4) Introduction to Data

Vitor Kamada

December 2018

Reference

Tables, Graphics, and Figures from

Introductory Statistics with Randomization and Simulation

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

50 E-mail Dataset

```
import pandas as pd
file = "https://github.com/VitorKamada/ECO5100/raw/master/Data/email50.csv"
email50 = pd.read_csv(file)
email50.head()
```

	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

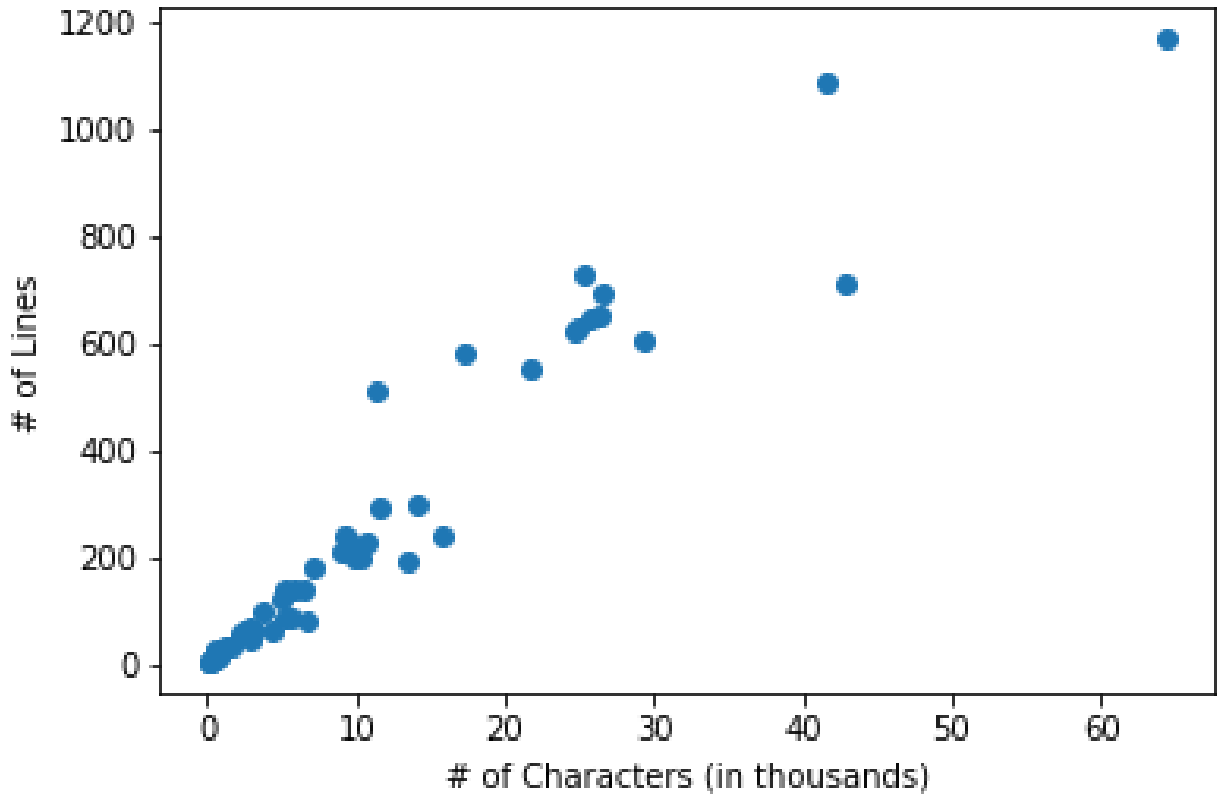
email50.describe()

	dollar	inherit	viagra	password	num_char	line_breaks
count	50.000000	50.0	50.0	50.000000	50.000000	50.00000
mean	0.900000	0.0	0.0	0.460000	11.598220	267.30000
std	3.518174	0.0	0.0	1.631451	13.125261	290.81983
min	0.000000	0.0	0.0	0.000000	0.057000	5.00000
25%	0.000000	0.0	0.0	0.000000	2.535500	60.25000
50%	0.000000	0.0	0.0	0.000000	6.889500	162.50000
75%	0.000000	0.0	0.0	0.000000	15.410750	459.00000
max	23.000000	0.0	0.0	8.000000	64.401000	1167.00000

	format	re_subj	exclaim_subj	urgent_subj	exclaim_mess
count	50.000000	50.000000	50.000000	50.0	50.000000
mean	0.740000	0.280000	0.060000	0.0	4.420000
std	0.443087	0.453557	0.239898	0.0	7.661433
min	0.000000	0.000000	0.000000	0.0	0.000000
25%	0.250000	0.000000	0.000000	0.0	1.000000
50%	1.000000	0.000000	0.000000	0.0	1.500000
75%	1.000000	1.000000	0.000000	0.0	4.000000
max	1.000000	1.000000	1.000000	0.0	43.000000

Scatterplot

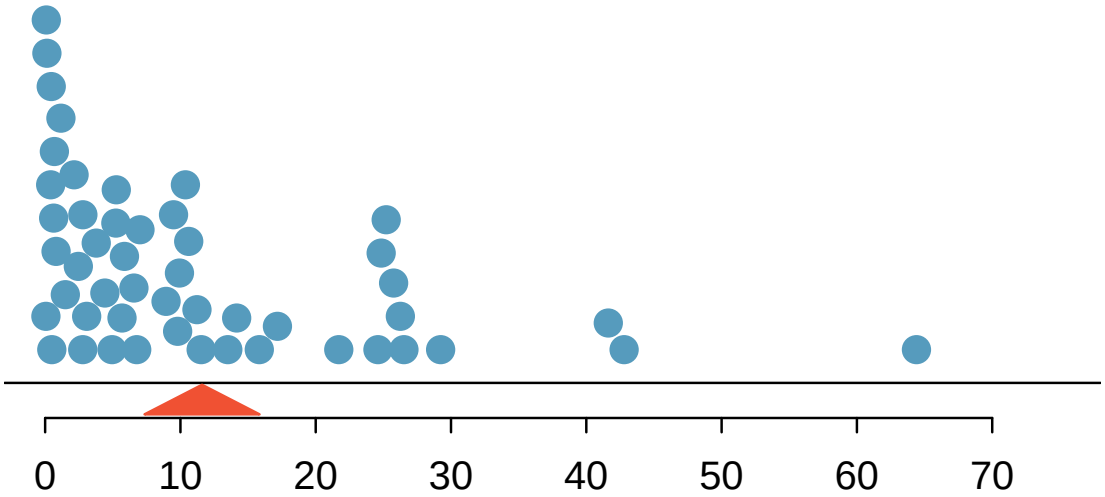
```
import matplotlib.pyplot as plt
plt.scatter(email50["num_char"], email50["line_breaks"])
plt.xlabel('# of Characters (in thousands)')
plt.ylabel('# of Lines')
```



Sample Mean ($\bar{x}$)

```
import numpy as np
np.mean(email50["num_char"])
```

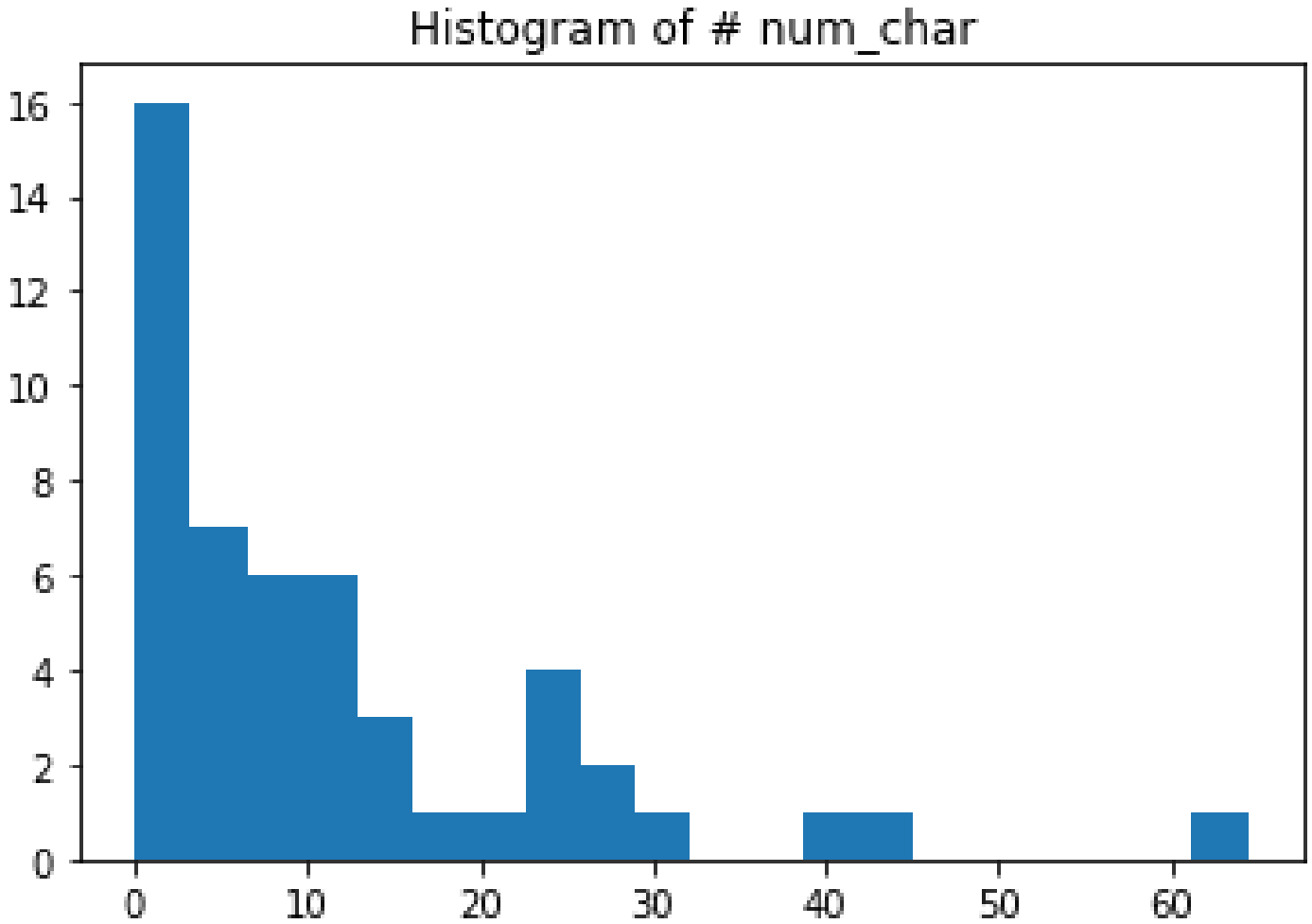
$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = 11.6$$



Histogram

plt.hist(email50["num_char"], bins=20)

plt.title('Histogram of # num_char')



Population Variance (σ^2) and Standard Deviation (σ)

```
np.var(email50["num_char"])
```

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = 168.83$$

```
np.std(email50["num_char"])
```

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = 12.98$$

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

from statistics import variance, stdev

variance(email50["num_char"])

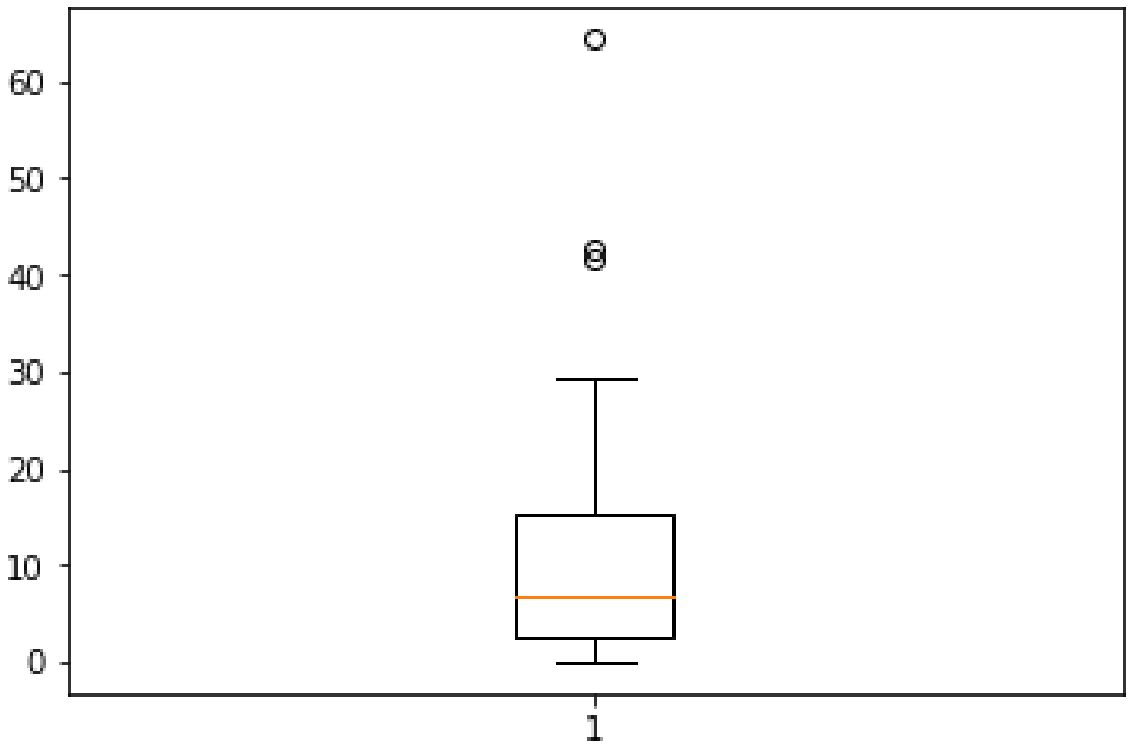
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = 172.27$$

stdev(email50["num_char"])

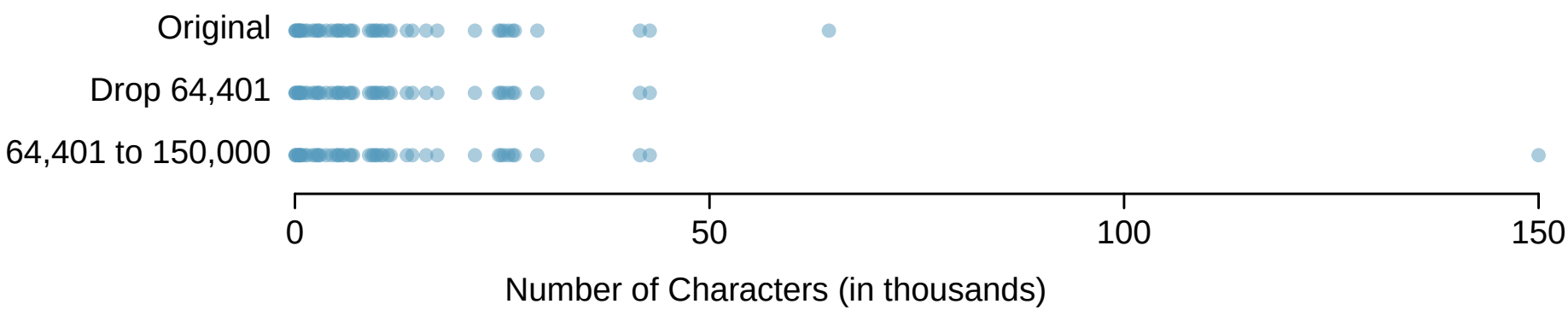
$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = 13.12$$

Boxplot

```
np.median(email50["num_char"])          6.9
from scipy import stats
scipy.stats.iqr(email50["num_char"])    12.87
plt.boxplot(email50["num_char"])
```



Robust Statistics

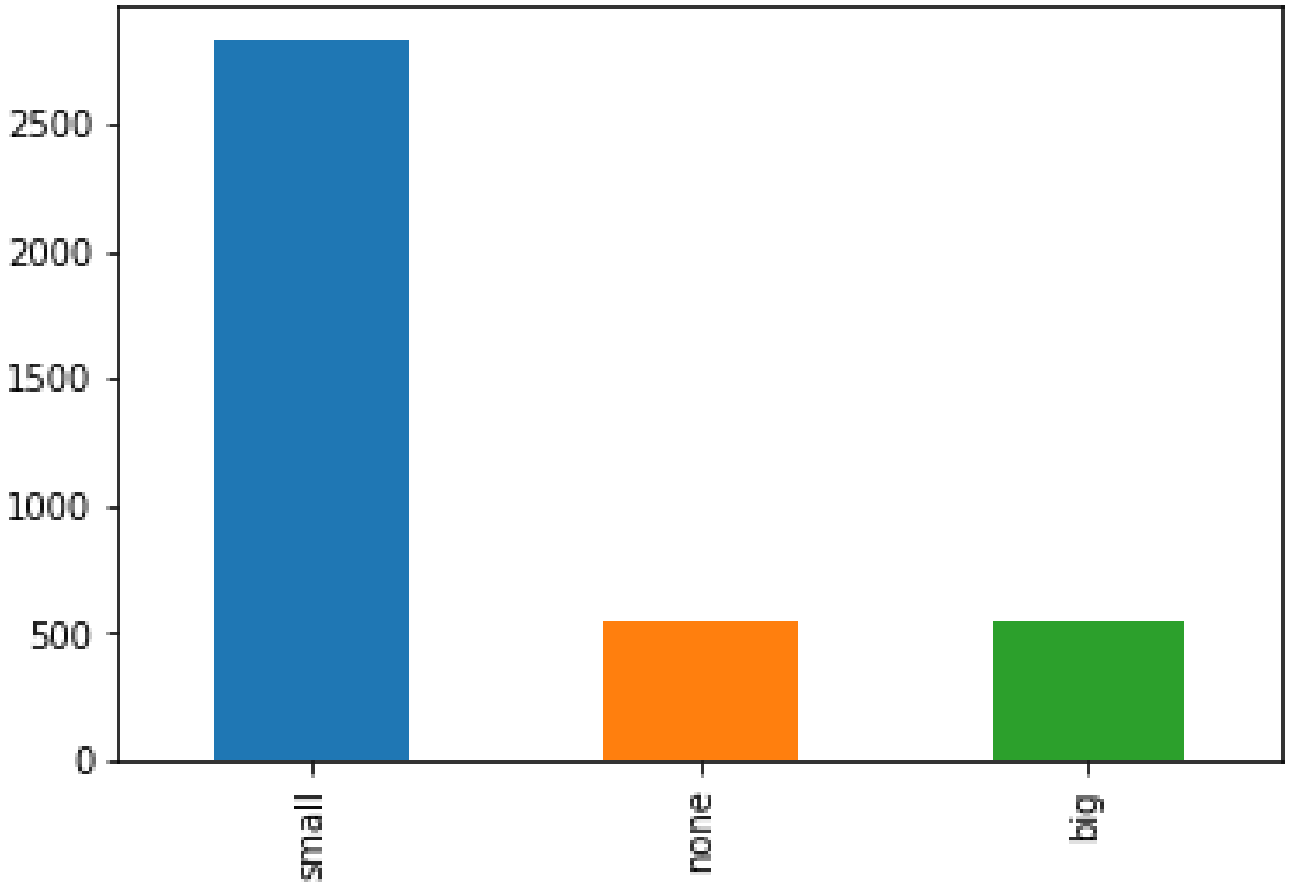


	robust		not robust	
scenario	median	IQR	$\bar{x}$	s
original `num_char` data	6,890	12,875	11,600	13,130
drop 66,924 observation	6,768	11,702	10,521	10,798
move 66,924 to 150,000	6,890	12,875	13,310	22,434

Barplot

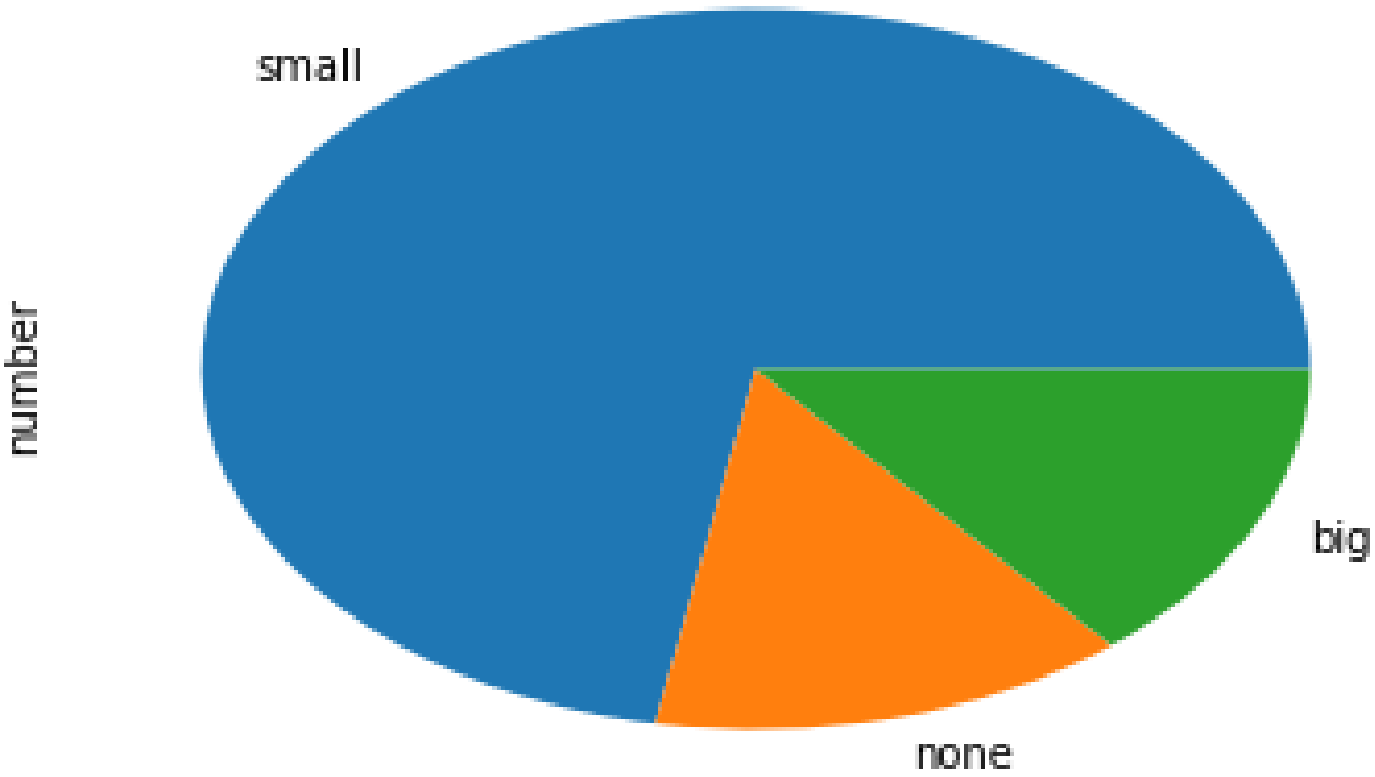
email = pd.read_csv("https://github.com/VitorKamada/ECO5100/raw/master/Data/email.csv")

email["number"].value_counts().plot(kind='bar')



Pie Chart

email["number"].value_counts().plot(kind='pie')



Frequency Table

pd.crosstab(email["spam"], columns="count")

pd.crosstab(email["number"], columns="count")

spam	count
0	3554
1	367

number	count
big	545
none	549
small	2827

Contingency Table

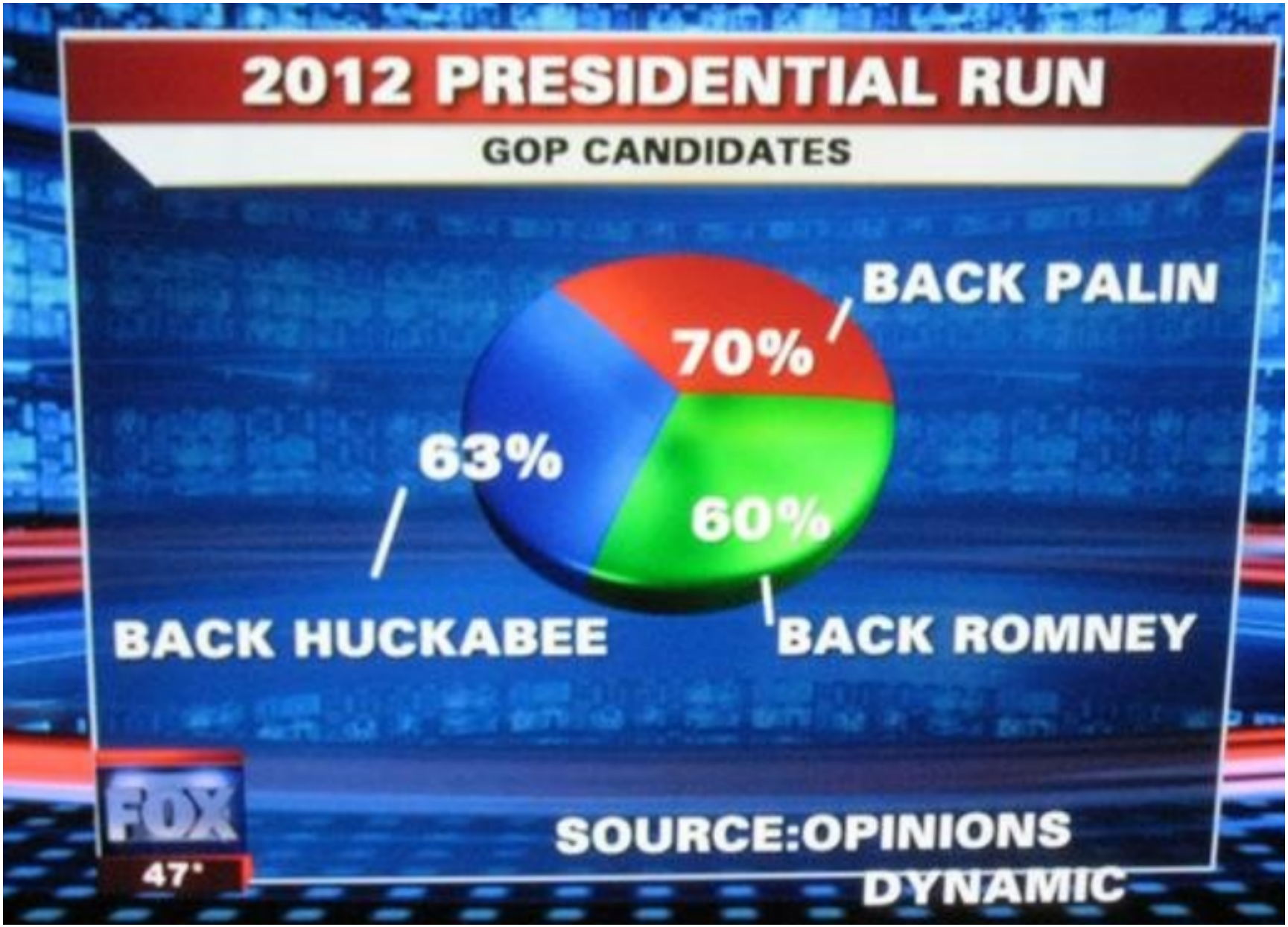
pd.crosstab(email["spam"], columns=email["number"], margins=True)

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

pd.crosstab(email["spam"], columns=email["number"], normalize='columns')

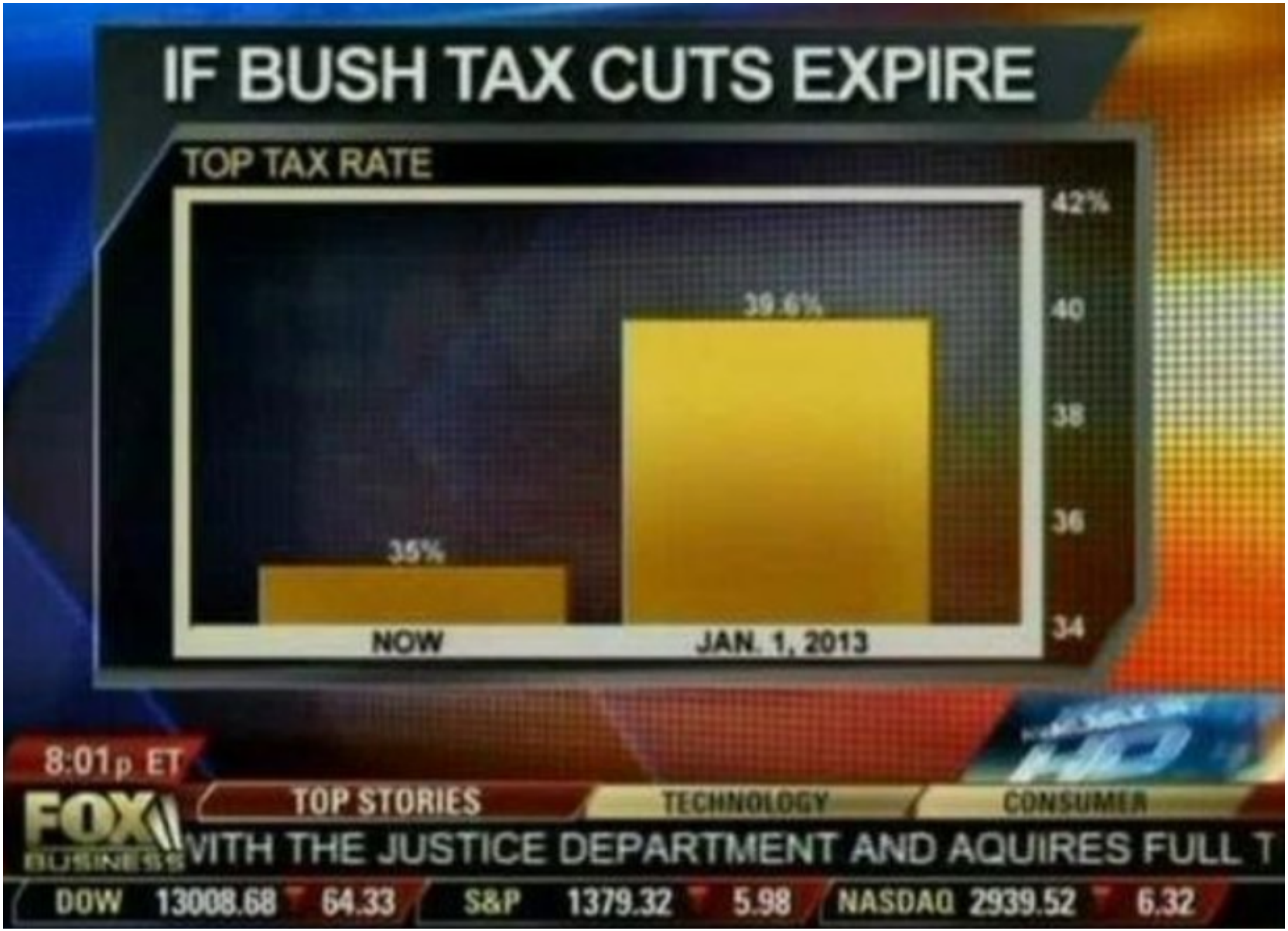
number	big	none	small
spam			
0	0.908257	0.728597	0.940573
1	0.091743	0.271403	0.059427

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