 2003

Uses of a scatterplot:

- Determining a relation between variables
- Making predictions
- Identifying outliers and subgroups

Relations between variables:

Predictability - y is predictable from x if the mean of y changes with x, when only x is known. Thus x reduces (but doesn't necessarily eliminate) the uncertainty in y. Note that x may or may not be predictable from y.

Linear correlation - y is predictable by a linear function of x, i.e. the mean of y changes linearly with x. In this case, the correlation coefficient measures predictive ability. This property is symmetric.

Statistical dependence - Some aspect of the distribution of y changes with x. This property is symmetric, and not the same as the everyday notion of dependence (causal dependence).

Permutation test - Rearrange the x values among the cases, also rearrange the y values, then plot. Looks the same → independent.

Checklist:

1. Transform to remove skew
2. Zoom in to show detail (or zoom out to show outliers)
3. Set symbol size (reduce overplotting)
4. Add trend line (with right smoothing factor)

Data-ink principle: The data should be prominent, the most prominent part of the plot

Common mistakes:

- Bad scale - Axes too big, lack of transformation
- Symbol too small or too large, not resistant to overplotting
- Obscuring boundary - no margin, ticks inward
- Making a lookup table - too many ticks/grid lines

Animals plot: data obscured by curves, big labels, no margin
Revision: rotated so labels don't clutter, shorter labels, bigger points, margin, fewer ticks

List of figures:

References

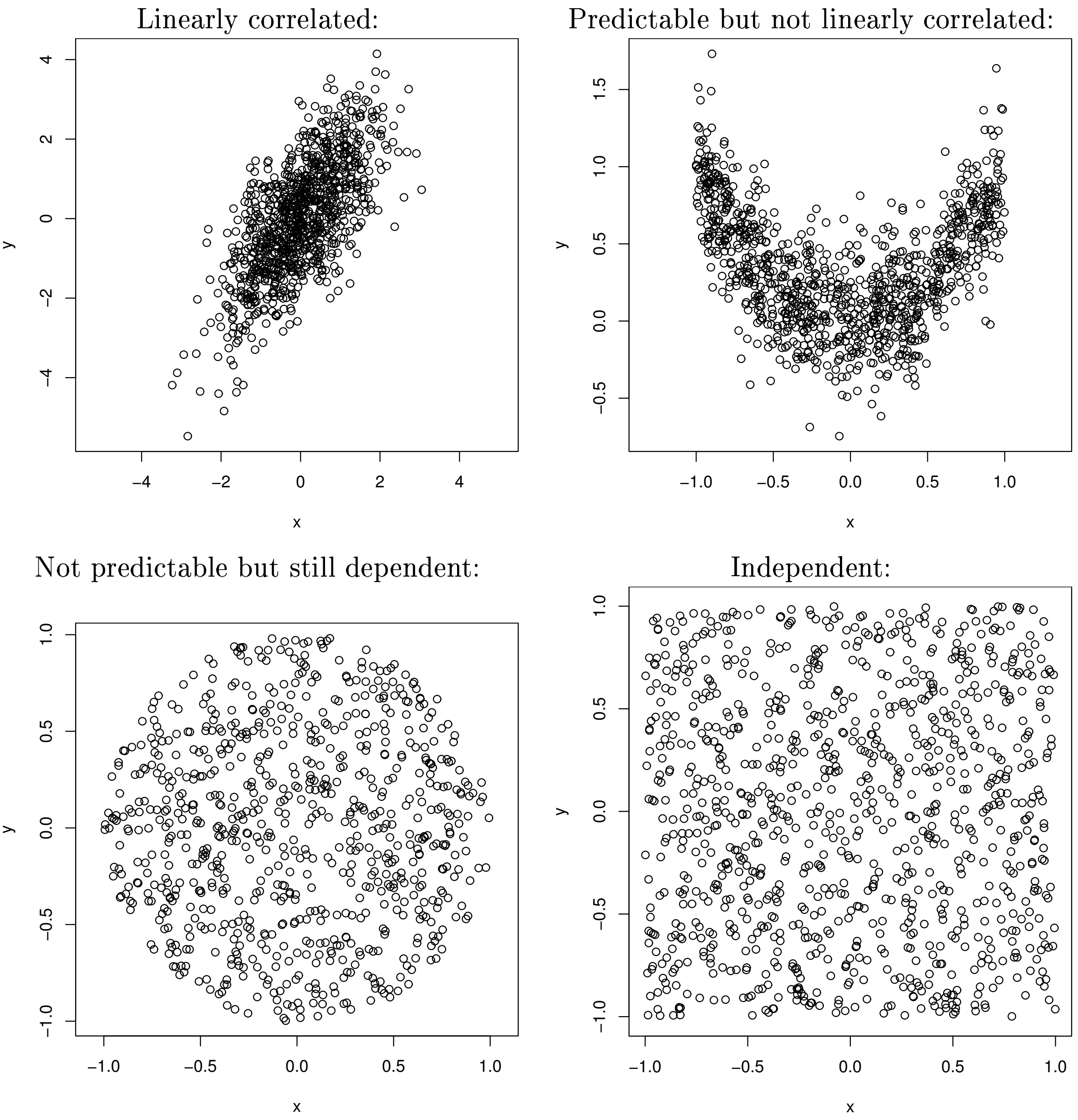
[1] J. M. Chambers, W. S. Cleveland, B. Kleiner, and P. A. Tukey. *Graphical methods for data analysis.* Wadsworth, 1983.

[2] William S. Cleveland. *The Elements of Graphing Data.* Hobart Press, NJ, 1994.

[3] CNN. "Butterfly ballot cost Gore White House"
`http://www.cnn.com/2001/ALLPOLITICS/03/11/palmbeach.recount/`

[4] Gary Klass. "Chart of the Week", Jan 1, 2002.
`http://lilt.ilstu.edu/gmklass/COW/archive/010102pbc.htm`

[5] Edward R. Tufte. *The Visual Display of Quantitative Information.* Graphics Press, Cheshire, CT 1983.



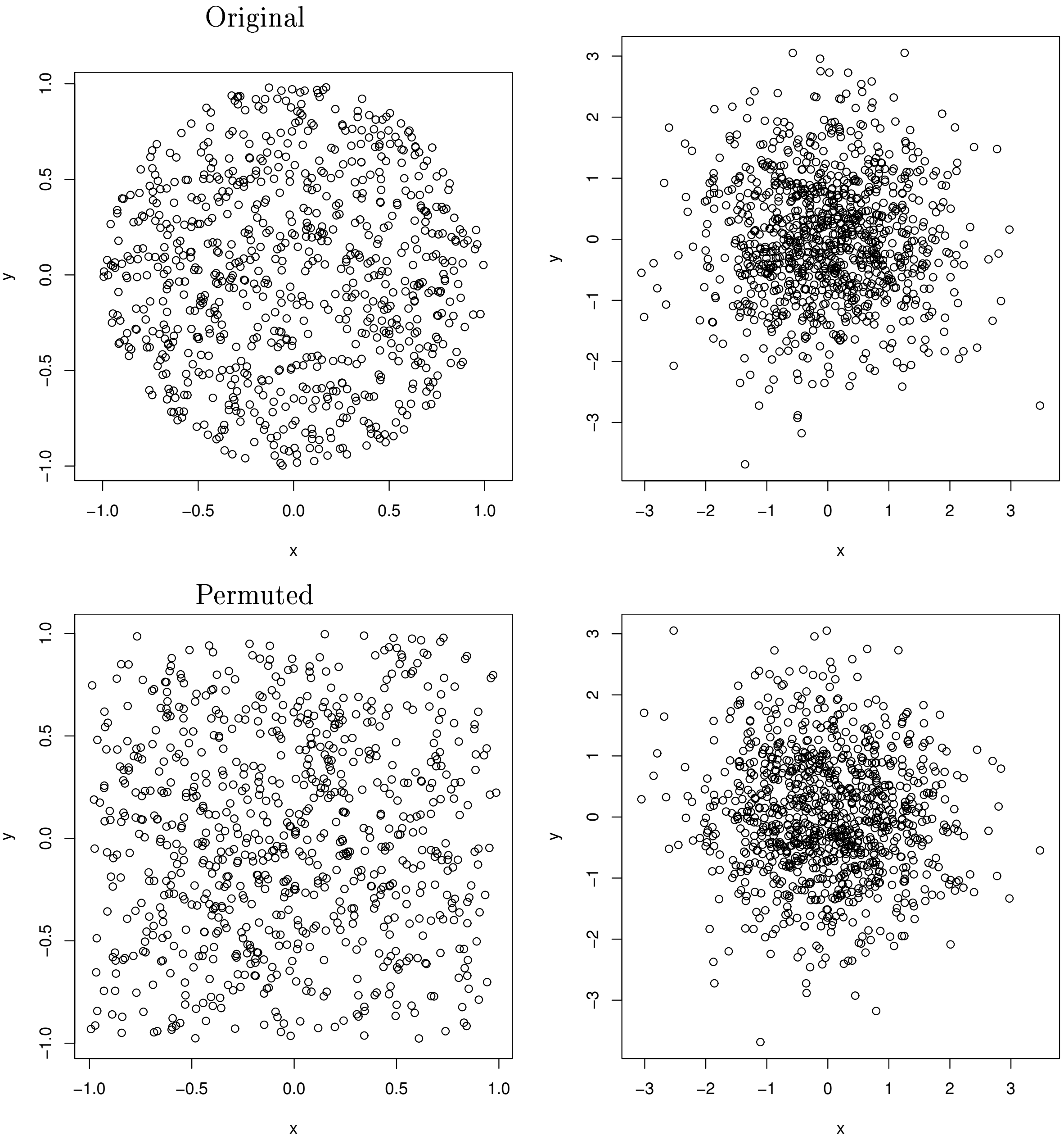
Linearly correlated:

Predictable but not linearly correlated:

Not predictable but still dependent:

Independent:

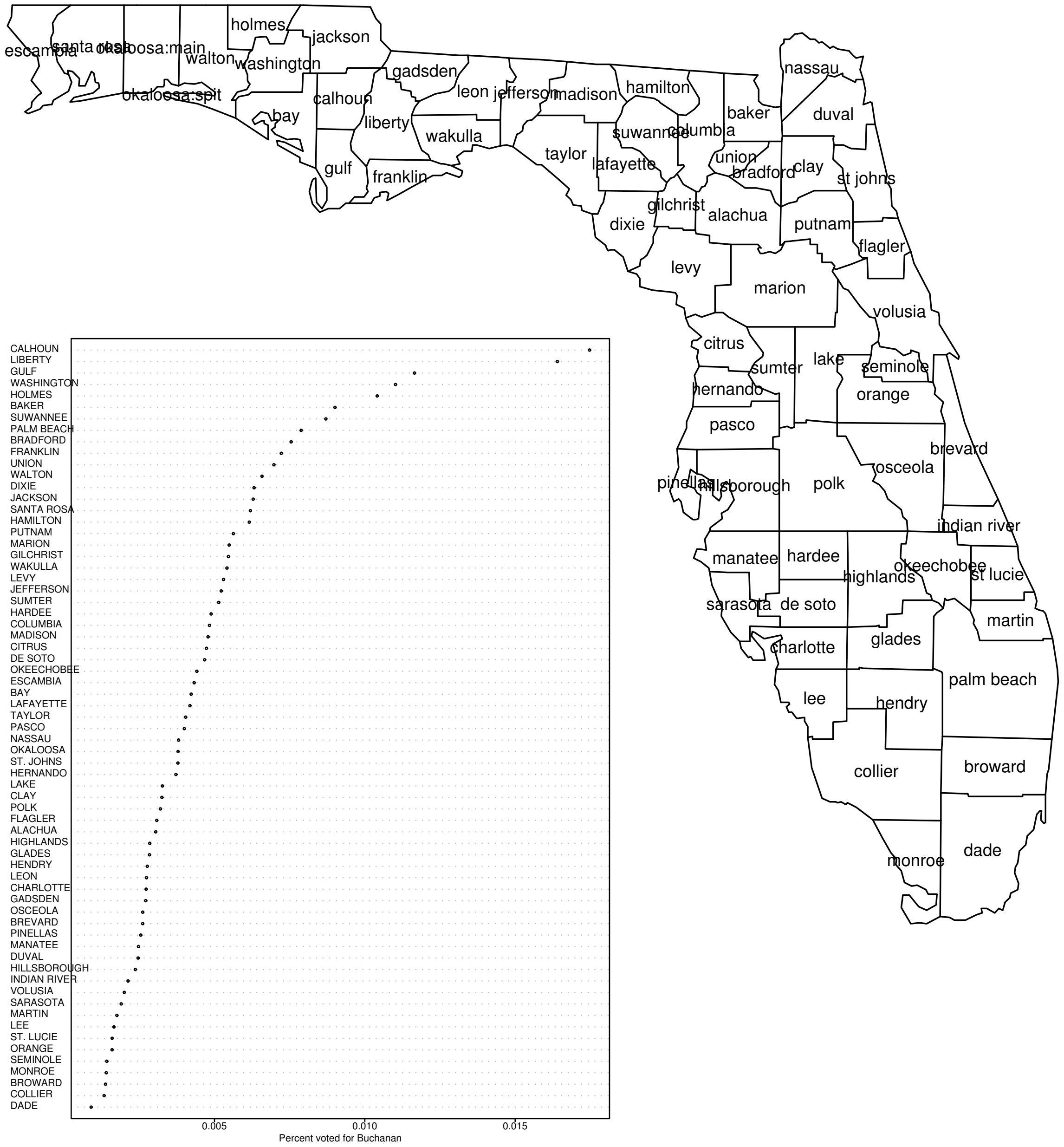
Permutation test for independence:



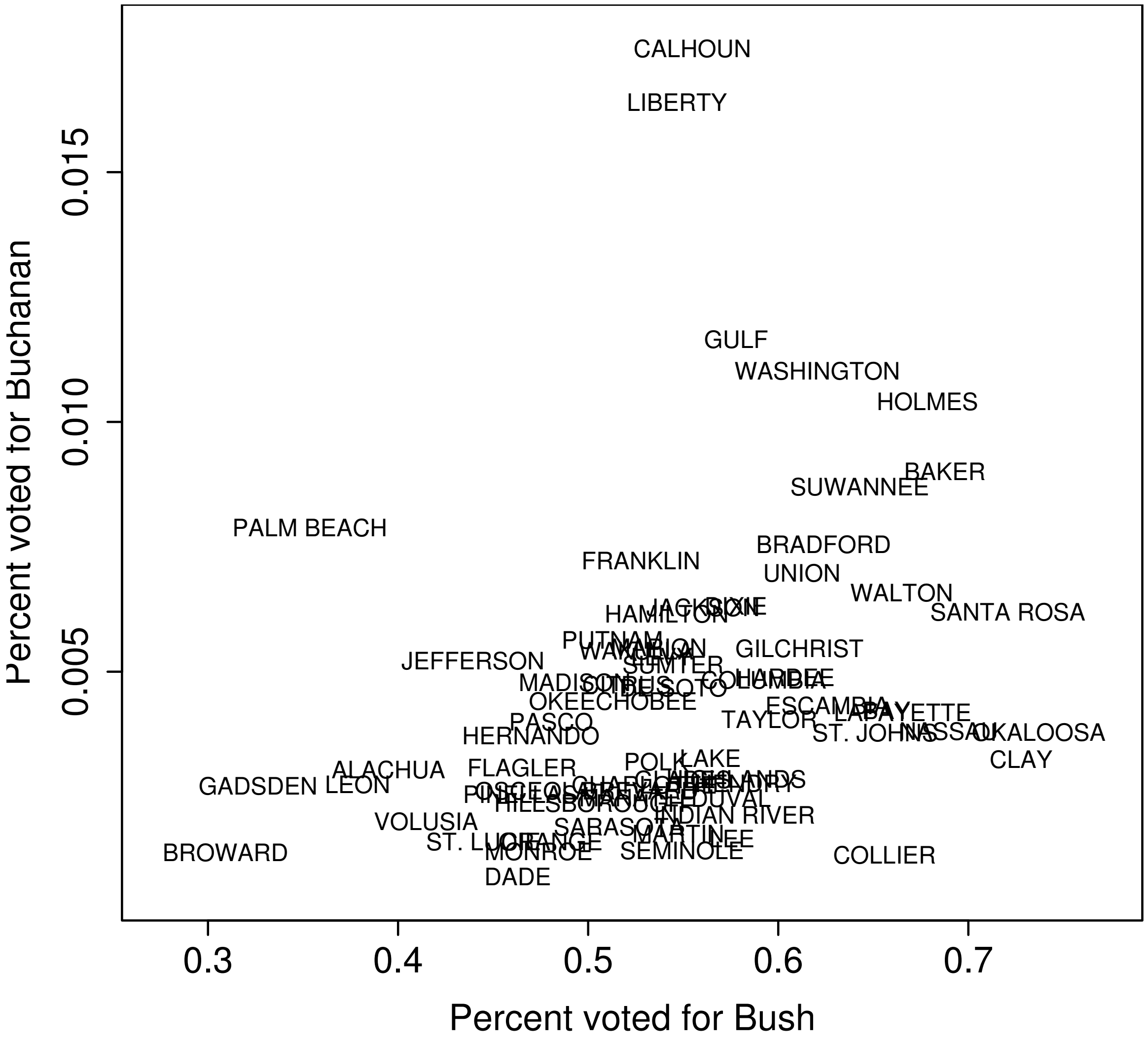
Original

Permuted

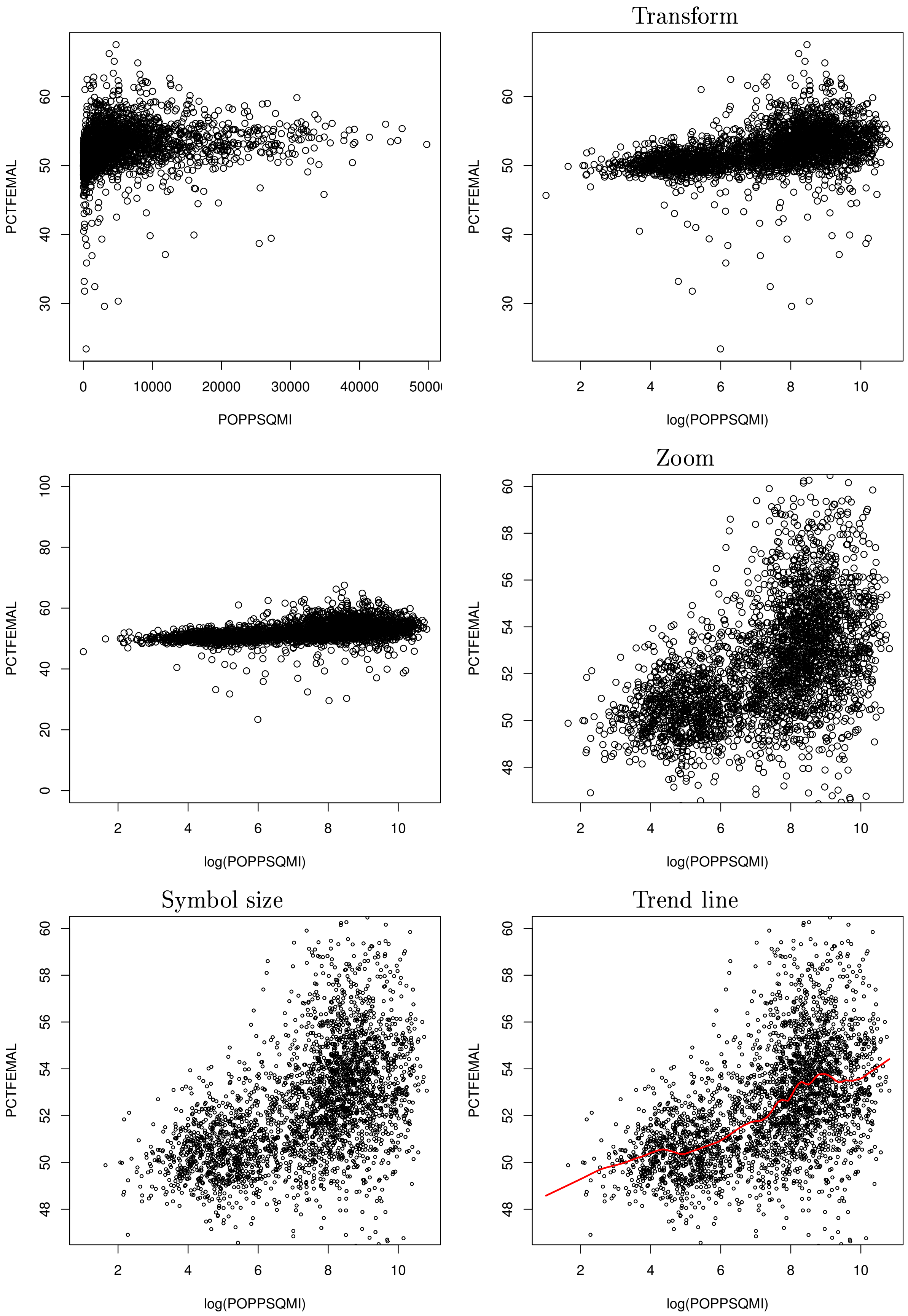
Voting in 2000 election:

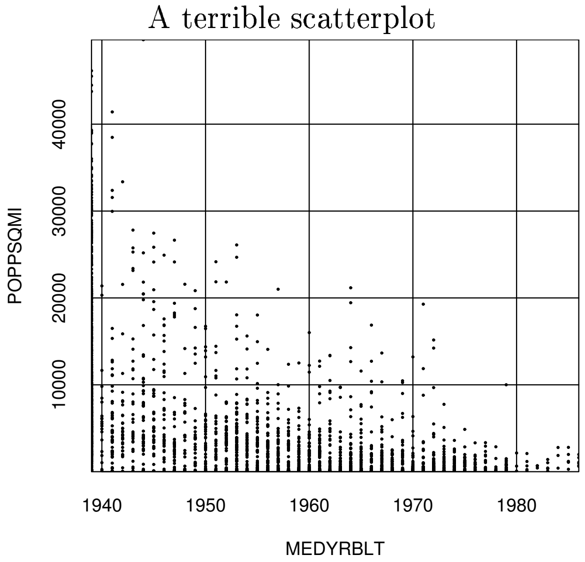


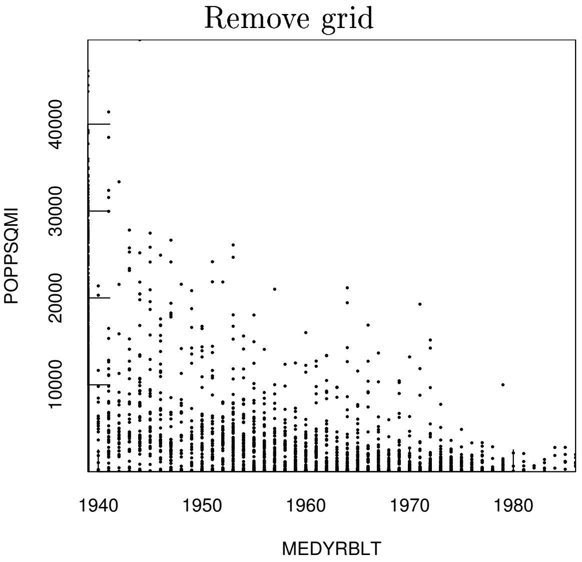
```
             GORE BUSH BUCHANAN
...
HILLSBOROUGH 0.47 0.50   0.0024
HOLMES       0.29 0.68   0.0104
INDIAN RIVER 0.40 0.58   0.0021
JACKSON      0.42 0.56   0.0063
JEFFERSON    0.54 0.44   0.0052
LAFAYETTE    0.31 0.67   0.0042
LAKE         0.41 0.56   0.0033
LEE          0.40 0.58   0.0017
LEON         0.60 0.38   0.0027
LEVY         0.42 0.54   0.0053
LIBERTY      0.42 0.55   0.0164
MADISON      0.49 0.49   0.0048
MANATEE      0.45 0.53   0.0025
MARION       0.44 0.54   0.0055
MARTIN       0.43 0.55   0.0018
MONROE       0.49 0.47   0.0014
NASSAU       0.29 0.69   0.0038
OKALOOSA     0.24 0.74   0.0038
OKEECHOBEE   0.47 0.51   0.0044
ORANGE       0.50 0.48   0.0016
OSCEOLA      0.51 0.47   0.0026
PALM BEACH   0.62 0.35   0.0079
PASCO        0.49 0.48   0.0040
PINELLAS     0.50 0.46   0.0025
POLK         0.45 0.54   0.0032
PUTNAM       0.46 0.51   0.0056
ST. JOHNS    0.32 0.65   0.0038
ST. LUCIE    0.53 0.44   0.0016
SANTA ROSA   0.25 0.72   0.0062
SARASOTA     0.45 0.52   0.0019
SEMINOLE     0.43 0.55   0.0014
SUMTER       0.43 0.54   0.0051
SUWANNEE     0.33 0.64   0.0087
TAYLOR       0.39 0.60   0.0040
UNION        0.37 0.61   0.0070
VOLUSIA      0.49 0.41   0.0020
WAKULLA      0.45 0.53   0.0054
WALTON       0.31 0.66   0.0066
WASHINGTON   0.35 0.62   0.0110
```

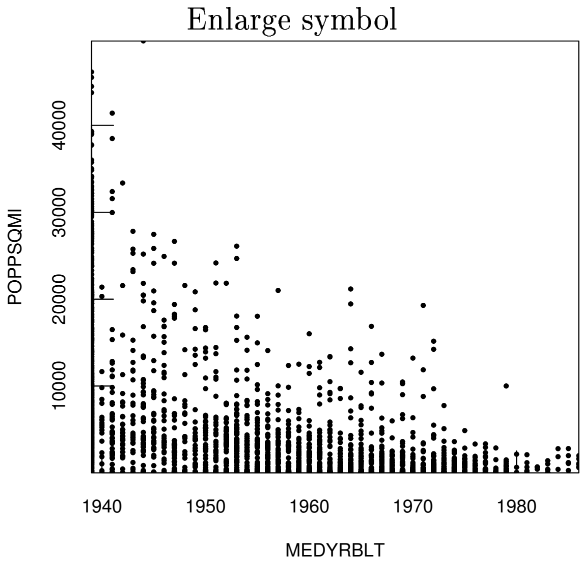


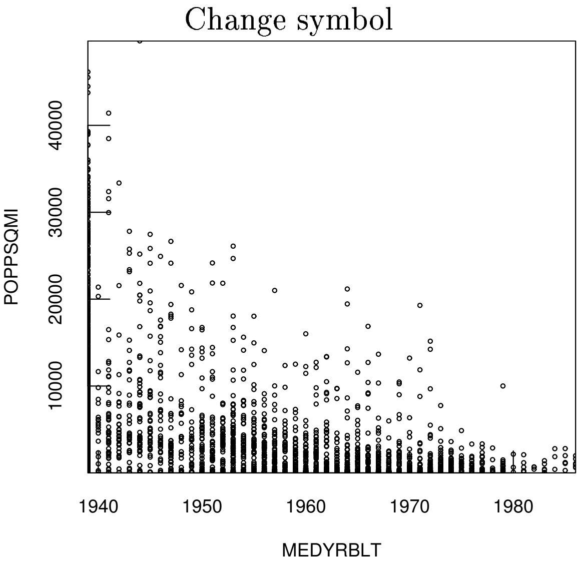
CALHOUN
LIBERTY
GULF
WASHINGTON
HOLMES
BAKER
SUWANNEE
PALM BEACH
BRADFORD
FRANKLIN
UNION
WALTON
SANTA ROSA
JACKSON
DIXIE
HAMILTON
PUTNAM
WAKULLA
LEVY
MARION
GILCHRIST
JEFFERSON
SUMTER
HARDEE
COLUMBIA
MADISON
CITRUS
DESOTO
OKEECHOBEE
ESCAMBIA
LAFAYETTE
BAY
TAYLOR
PASCO
ST. JOHNS
NASSAU
OKALOOSA
HERNANDO
CLAY
LAKE
POLK
ALACHUA
FLAGLER
GLADES
HIGHLANDS
HENDRY
GADSDEN
LEON
OSCEOLA
CHARLOTTE
BREVARD
PINELLAS
MANATEE
HILLSBOROUGH
DUVAL
INDIAN RIVER
VOLUSIA
SARASOTA
MARTIN
LEE
ST. LUCIE
ORANGE
BROWARD
MONROE
SEMINOLE
COLLIER
DADE
0.015
0.010
0.005
Percent voted for Buchanan
0.3
0.4
0.5
0.6
0.7
Percent voted for Bush

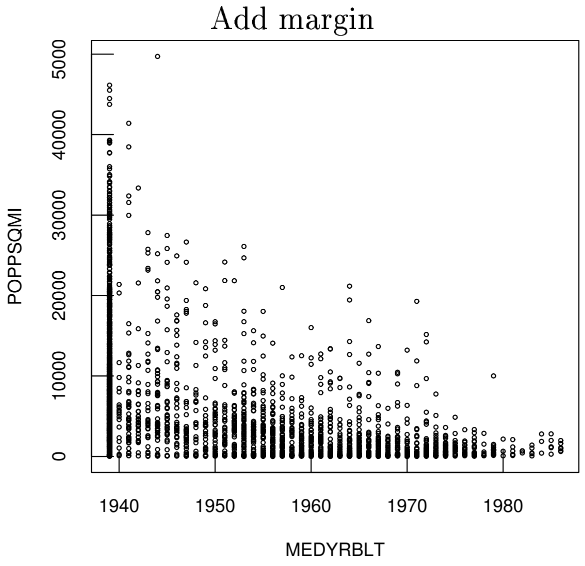

Transform
Zoom
Symbol size
Trend line
PCTFEMAL
POPPSQMI
log(POPPSQMI)
0
10000
20000
30000
40000
50000
2
4
6
8
10
30
40
50
60
0
20
40
60
80
100
48
50
52
54
56
58
60

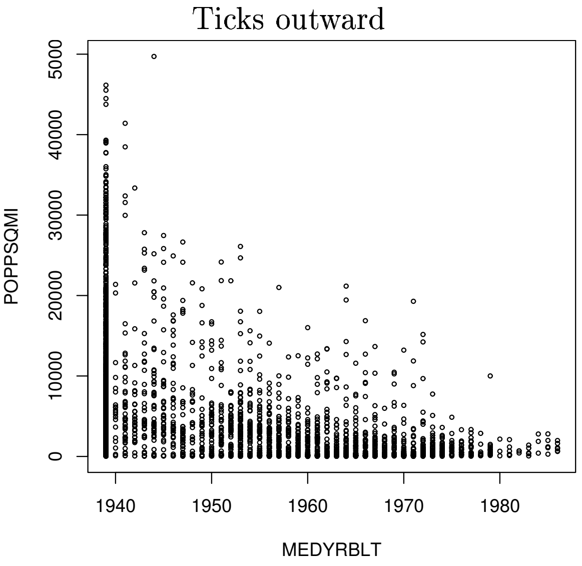

A terrible scatterplot


Remove grid


Enlarge symbol


Change symbol


Add margin


Ticks outward