4) Introduction to Data

Vitor Kamada

January 2018

Reference

Tables, Graphics, and Figures from

Introductory Statistics with Randomization and Simulation

Diez et al. (2014): Ch 1 - Introduction to Data

E-mail Dataset

rm(list = ls()); library(openintro)

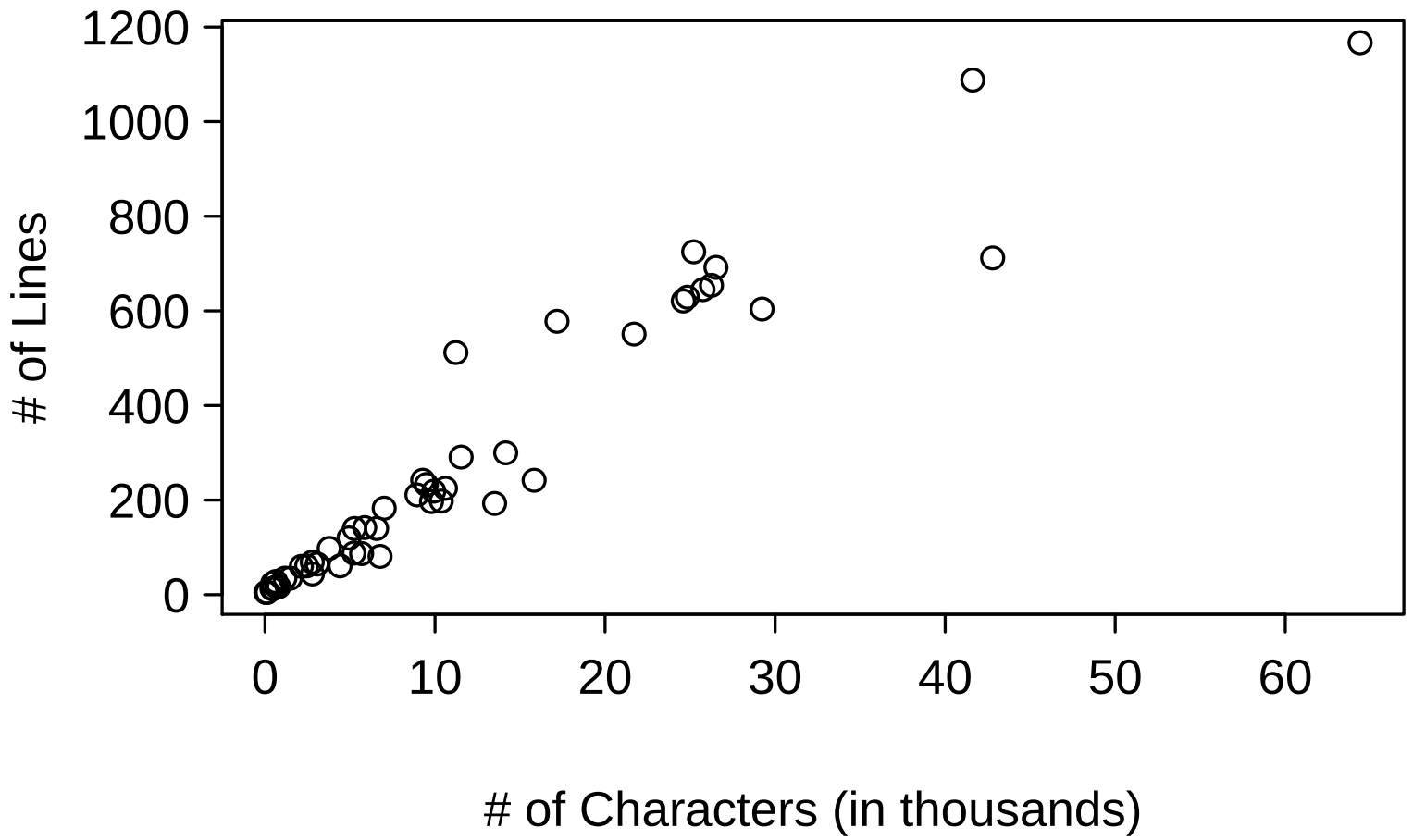
data(email50); attach(email50)

	spam	num_char	line_breaks	format	number
1	no	21,705	551	html	small
2	no	7,011	183	html	big
3	yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮	⋮
50	no	15,829	242	html	small

Scatterplot

plot(num_char, line_breaks,

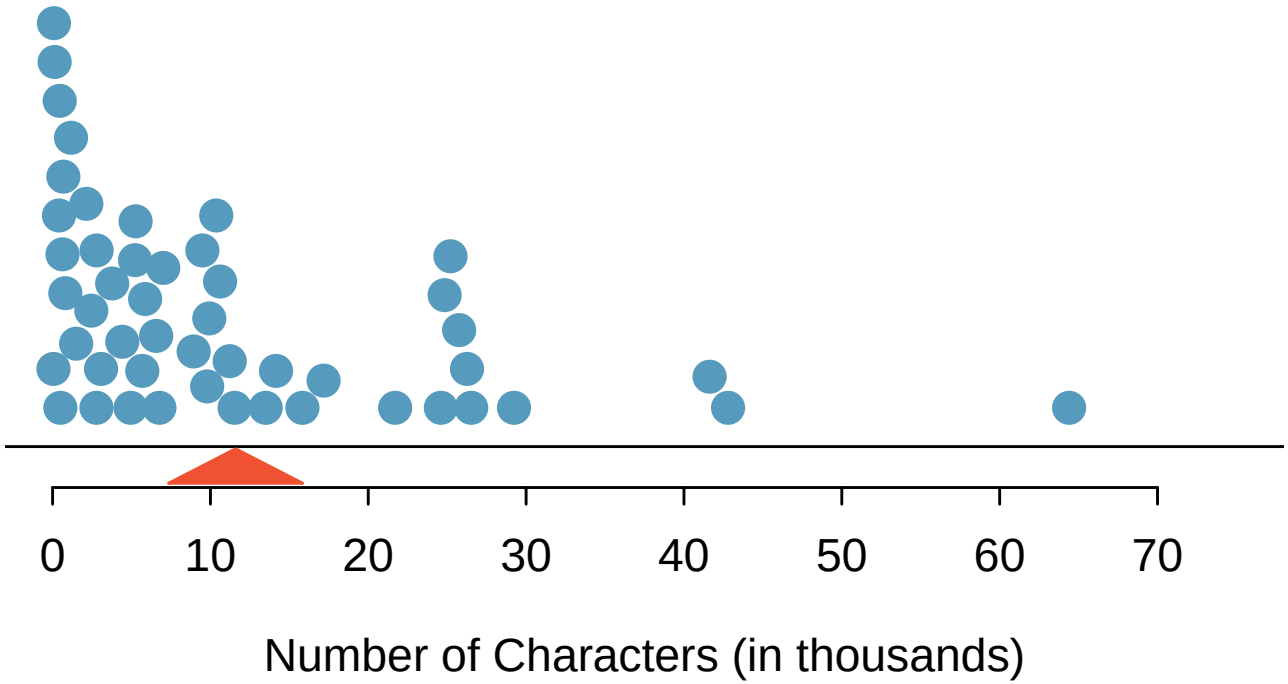
xlab="# of Characters (in thousands)", ylab="# of Lines")



Sample Mean ($\bar{x}$)

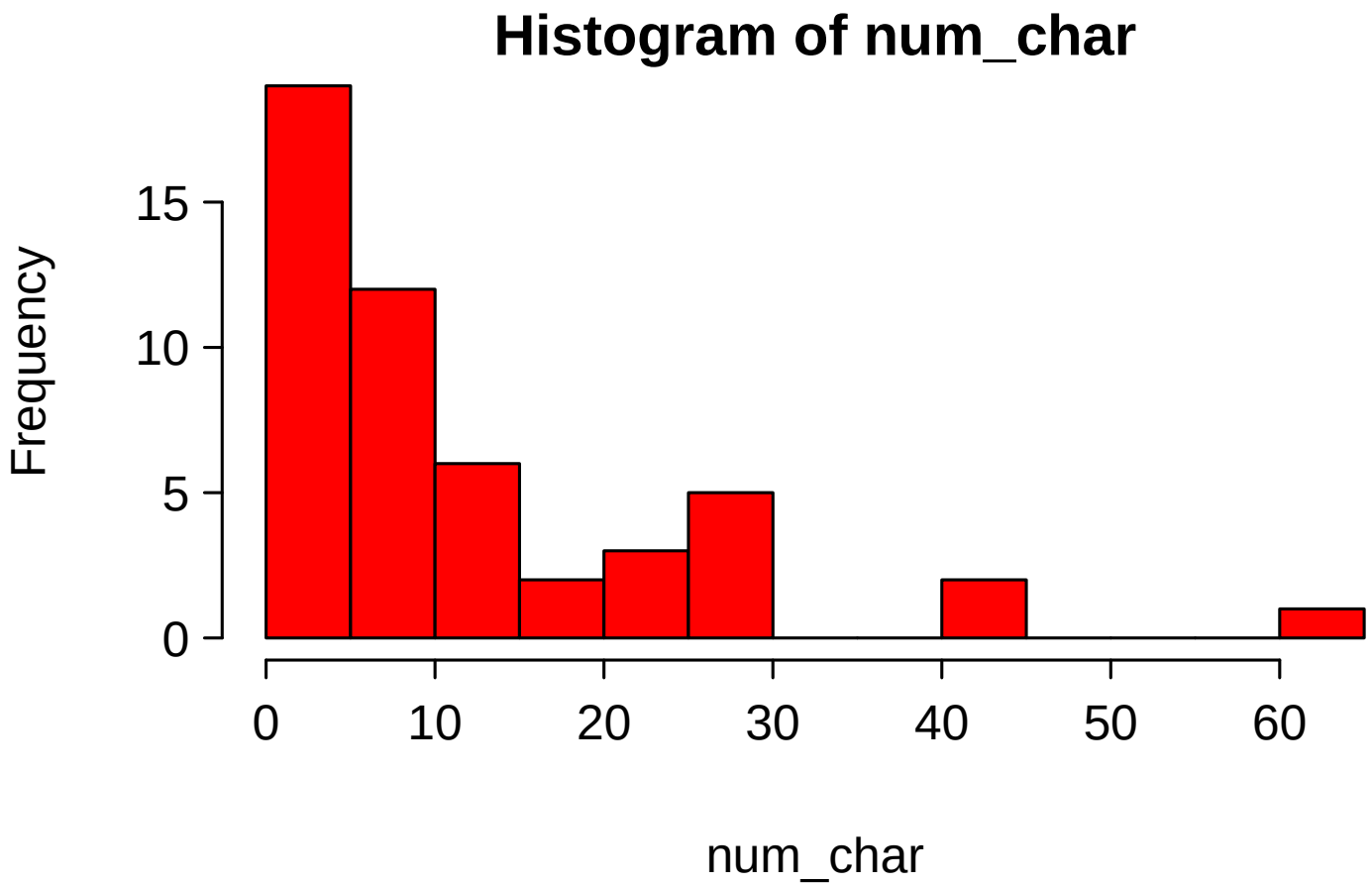
mean(num_char)

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \sum_{i=1}^{n} x_i = 11.6$$



Histogram

```
myPDF("Histogram.pdf")
hist(num_char, breaks=12, col="red")
dev.off()
```



Sample Variance (s^2) and Standard Deviation (s)

var(num_char)

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = 172.27$$

sd(num_char)

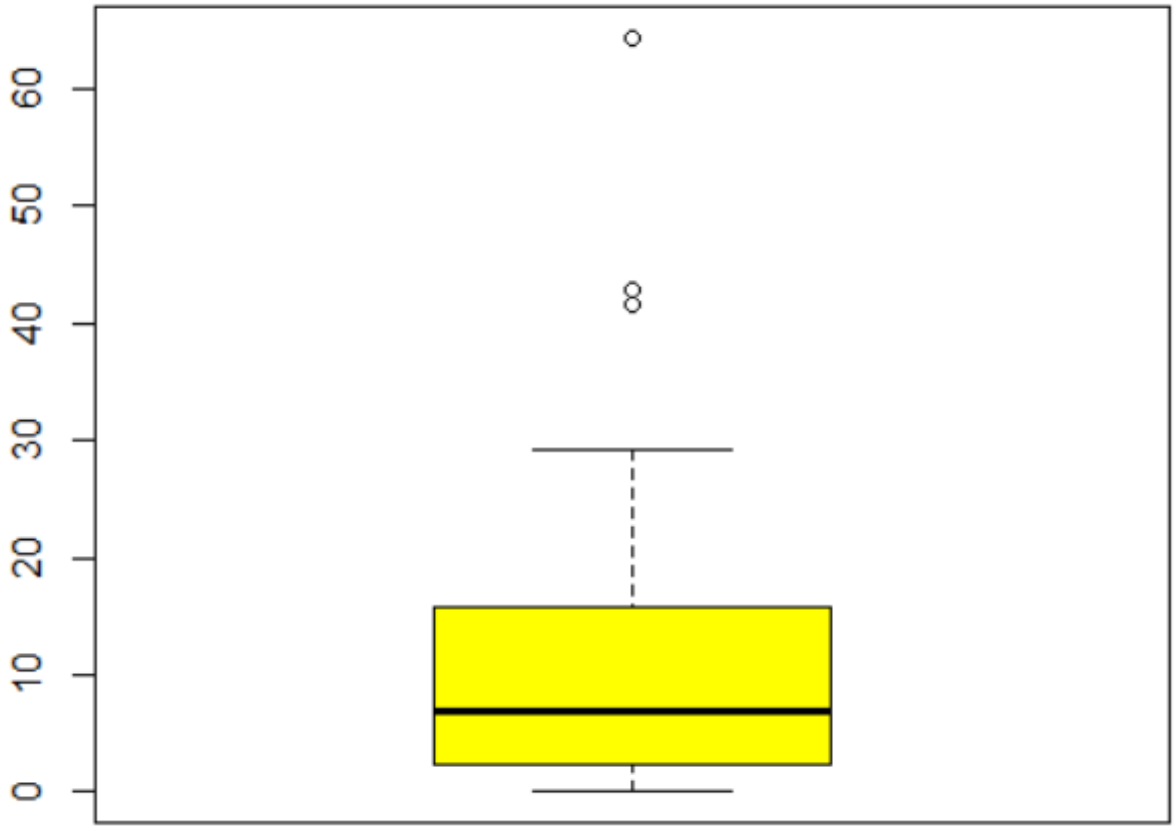
$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = 13.12$$

Boxplot

median(num_char) 6.9

IQR(num_char) 12.87

boxplot(num_char, col="yellow")

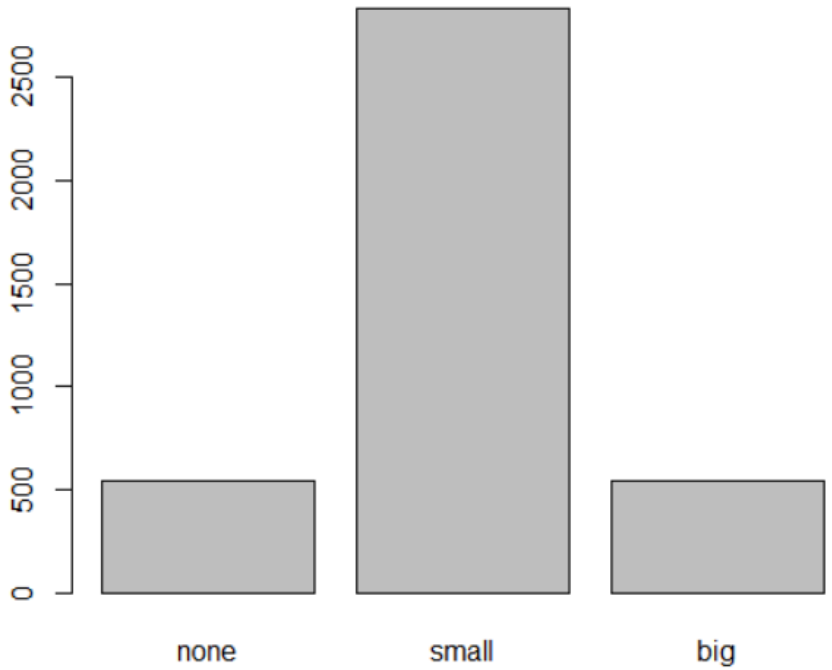


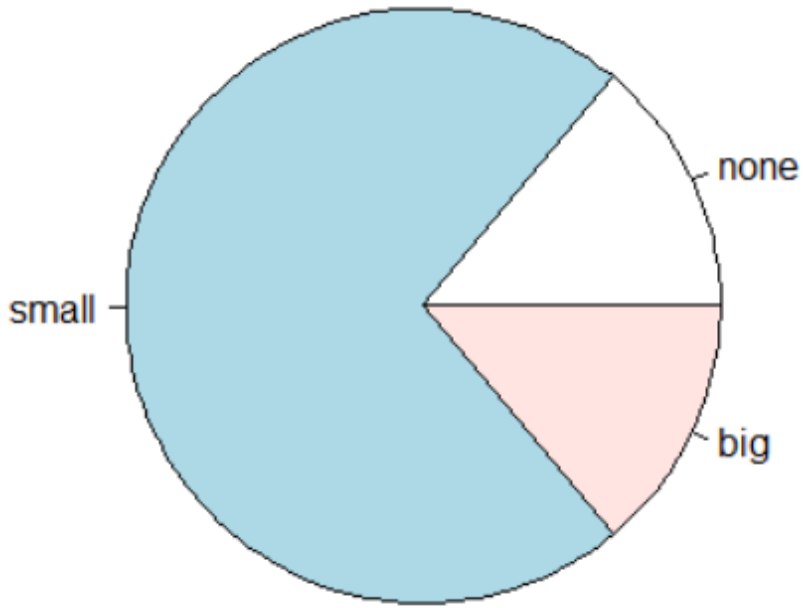
Robust Statistics

	robust		not robust	
scenario	median	IQR	$\bar{x}$	s
original `num_char` data	6,890	12,875	11,600	13,130
drop 66,924 observation	6,768	11,702	10,521	10,798
move 66,924 to 150,000	6,890	12,875	13,310	22,434

Barplot

```
t <- table(email$number)
barplot(t); pie(t)
```





Contingency Table

data(email); attach(email)

tb<-table(spam, number); addmargins(tb)

spam	none	small	big	Sum
0	400	2659	495	3554
1	149	168	50	367
Sum	549	2827	545	3921

round(prop.table(tb,2),2) # column %

	number		
spam	none	small	big
0	0.73	0.94	0.91
1	0.27	0.06	0.09

US Census Website

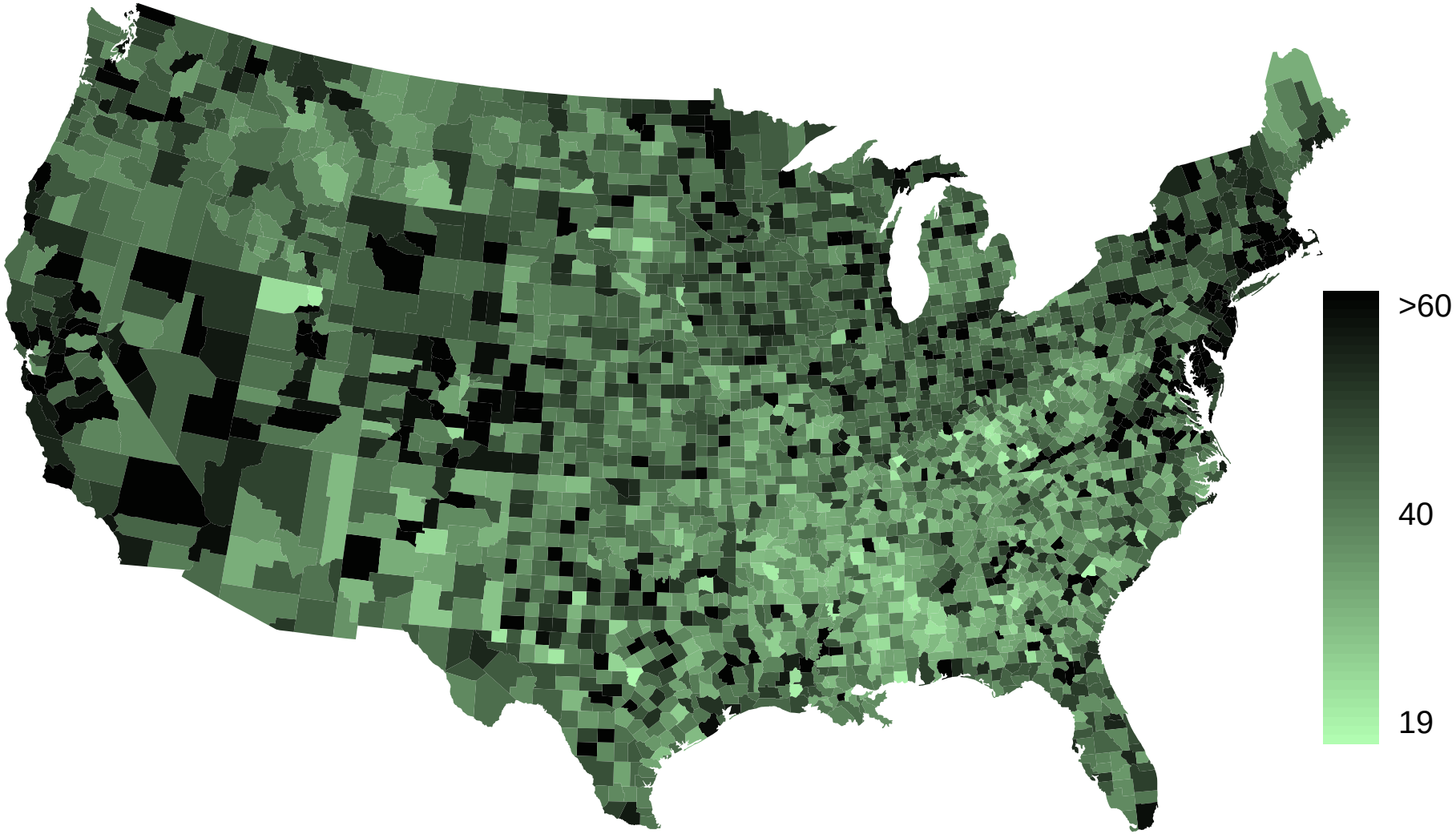
http://quickfacts.census.gov/qfd/index.html

	name	state	pop2000	pop2010	fed_spend	poverty
1	Autauga County	Alabama	43671	54571	6.068095	10.6
2	Baldwin County	Alabama	140415	182265	6.139862	12.2
3	Barbour County	Alabama	29038	27457	8.752158	25.0
4	Bibb County	Alabama	20826	22915	7.122016	12.6
5	Blount County	Alabama	51024	57322	5.130910	13.4
6	Bullock County	Alabama	11714	10914	9.973062	25.3
7	Butler County	Alabama	21399	20947	9.311835	25.0
8	Calhoun County	Alabama	112249	118572	15.439218	19.5
9	Chambers County	Alabama	36583	34215	8.613707	20.3
10	Cherokee County	Alabama	23988	25989	7.104621	17.6

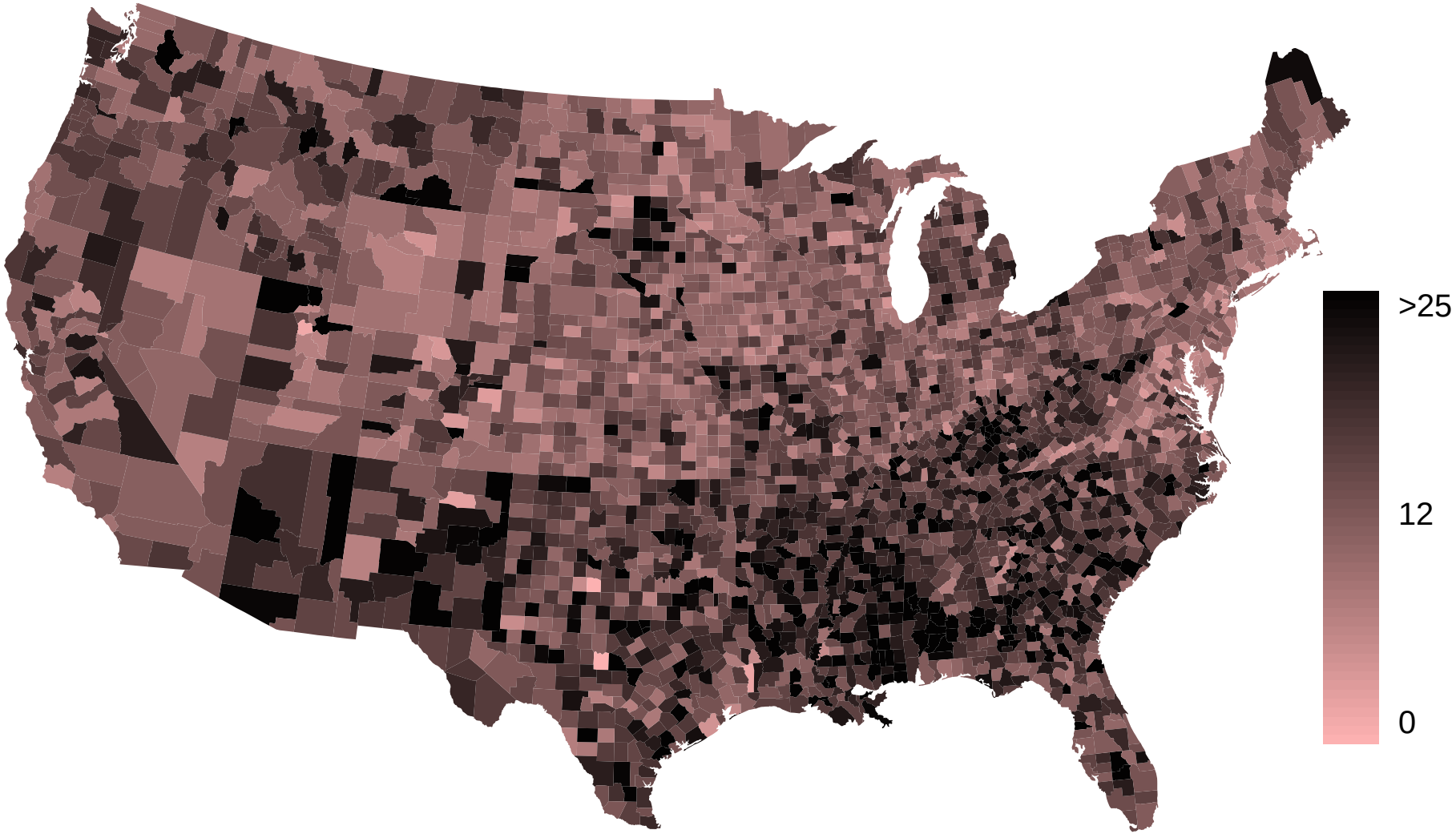
Intensity Map

```
library(openintro)
data(countyComplete); data(county)
source("countyMap.R")
myPDF("MedIncomeMap.pdf", 7.8, 4.5)
val <- county$med_income/1000
val[val > 60] <- 60
countyMap(val, countyComplete$FIPS, "green",
gtlt=">")
dev.off()
```

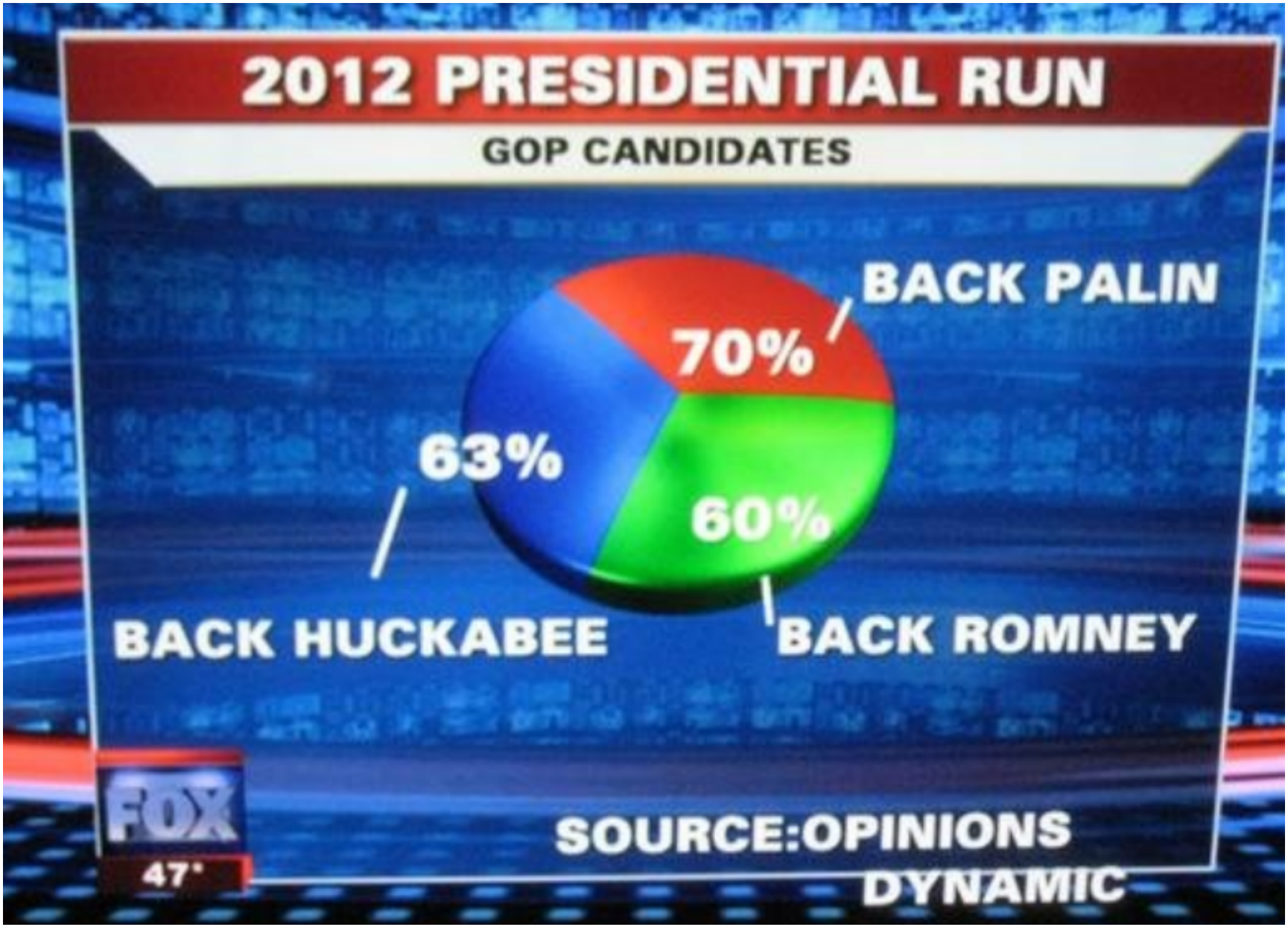
Median Household Income ($1000s).



Poverty Rate (%)

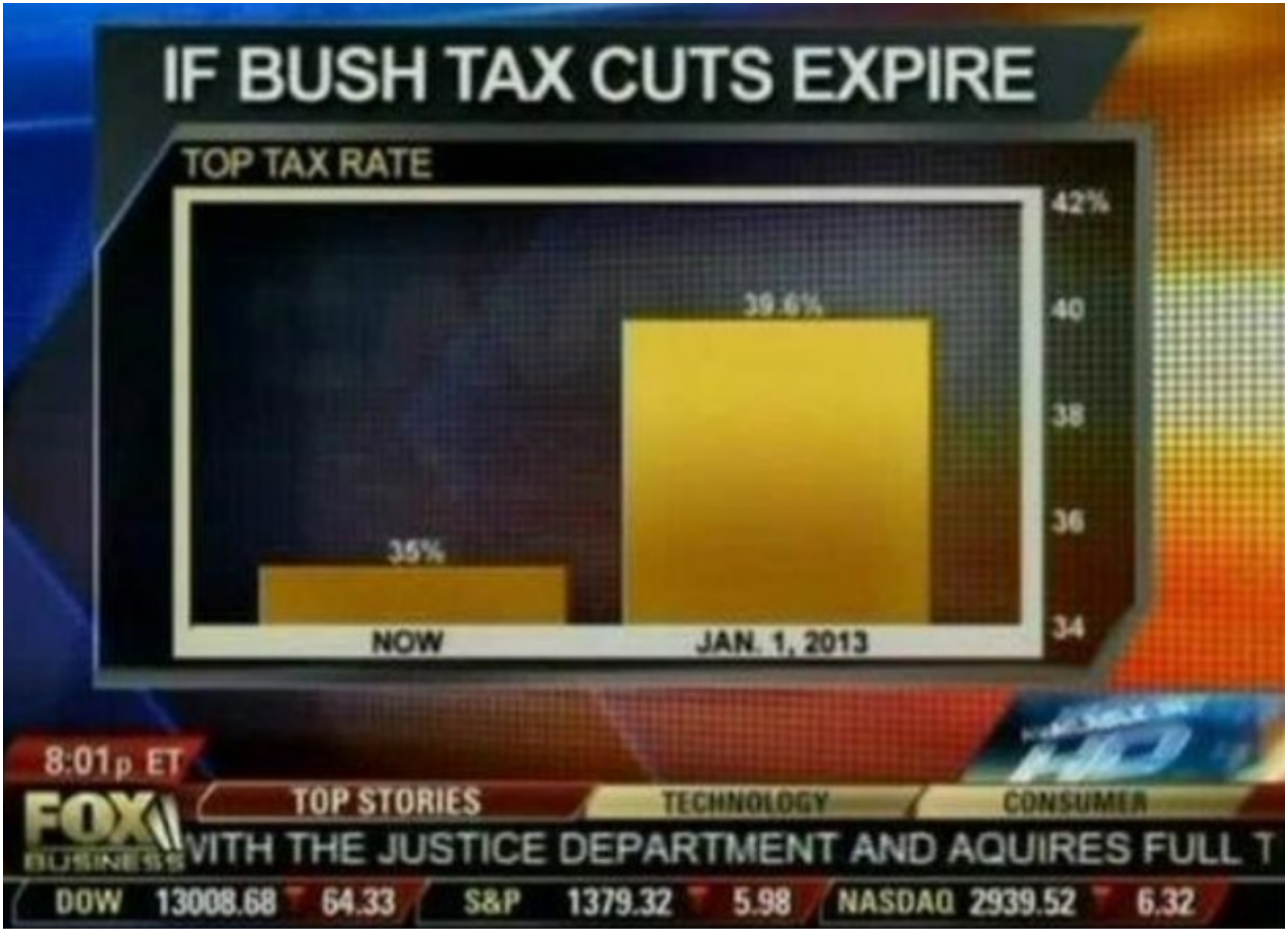


Presidential Run



Global Warming

Tax Cuts


IF BUSH TAX CUTS EXPIRE
TOP TAX RATE
42%
40
38
36
34
39.6%
35%
NOW
JAN. 1, 2013
8:01p ET
FOX BUSINESS
TOP STORIES
TECHNOLOGY
CONSUMER
VITH THE JUSTICE DEPARTMENT AND AQUIRES FULL T
DOW 13008.68 64.33
S&P 1379.32 5.98
NASDAQ 2939.52 6.32

Unemployment Rate