GTK CYBER

GET ON THE FAST TRACK

Data Science for Security Professionals - Day 2

Statistical Summaries and Data Visualization

GTK CYBER
GET ON THE FAST TRACK

Explore Your Data

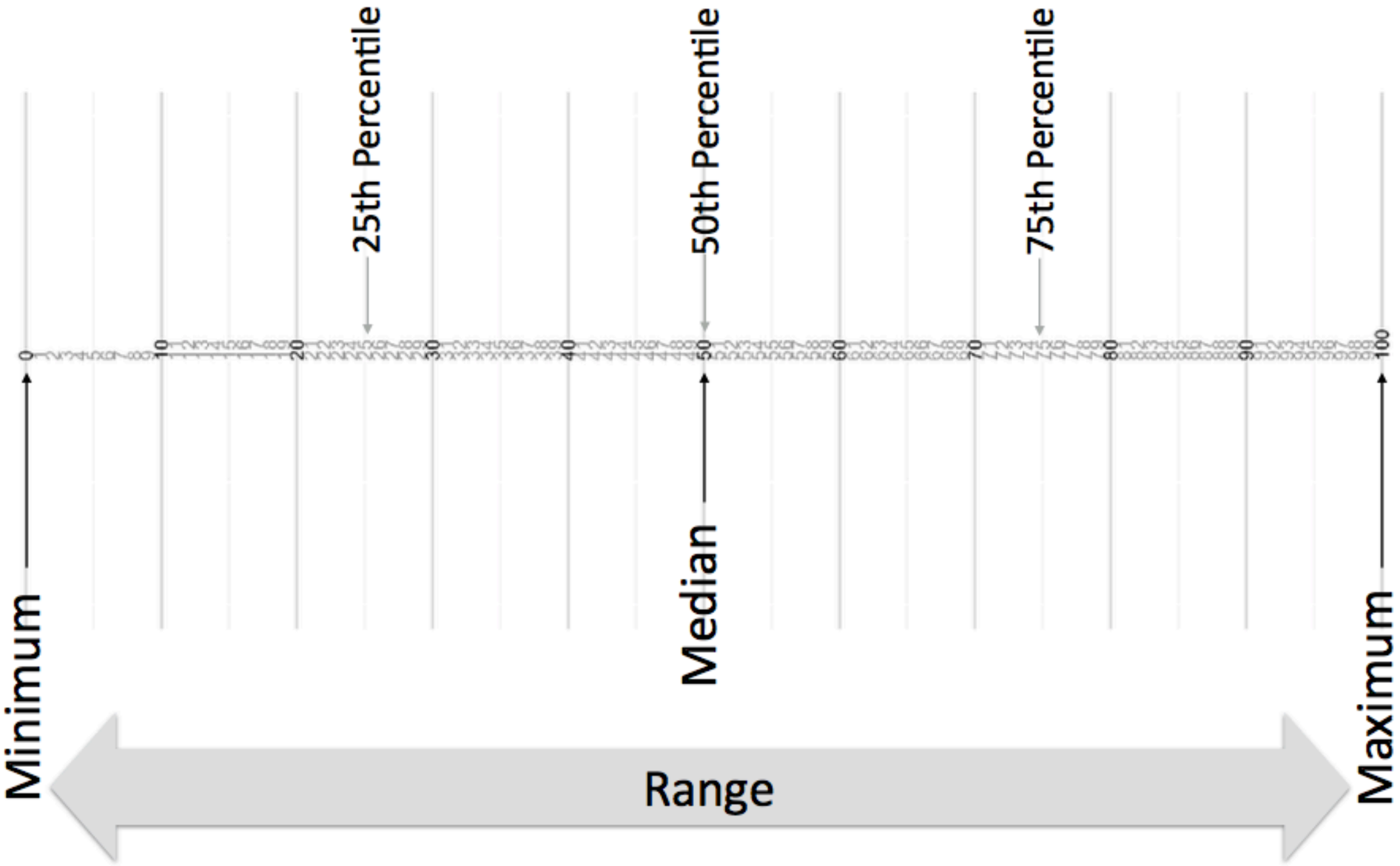
- **Mean**: The mean is the average value of a set of numbers

- **Median**: The median is the middle value of an ordered set of numbers

- **Mode**: The mode is the value from a set which is repeated the most frequently.

- **Range**: Difference between minimum and maximum

- **Standard Deviation**: Measures dispersion in a data set. Close to zero indicates little dispersion.


Minimum
0
10
20
25th Percentile
30
40
Median
50
60
70
75th Percentile
80
90
100
Maximum
Range

`Series.abs()`	Absolute Value of the Series
`Series.count()`	Returns number of non-empty values in the series
`Series.max()`	Returns maximum value in the Series
`Series.mean()`	Returns the mean of a Series
`Series.median()`	Returns the median of a Series
`Series.min()`	Returns the minimum value in a Series
`Series.mode()`	Take a guess..
`Series.quantile([q])`	Returns the quantiles of a Series
`Series.sum`	Returns the sum of a series
`Series.std`	Returns the standard deviation of a Series

Tukey 5 Number Summary

- **Minimum:** The smallest value in the dataset
- **Lower Quartile:** Smallest 25% of the dataset
- **The Median:** The middle value of the dataset
- **Upper Quartile:** The largest 25% of the dataset
- **Maximum:** The largest value in the dataset

```
Series.describe()
```

Series.describe()

```
>>> random_numbers.describe()
count    50.000000
mean     50.620000
std      30.102471
min       1.000000
25%      25.500000
50%      54.000000
75%      73.000000
max      99.000000
dtype: float64
```

```
names = pd.Series(
 ['Jim','Bob',
 'Bob','Steve','Jim,'Jane','Steph','Emma','Rachel
 '])

names.describe()
```

```
names = pd.Series(
 ['Jim','Bob',
 'Bob','Steve','Jim,'Jane','Steph','Emma','Rachel
 '])

names.describe()
```

```
count        9
unique       7
top        Jim
freq         2
dtype: object
```

Finding Unique Values

```
unique_nums = []
for key, value in random_numbers.iteritems():
    if value not in unique_nums:
        unique_nums.append(value)
```

NO!!!

```
series.drop_duplicates()
series.unique()
```

Finding Unique Values

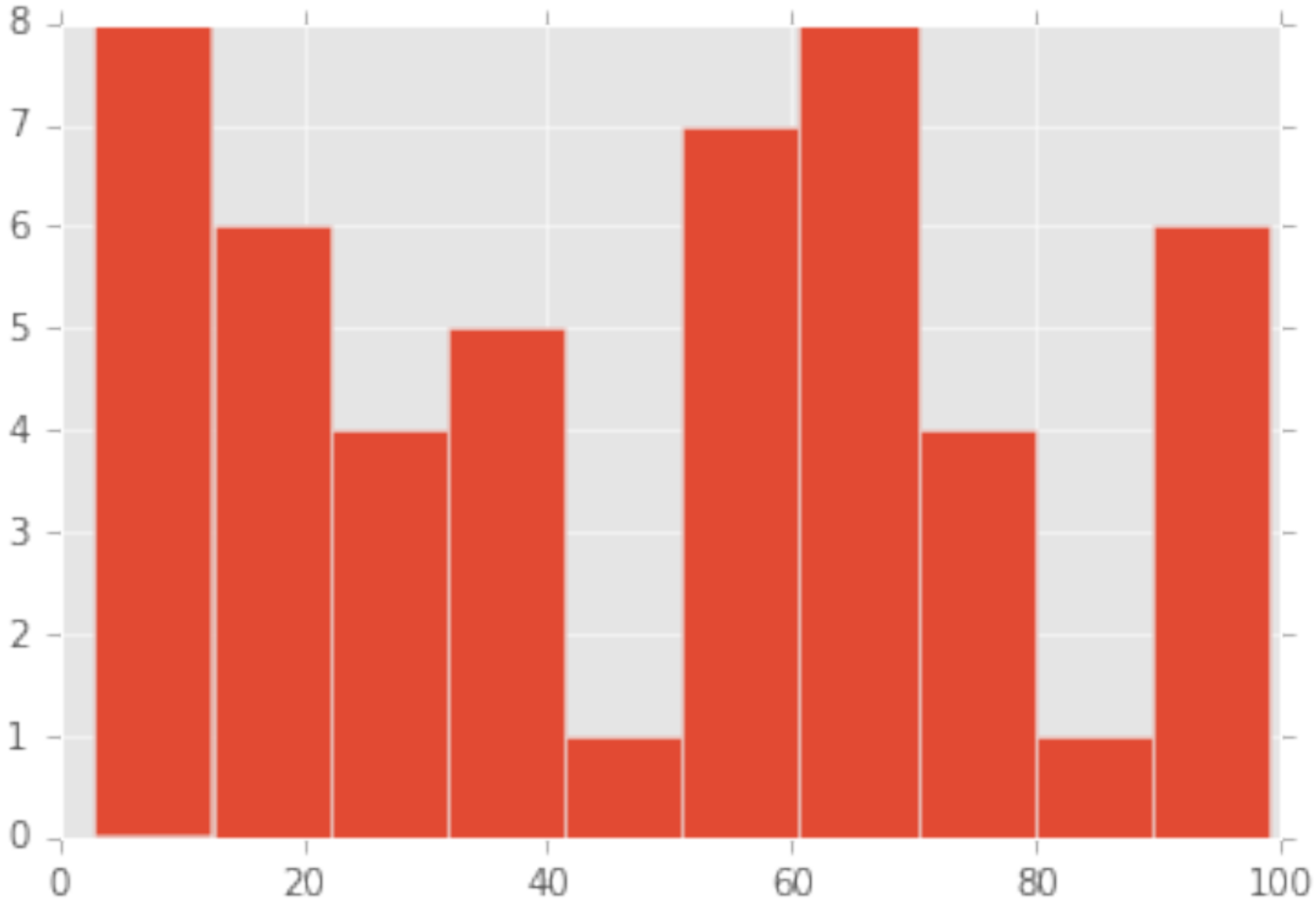
```
series.value_counts()
```

```
series.value_counts(bins=5)
```

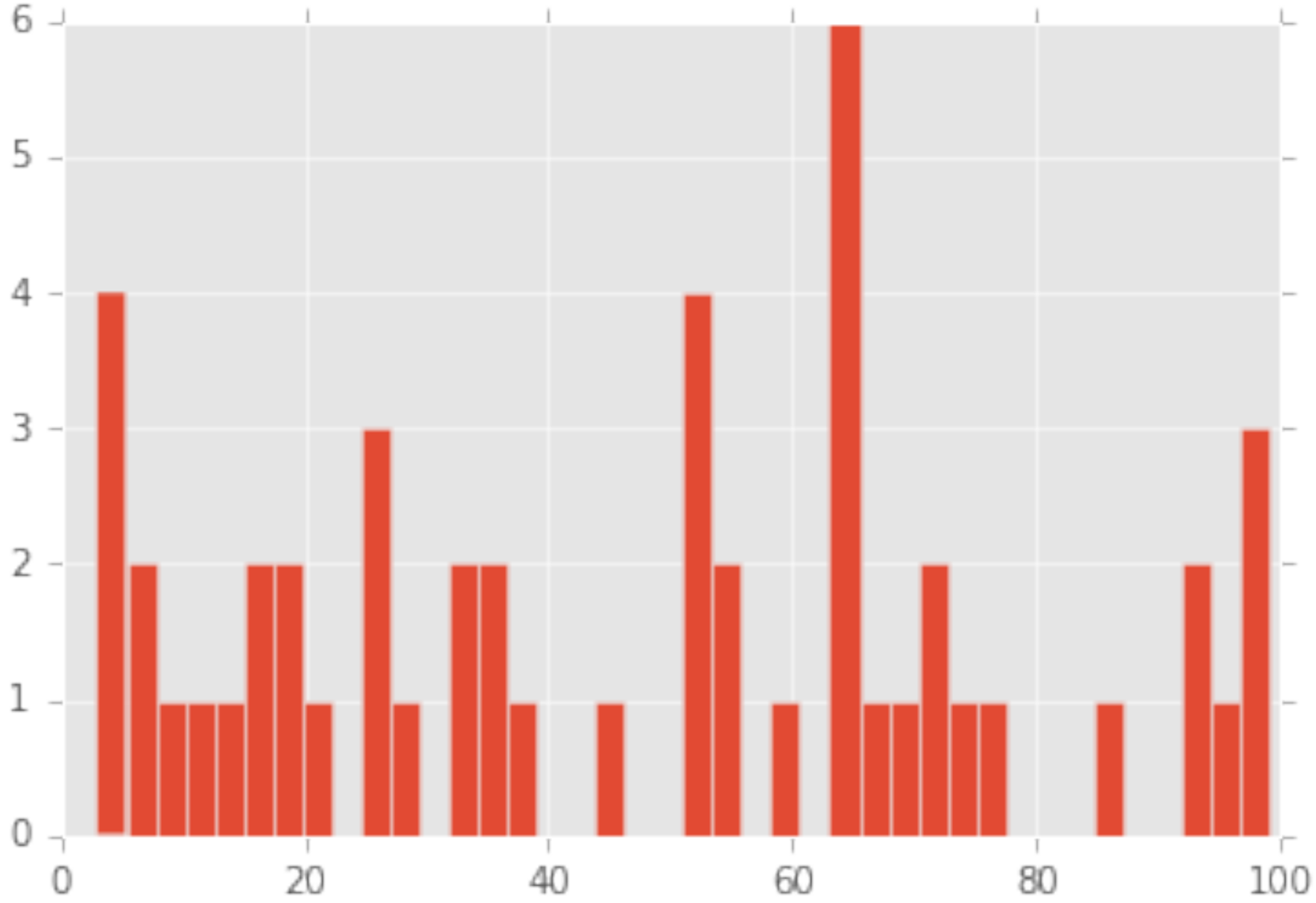
```
series.value_counts(bins=5,
normalize=True)
```

```
series.hist()
```

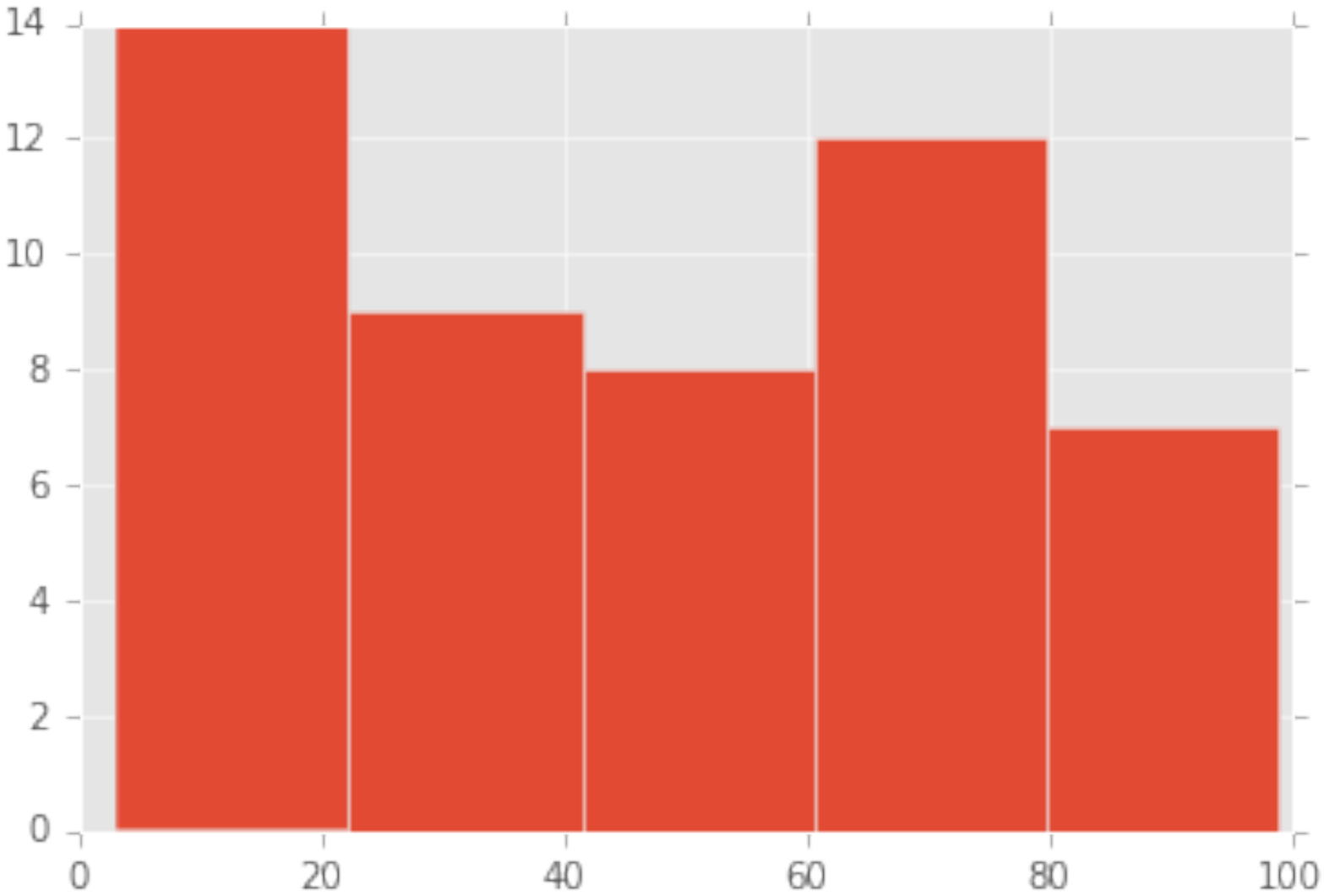
series.hist()



```
series.hist(bins=40)
```

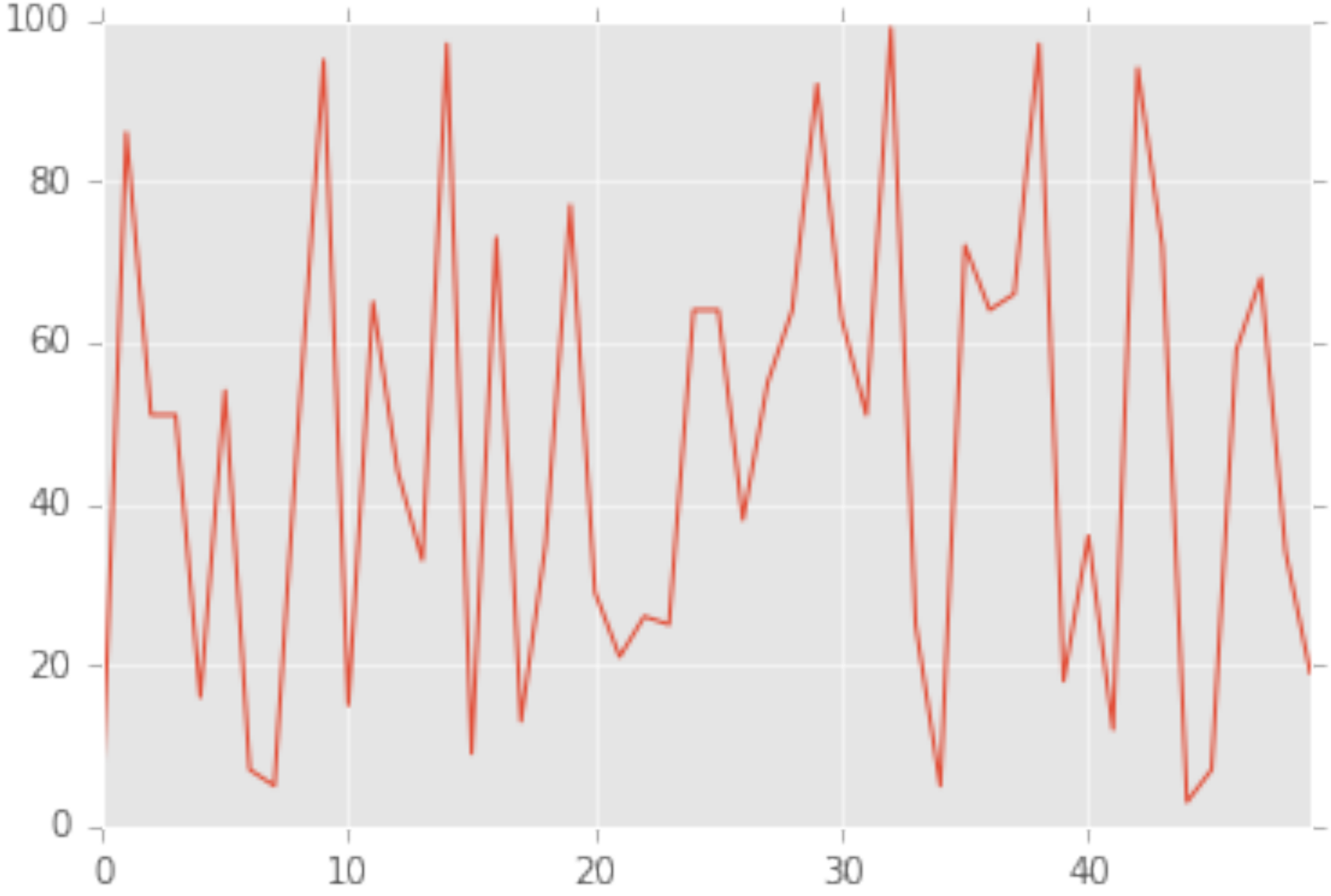


series.hist(bins=5)

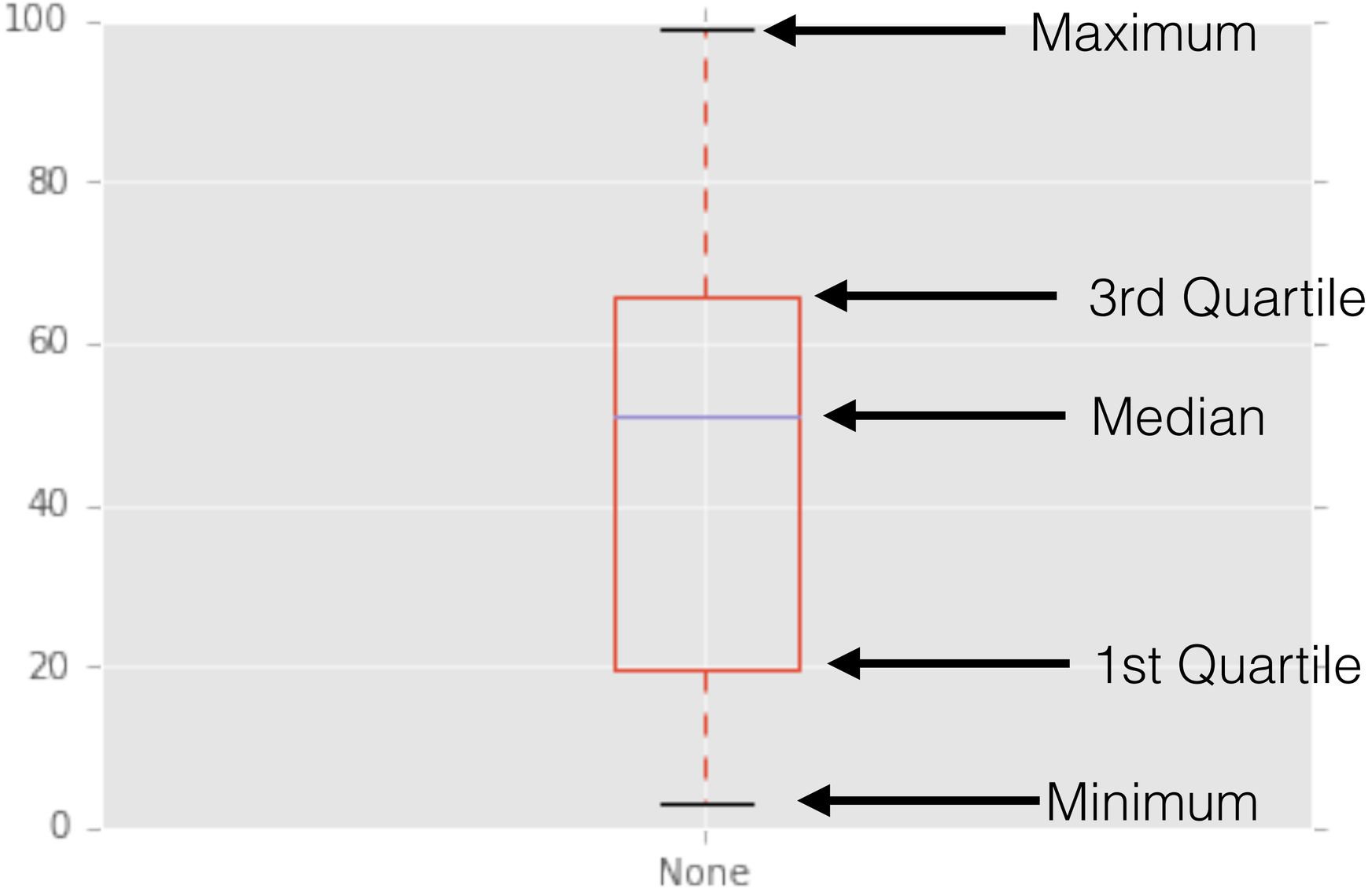


```
series.plot()
```

series.plot()

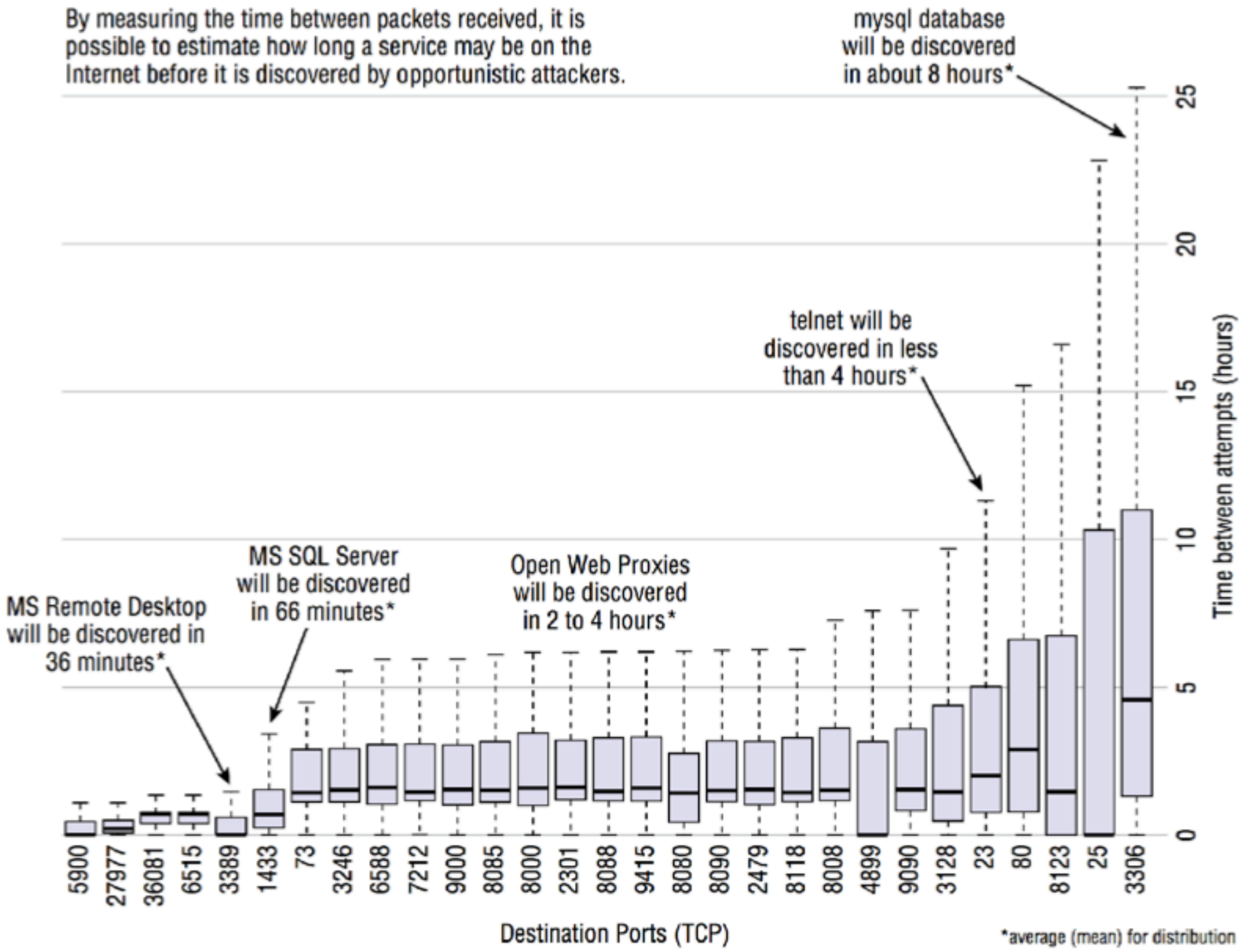


series.plot(kind='box')



How long will a service go undiscovered by opportunistic attackers?

By measuring the time between packets received, it is possible to estimate how long a service may be on the Internet before it is discovered by opportunistic attackers.

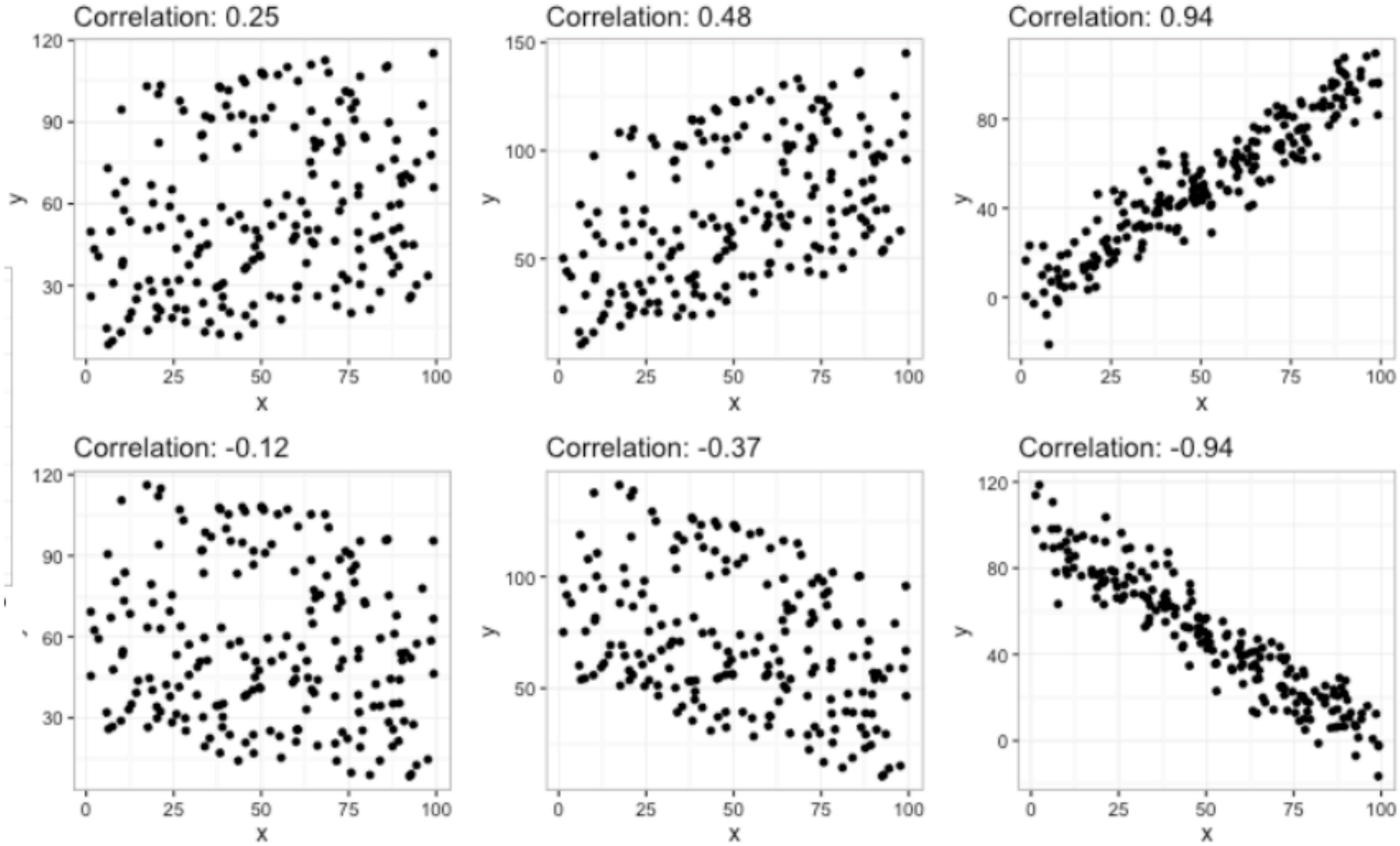


Correlations between Datasets

Correlations

- Measurement of the relationship between **two continuous variables**
- Output is between -1 and 1, with 1 being perfect correlation, -1 is perfect negative correlation, 0 is no correlation.
- Correlation measurement is referred to as r.

Correlations



Correlations

```
series1.corr( series2 )
```

In Class Exercise:

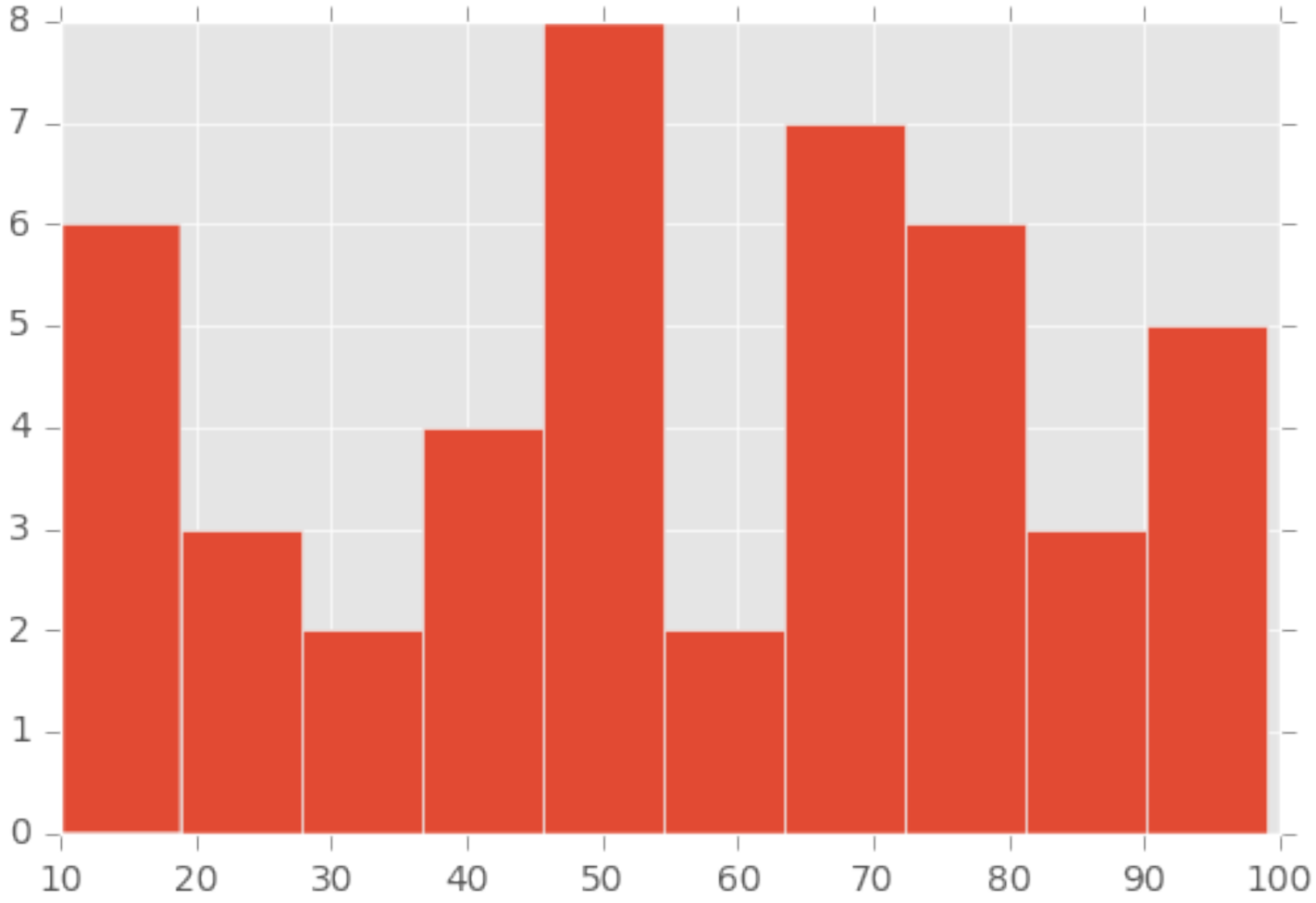
Please complete Worksheet 3: EDA Worksheet

```
#Filter the Series
random_numbers = random_numbers[random_numbers >= 10]

#Sort the Series
random_numbers.sort_values(inplace=True)

#Calculate the Tukey 5 Number Summary
random_numbers.describe()

#Count the number of even and odd numbers
even_numbers = random_numbers[random_numbers % 2 == 0].count()
odd_numbers = random_numbers[random_numbers % 2 != 0].count()
print( "Even numbers: " + str(even_numbers))
print( "Odd numbers: " + str(odd_numbers))

#Find the five largest and smallest numbers
print( "Smallest Numbers:")
print( random_numbers.head(5))

print( "Largest Numbers:")
print( random_numbers.tail(5))
```

```
Even numbers: 22
Odd numbers: 20
Smallest Numbers:
19    10
36    13
48    15
4     15
23    15
dtype: int64
Largest Numbers:
26    88
41    88
3     90
32    92
27    95
dtype: int64
```

Exercise 2

```
random_numbers.hist(bins=10)
```



Exercise 3

```
phone_number_series = pd.Series(phone_numbers)
area_codes = phone_number_series.str.slice(1,4)
area_codes2 = phone_number_series.str.extract( '\((\d{3})\)',
expand=False)
area_codes2.value_counts()
```

```
833    7
811    5
822    5
899    3
844    3
855    2
dtype: int64
```

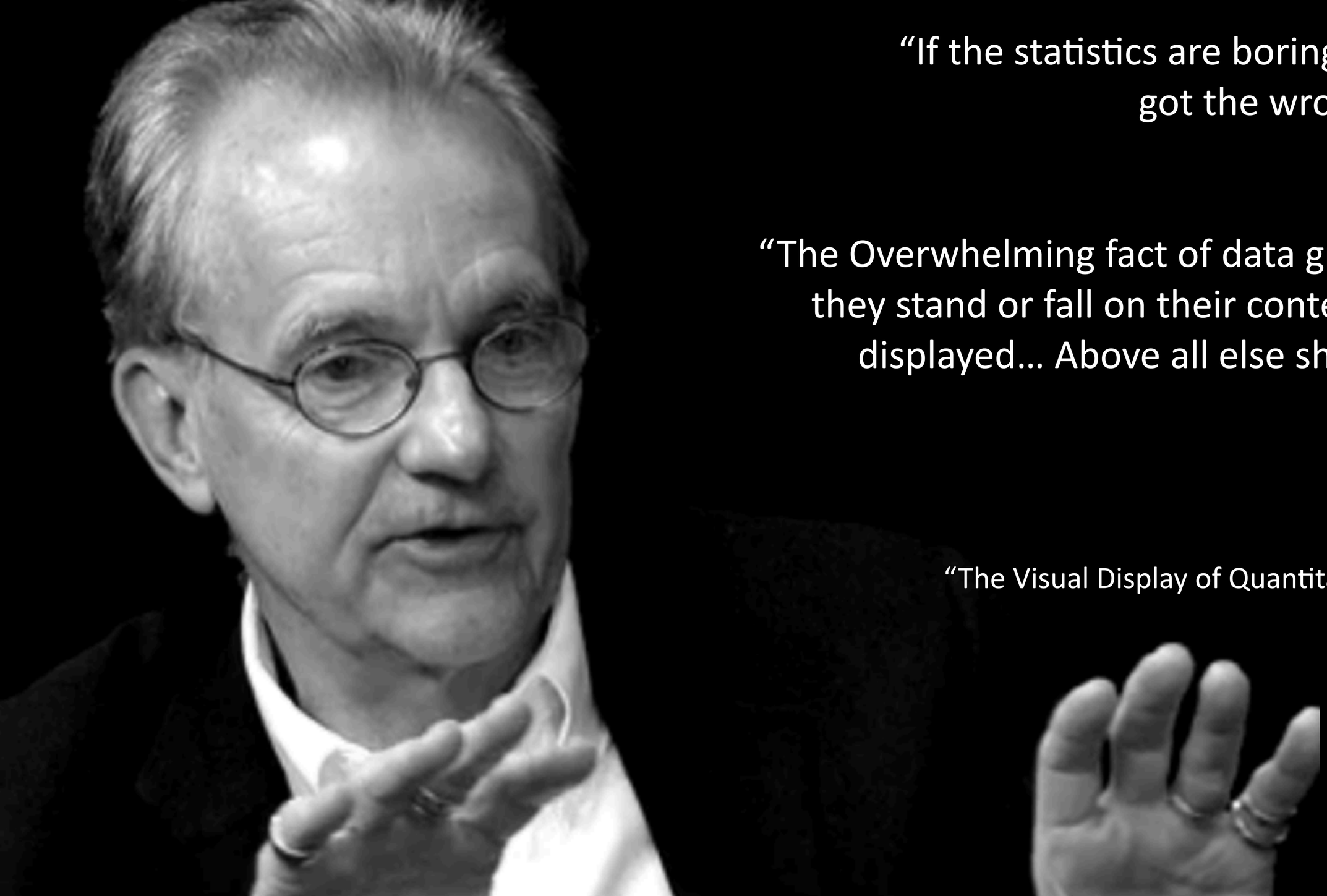
Questions?

Exploring Data Through Visualization

Visualization Goals

- Analyze
- Explore
- Assess
- Determine
- Decide

- Communicate
- Explain
- Present
- Prove
- Persuade



“If the statistics are boring, then you’ve got the wrong numbers.”

“The Overwhelming fact of data graphics is that they stand or fall on their content, gracefully displayed… Above all else show the data.”

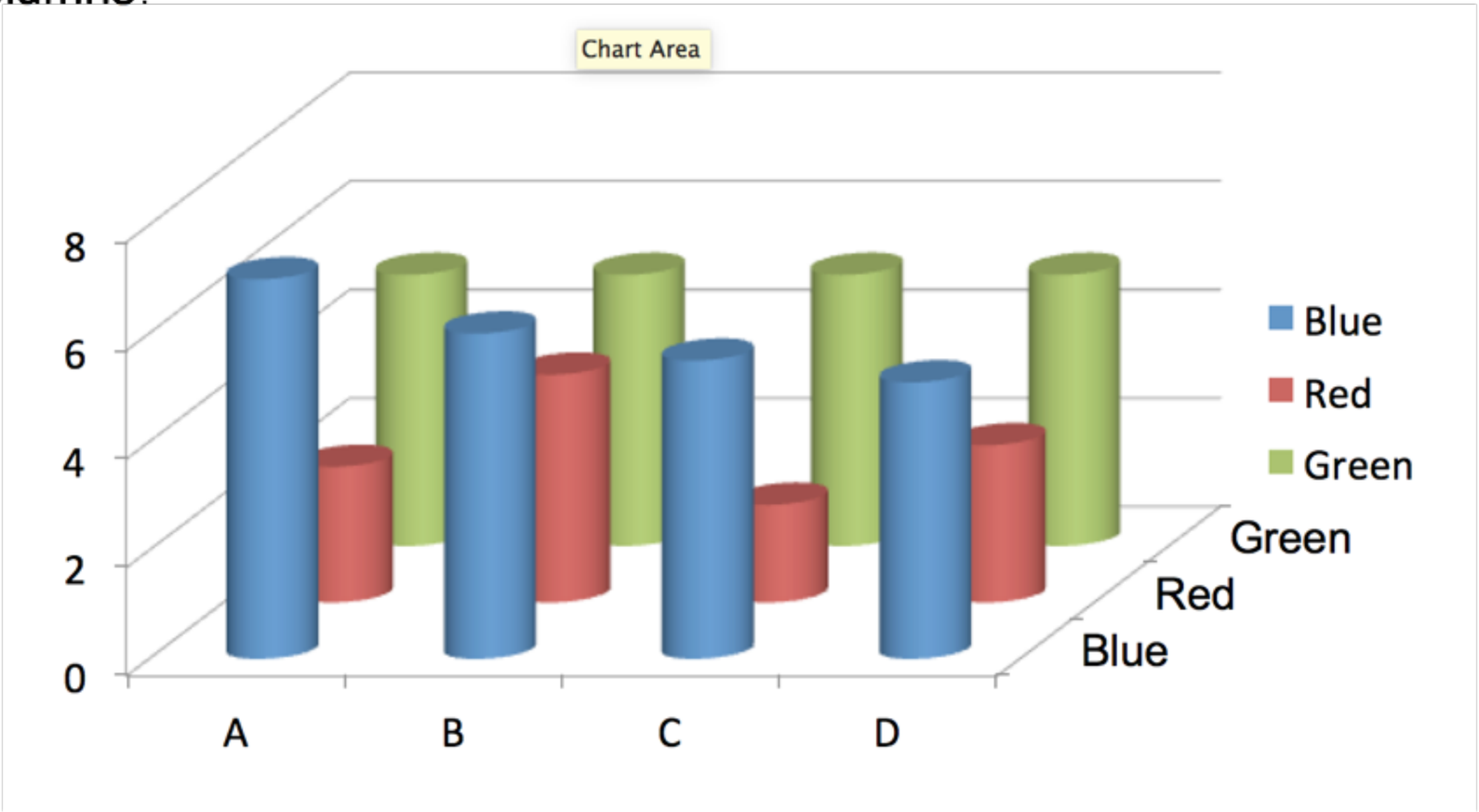
Edward R. Tufte
“The Visual Display of Quantitative Information"

Elements of Good Visualizations

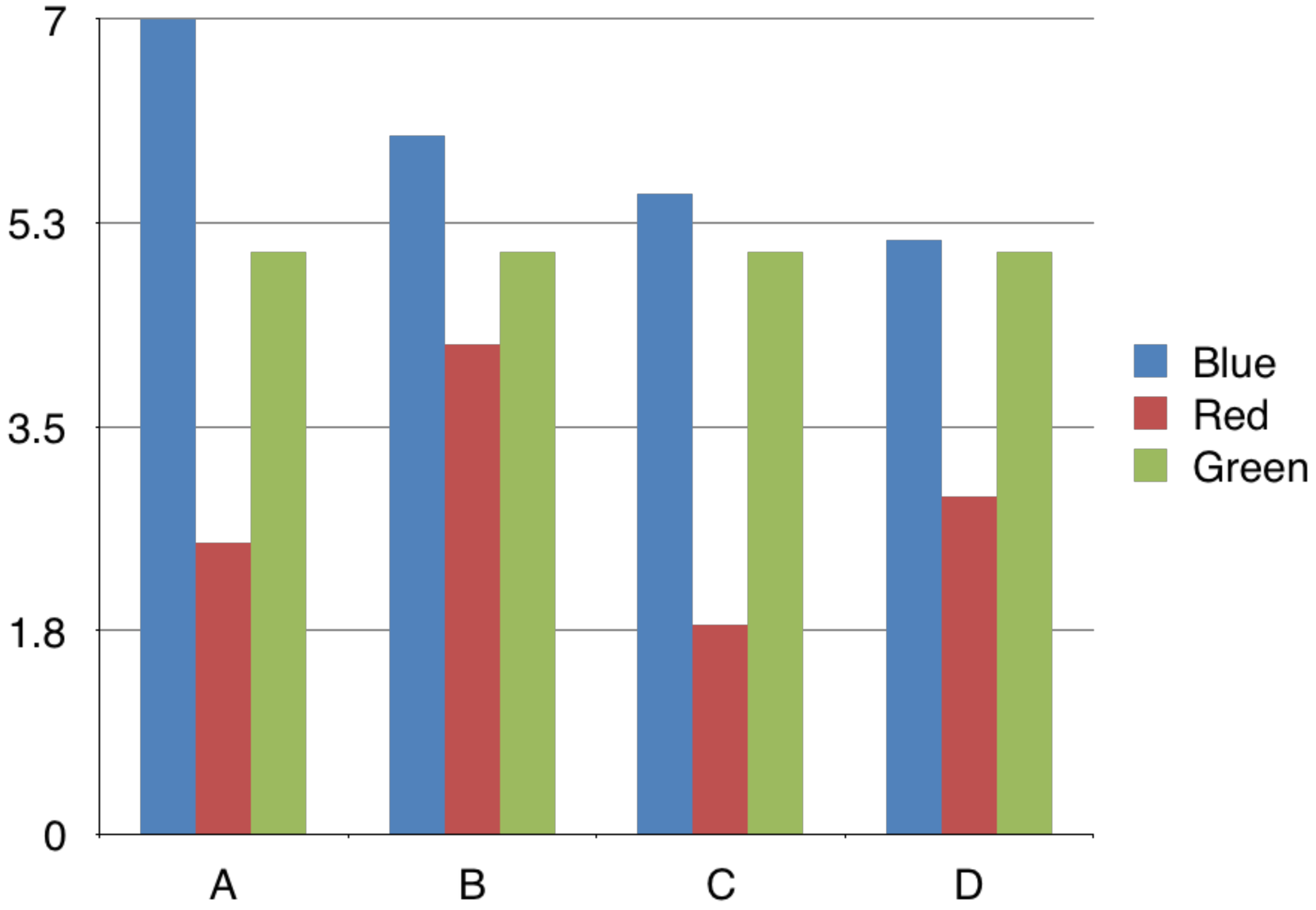
1. Graphical Integrity
2. Simple
3. Proper Display
4. Proper Color
5. Tells a story

Proper Display

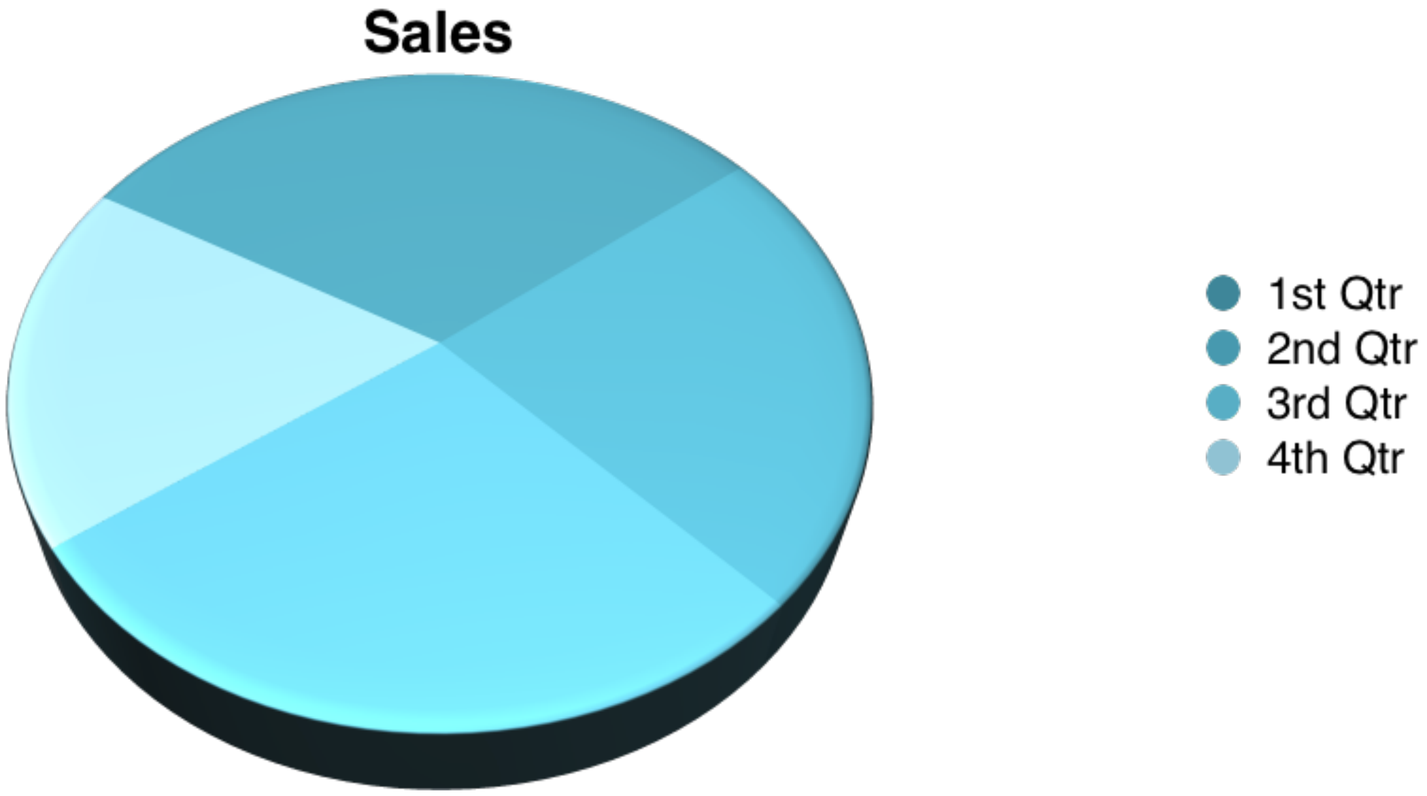
Questions: What is the height of the green columns? For which categories (A,B,C,D) are the blue columns taller than the green columns?



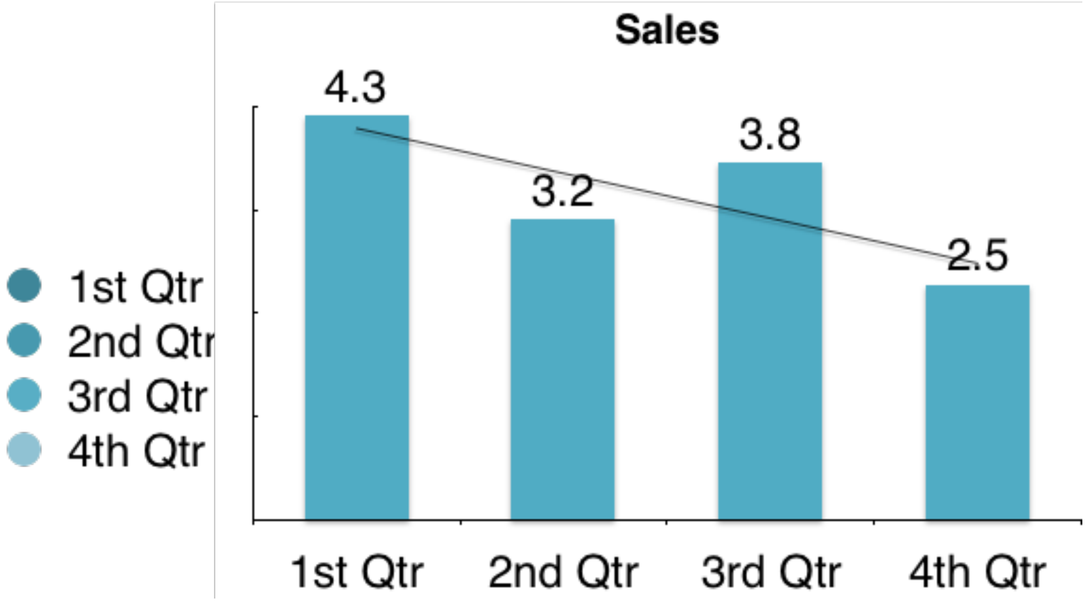
Proper Display



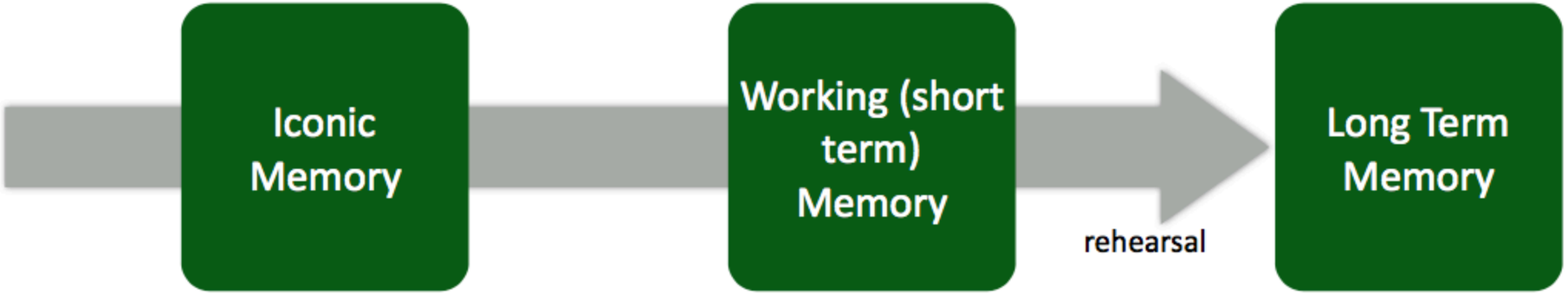
Proper Display



Proper Display



Visual Processing System



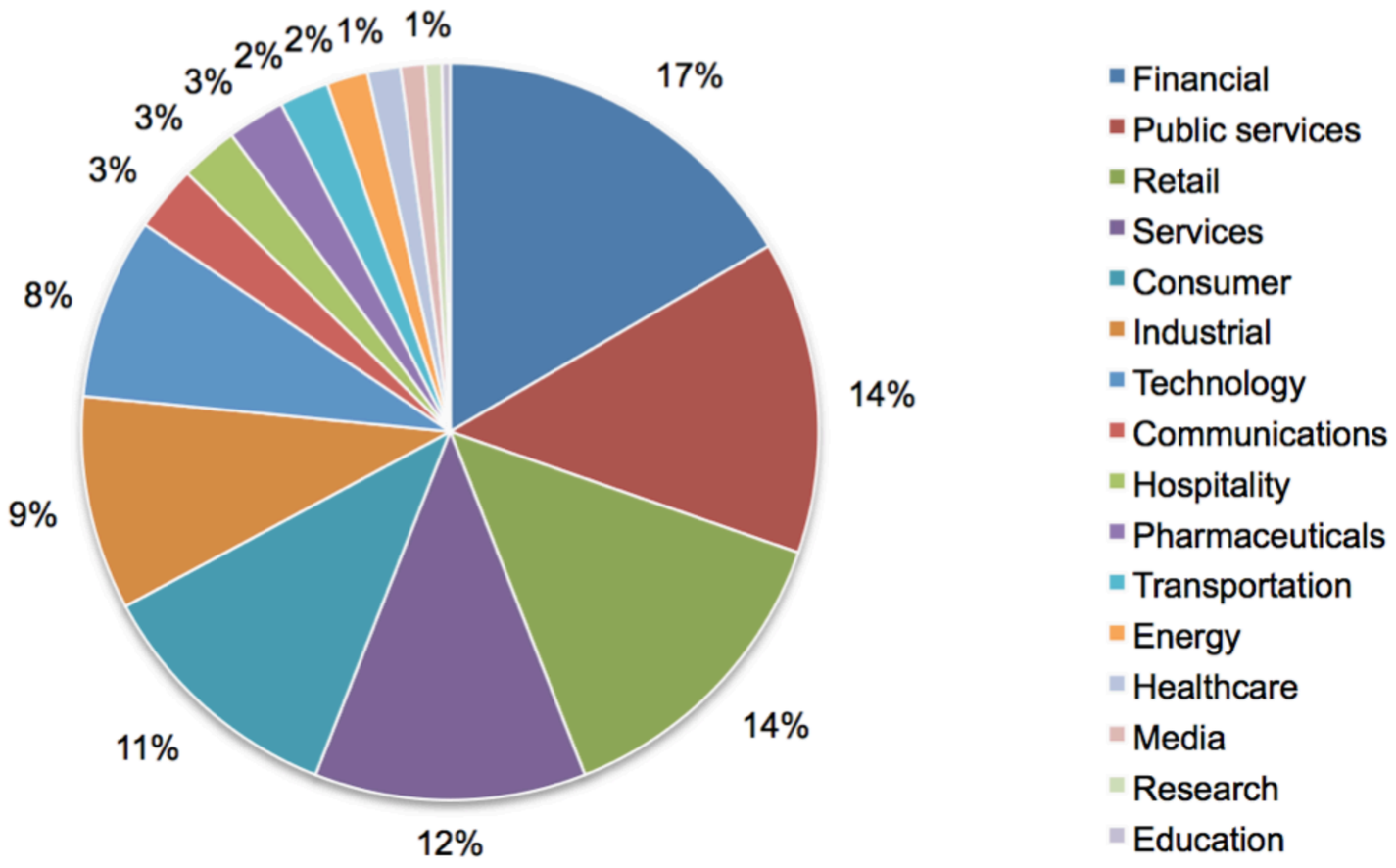
- < 1 second (or until replaced)
- Preattentive processing
- Saccadic movement
- ~20 objects

- Visual groupings
- Flexible, conscious
- 3-5 objects

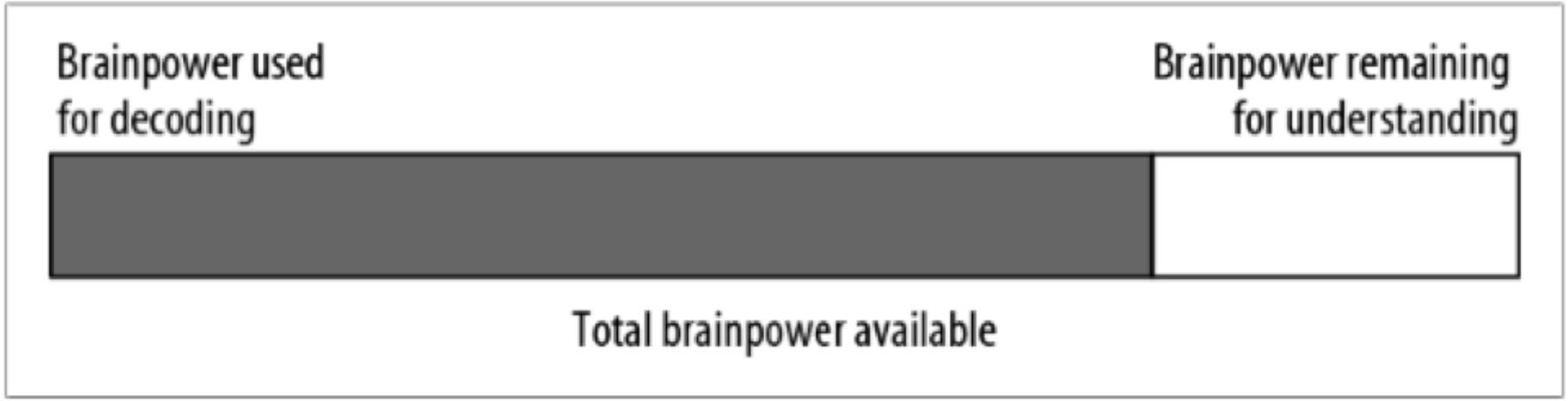
- Affects dataviz through expectations

Overworking Visual Memory

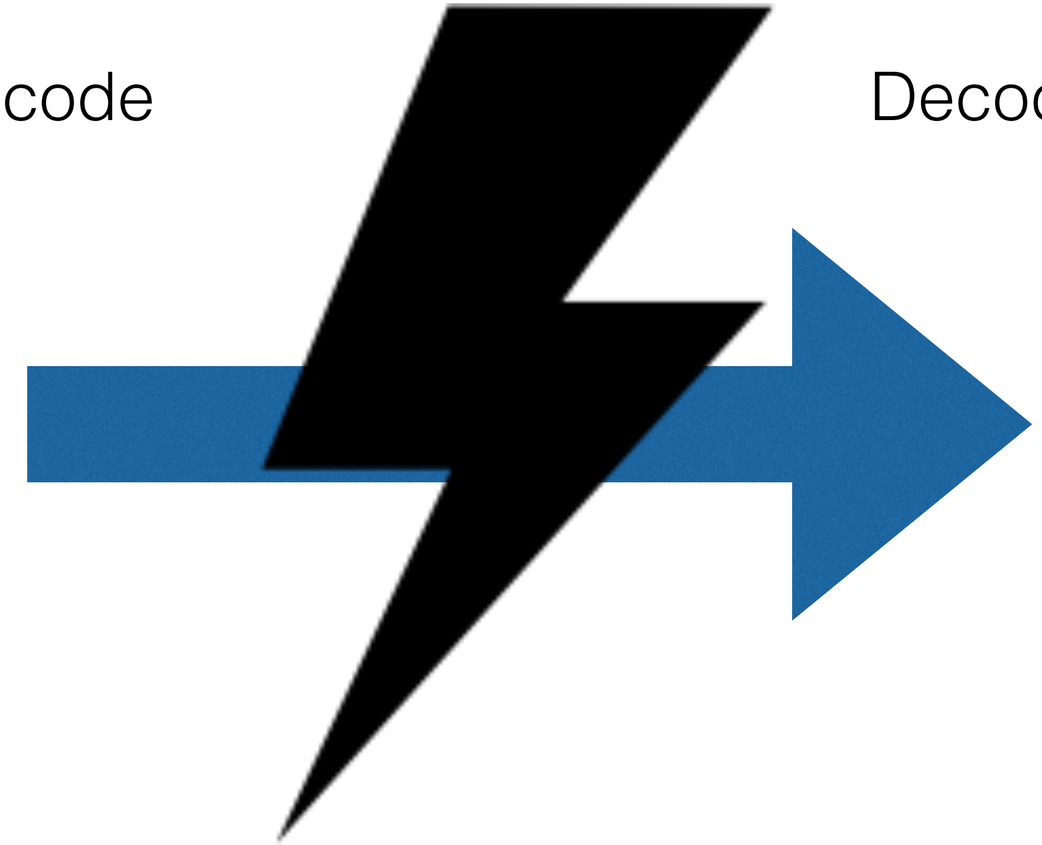
Figure 20. Distribution of the benchmark sample by industry segment
Consolidated (n = 277 organizations)



https://www4.symantec.com/mktginfo/whitepaper/053013_GL_NA_WP_Ponemon-2013-Cost-of-a-Data-Breach-Report_daiNA_cta72382.pdf

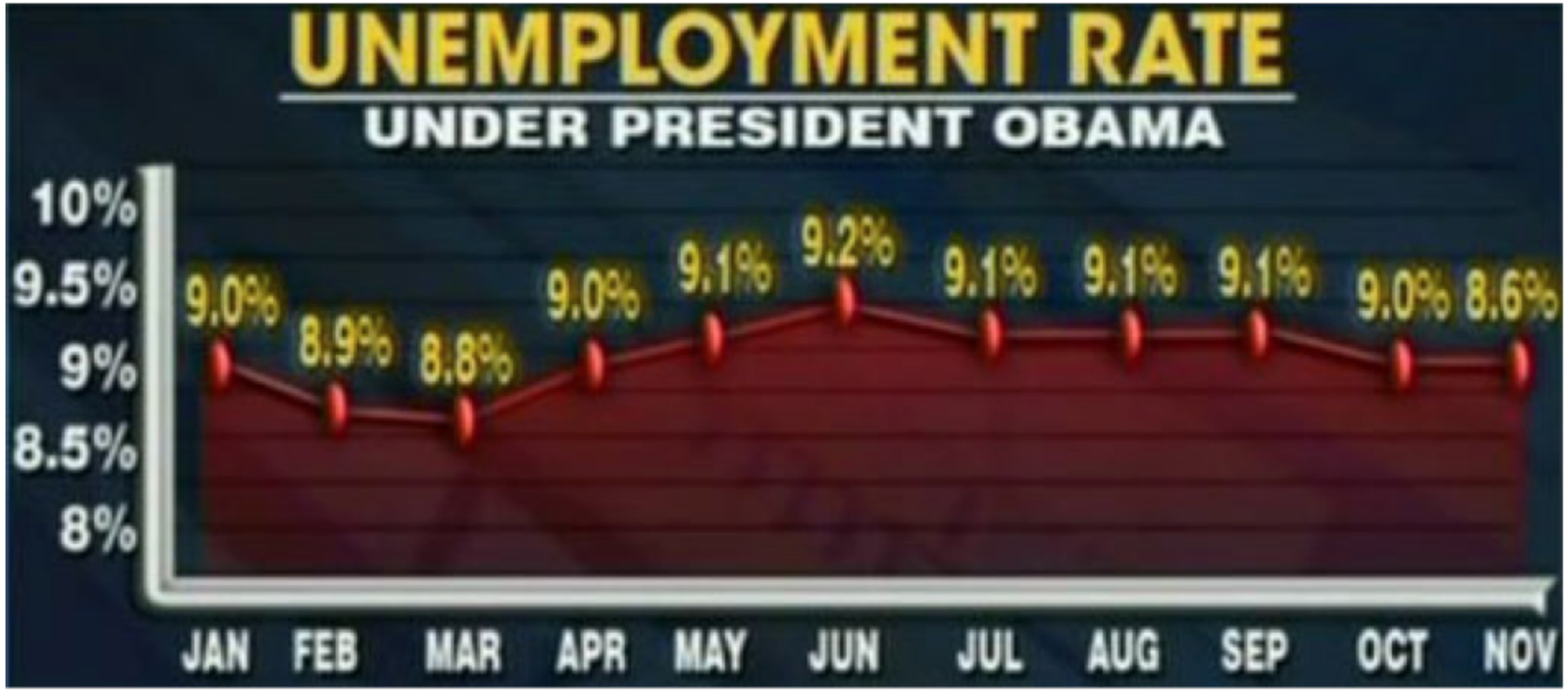

Brainpower used
for decoding
Brainpower remaining
for understanding
Total brainpower available

Encode



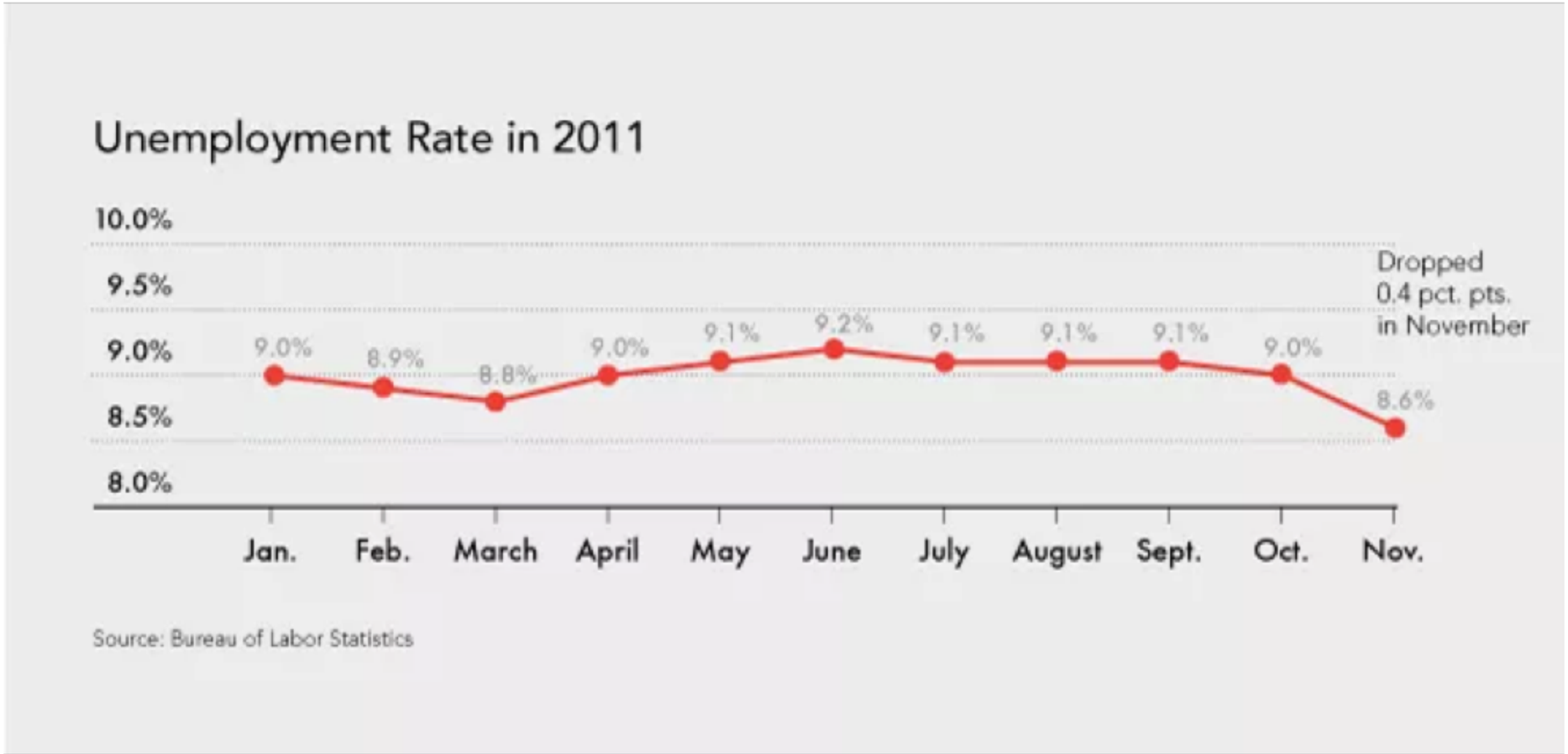
Decode

Graphical Integrity



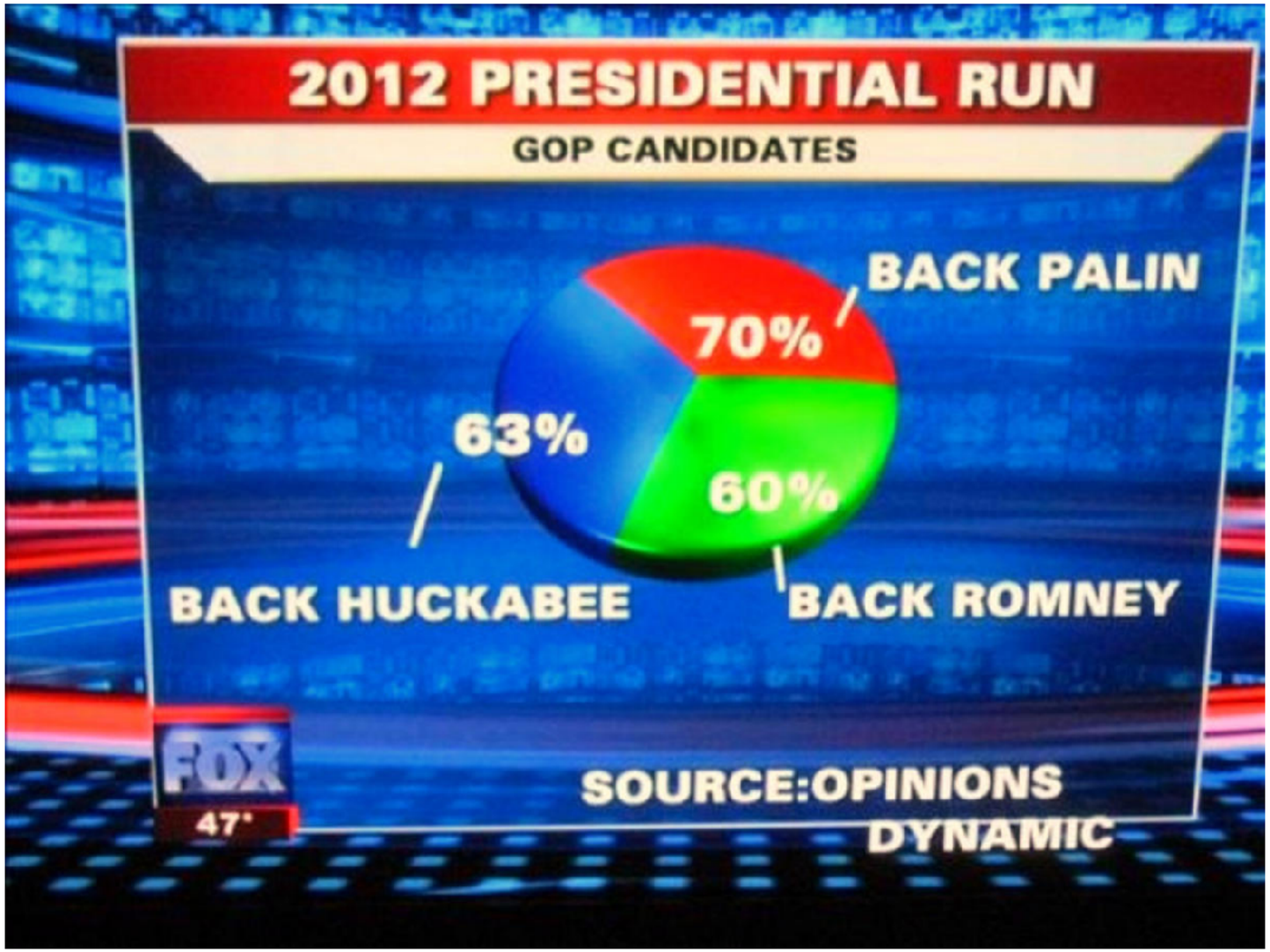
flowingdata.com

Graphical Integrity



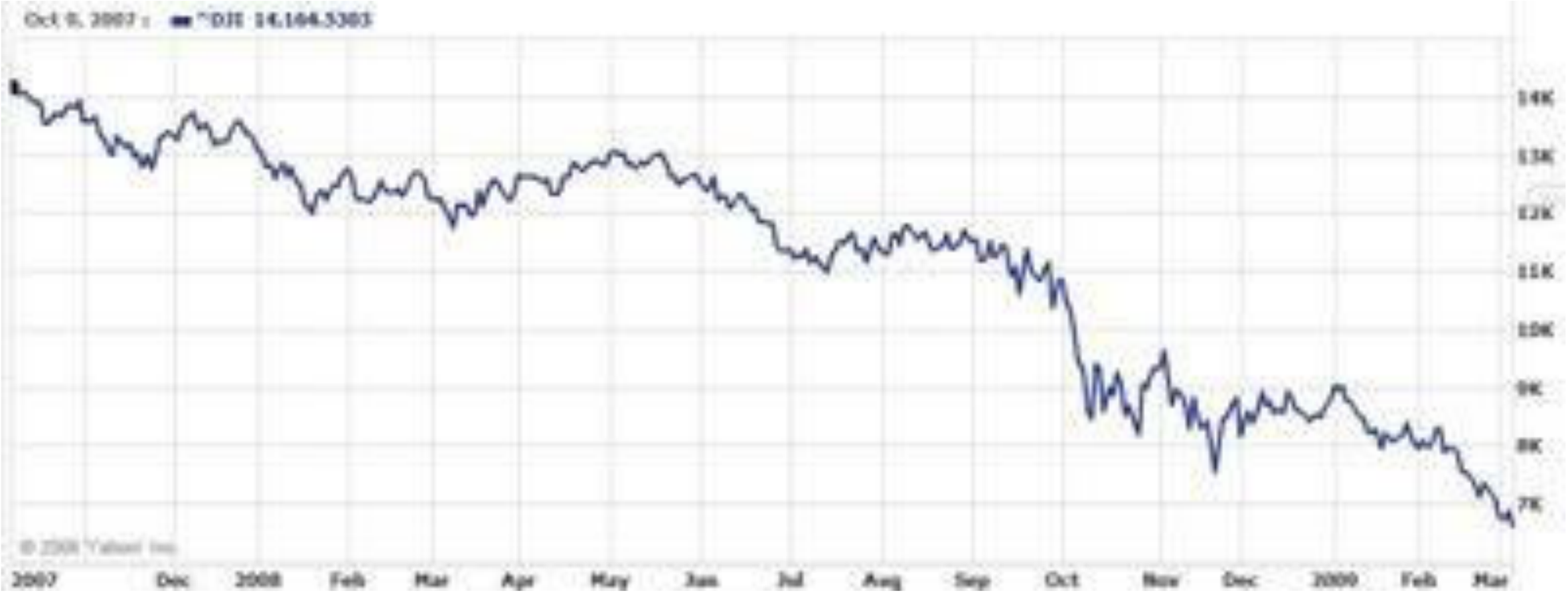
flowingdata.com

Graphical Integrity?

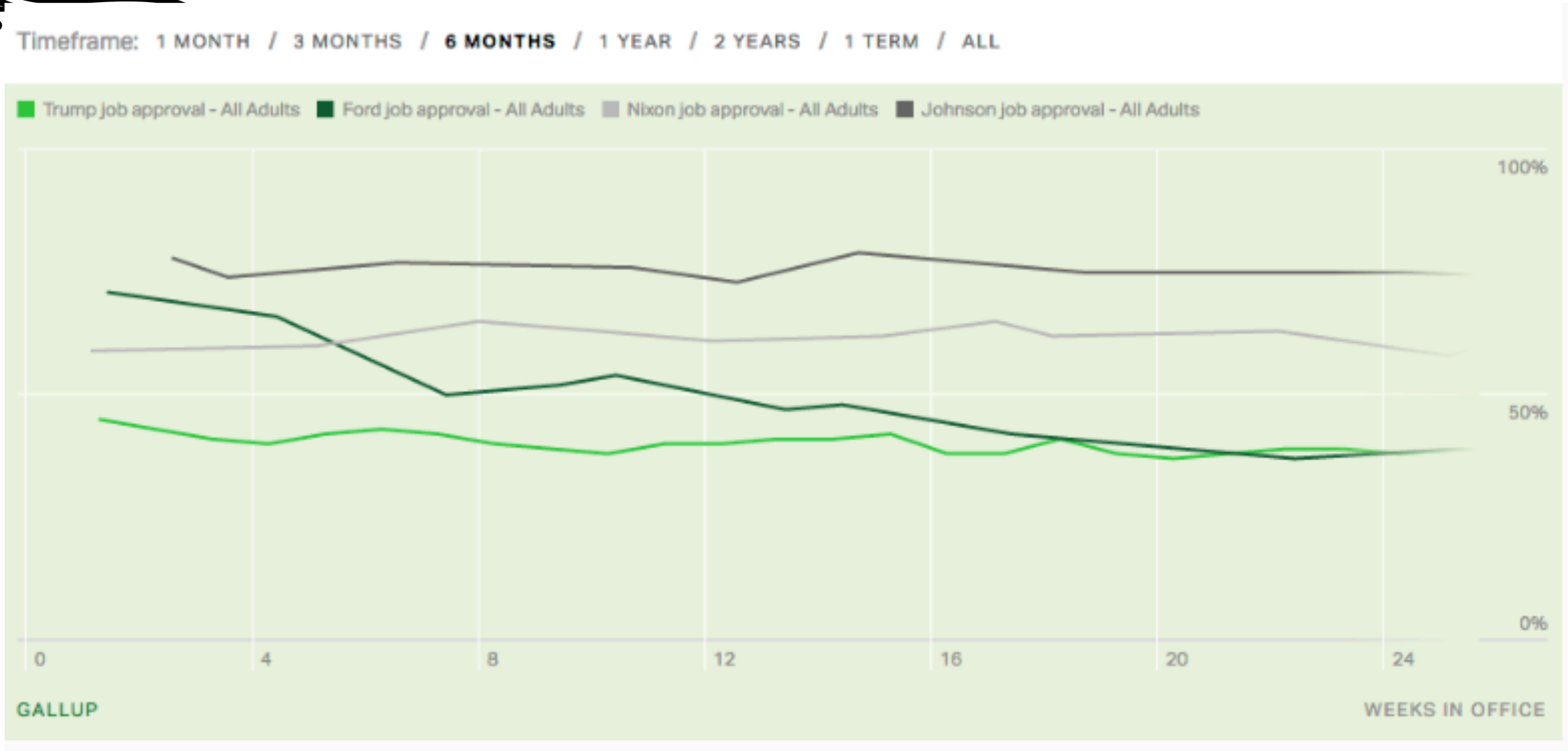


Graphical Integrity

STOCK MARKET SLIDE
NOV. 4, 2008
ELECTION DAY
9625.30
FRIDAY
MAR. 5, 2009
6626.94
JAN. 20, 2009
INAUGURATION DAY
7949.10
SOURCE: BUSINESSWEEK
6:31 CT
MSNBC

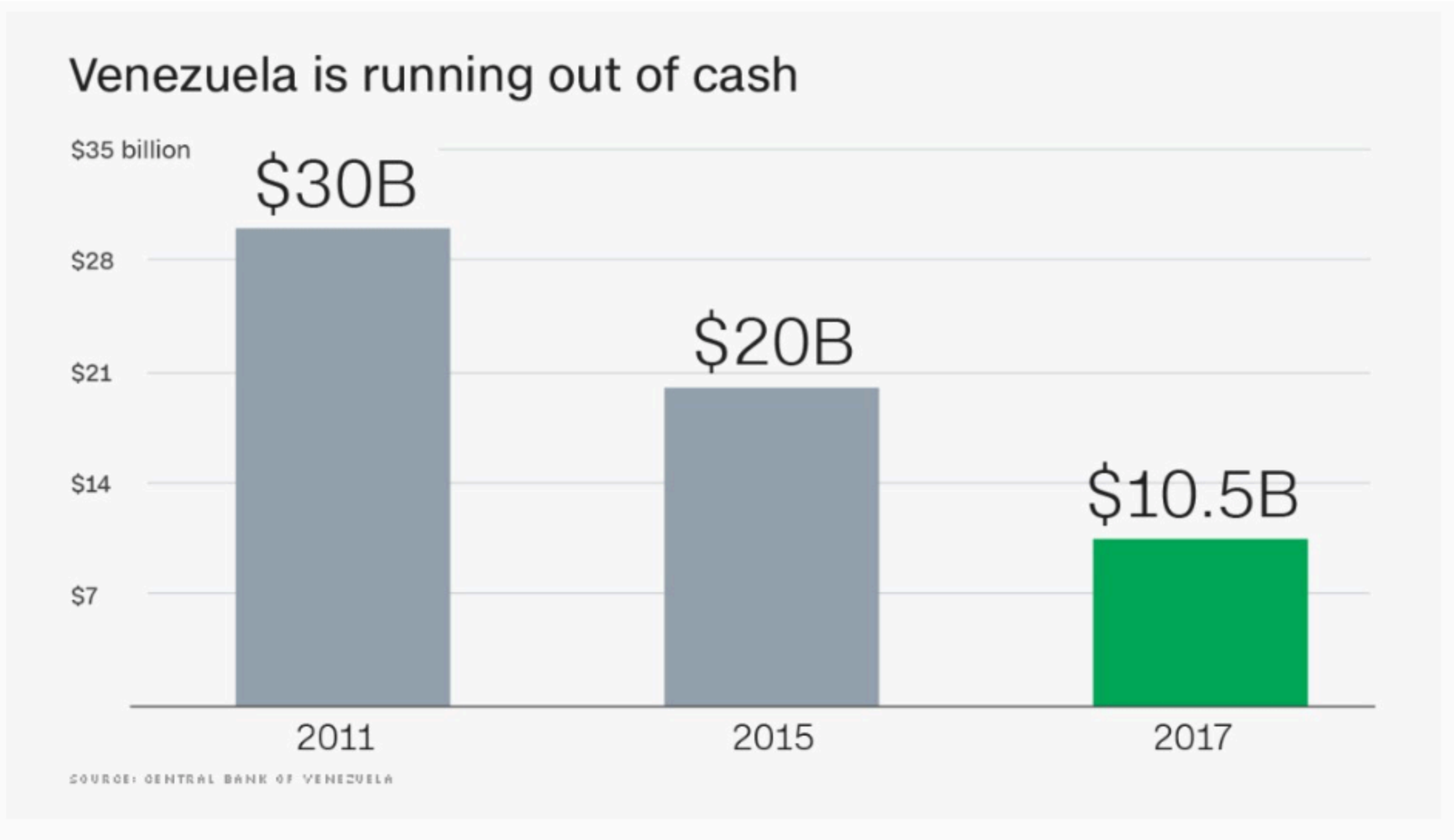

14K
13K
12K
11K
10K
9K
8K
7K
2007
Dec
2008
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2009
Feb
Mar

GEORGIA
PRESIDENT
FEELINGS ABOUT THE FEDERAL GOVERNMENT
SATISFIED
D CLINTON 81%
R TRUMP 14%
L JOHNSON 4%
CNN EXIT POLL
26:23
NEXT POLLS CLOSE
LIVE
CNN
CNN ELECTORAL MAP
19 TRUMP
3 CLINTON
IN PRESIDENT
R TRUMP 126,791 65.4%
D CLINTON 59,284 30.6%

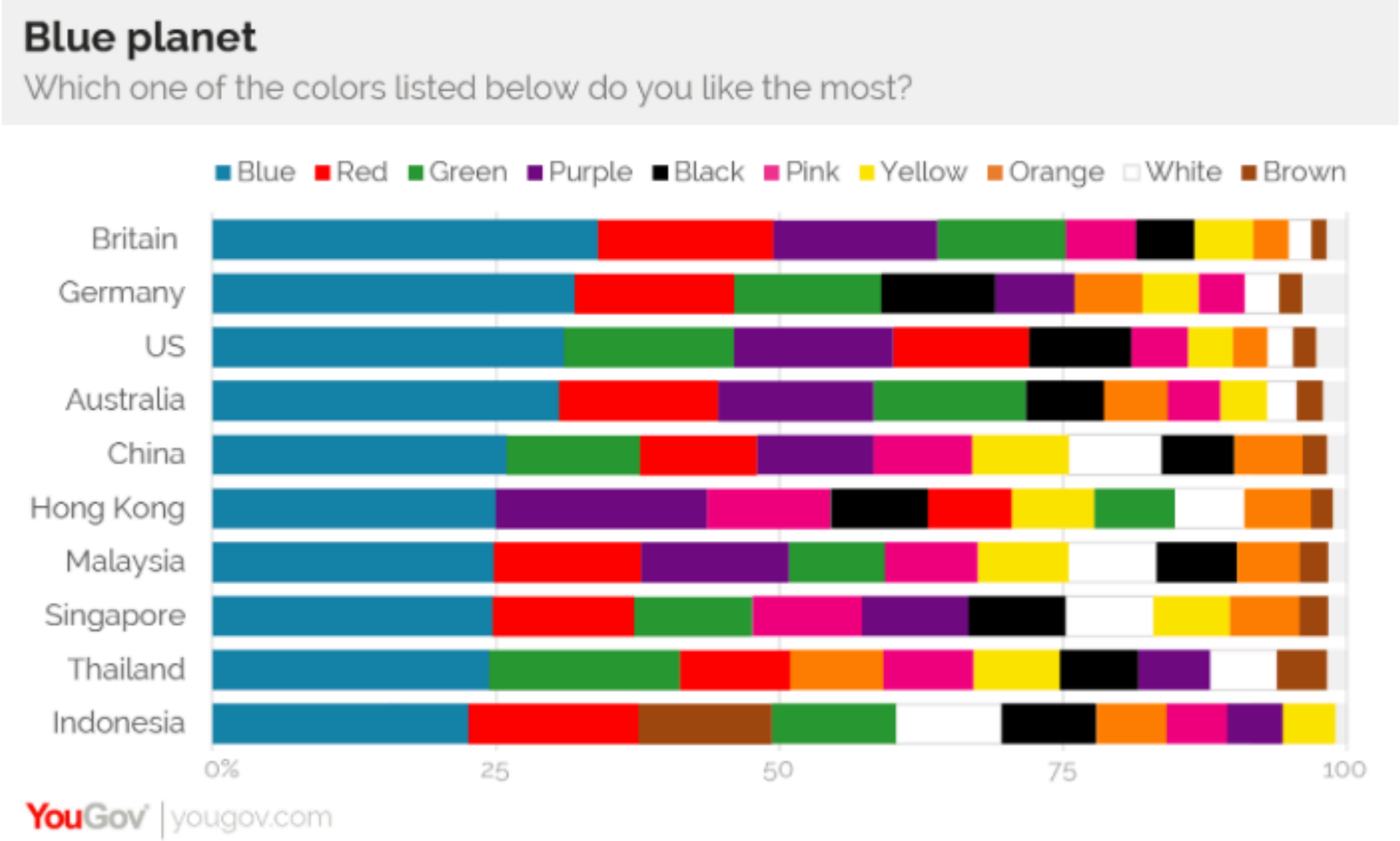


http://www.cnn.com/2017/07/17/politics/trump-approval-ratings/index.html

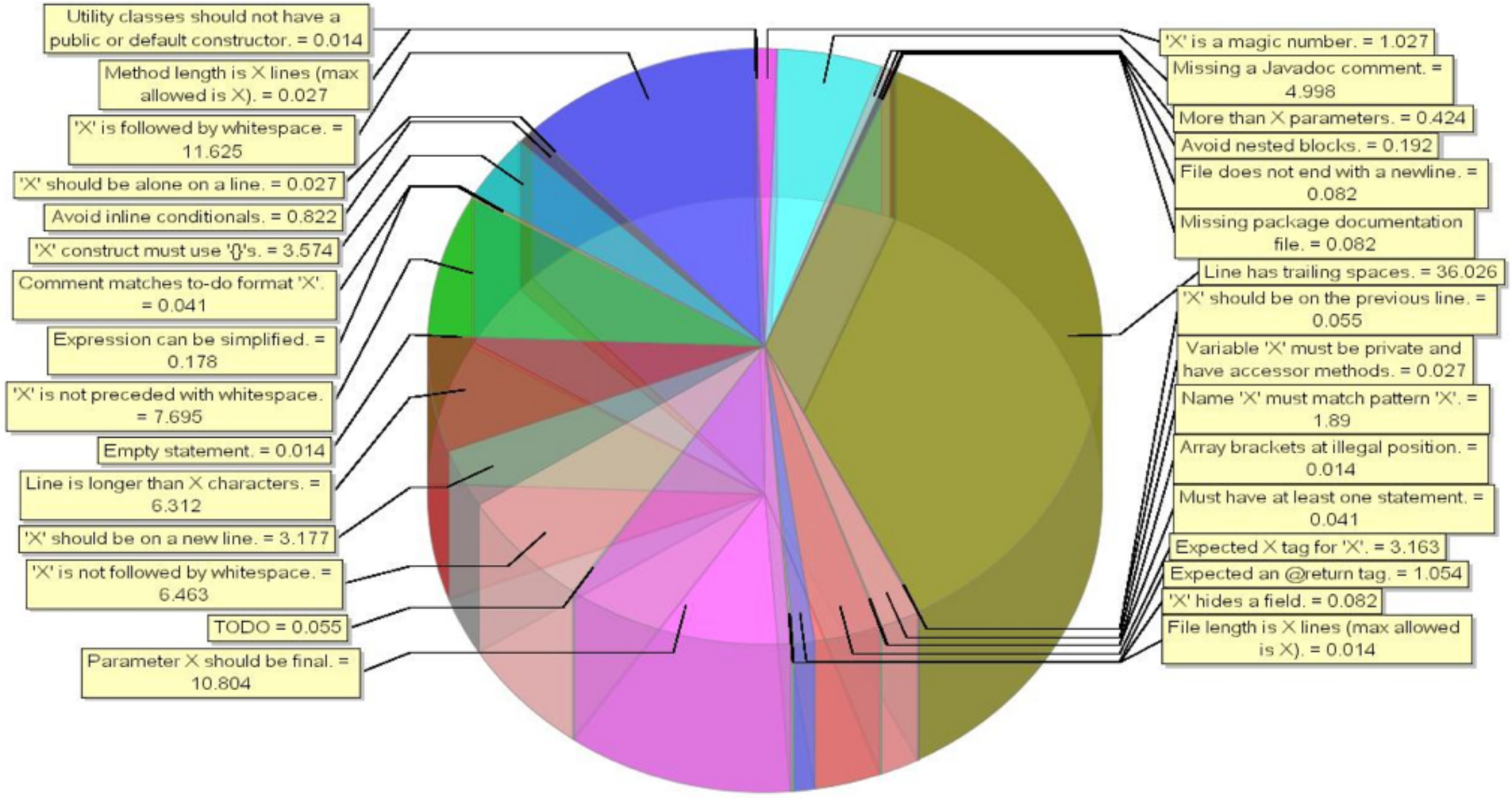
Graphical Integrity



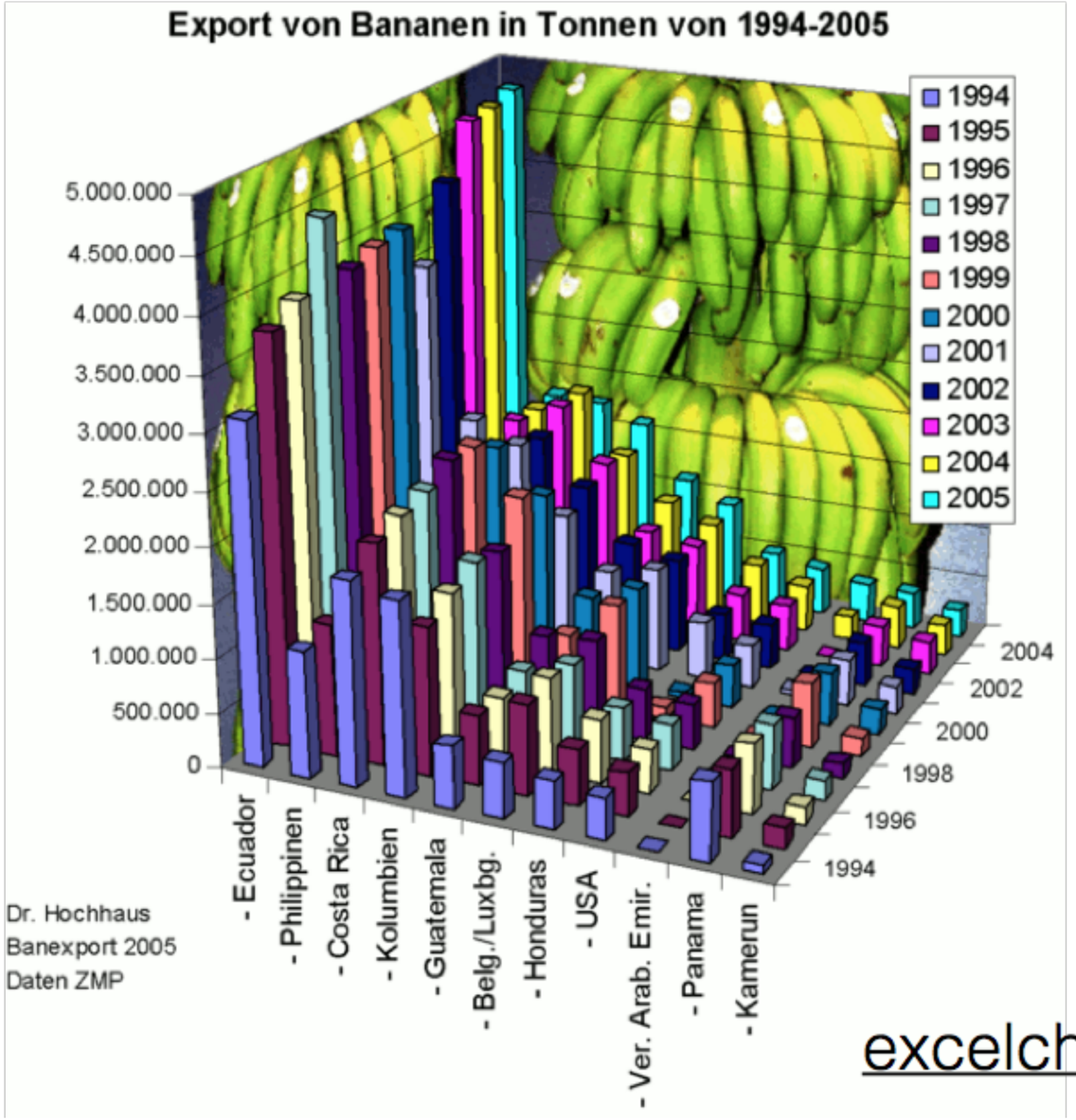
Graphical Integrity



Simple

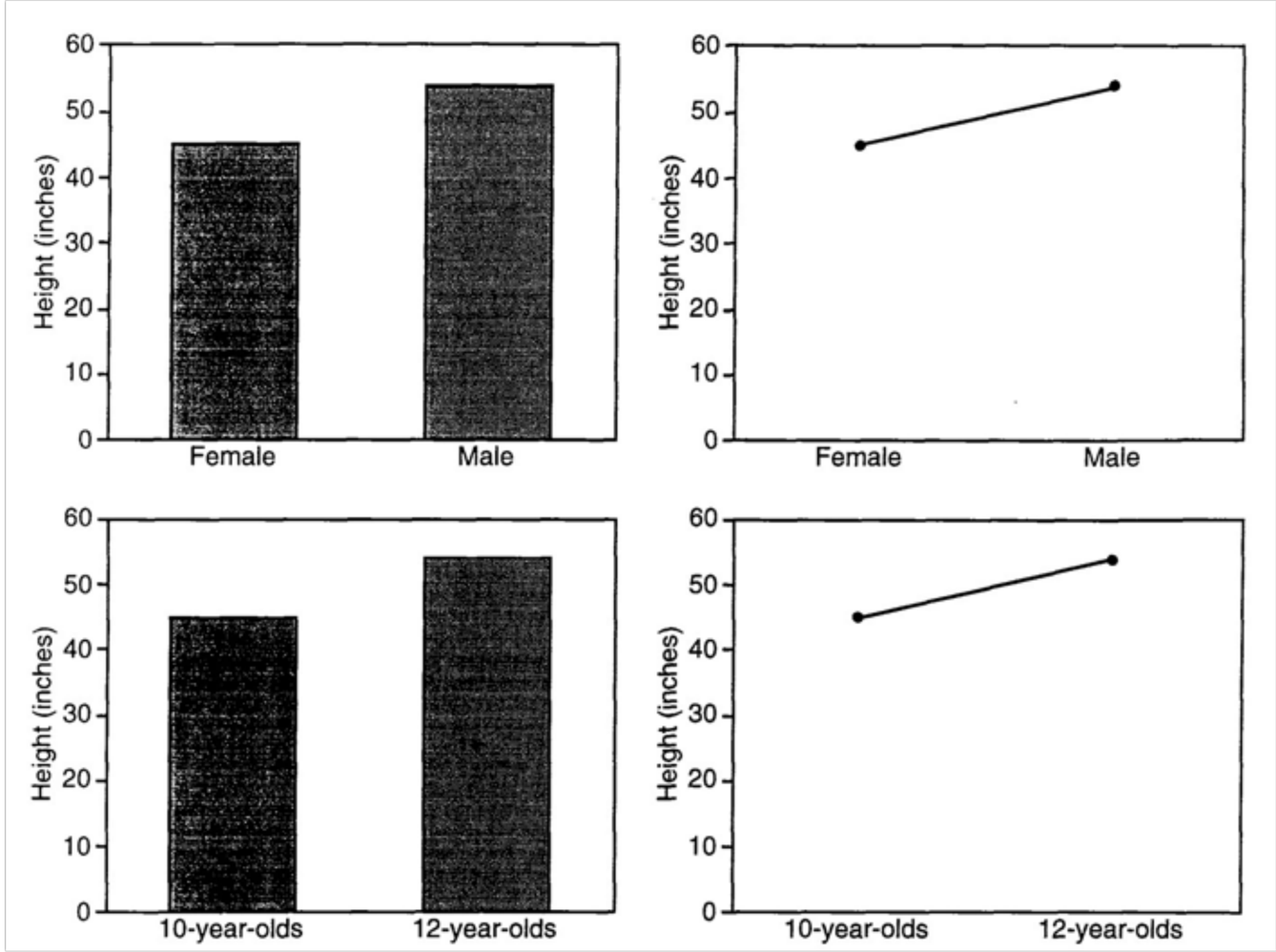

Utility classes should not have a public or default constructor. = 0.014
Method length is X lines (max allowed is X). = 0.027
'X' is followed by whitespace. = 11.625
'X' should be alone on a line. = 0.027
Avoid inline conditionals. = 0.822
'X' construct must use '{}'s. = 3.574
Comment matches to-do format 'X'. = 0.041
Expression can be simplified. = 0.178
'X' is not preceded with whitespace. = 7.695
Empty statement. = 0.014
Line is longer than X characters. = 6.312
'X' should be on a new line. = 3.177
'X' is not followed by whitespace. = 6.463
TODO = 0.055
Parameter X should be final. = 10.804
'X' is a magic number. = 1.027
Missing a Javadoc comment. = 4.998
More than X parameters. = 0.424
Avoid nested blocks. = 0.192
File does not end with a newline. = 0.082
Missing package documentation file. = 0.082
Line has trailing spaces. = 36.026
'X' should be on the previous line. = 0.055
Variable 'X' must be private and have accessor methods. = 0.027
Name 'X' must match pattern 'X'. = 1.89
Array brackets at illegal position. = 0.014
Must have at least one statement. = 0.041
Expected X tag for 'X'. = 3.163
Expected an @return tag. = 1.054
'X' hides a field. = 0.082
File length is X lines (max allowed is X). = 0.014

Simple



excelcharts.com

Proper Display

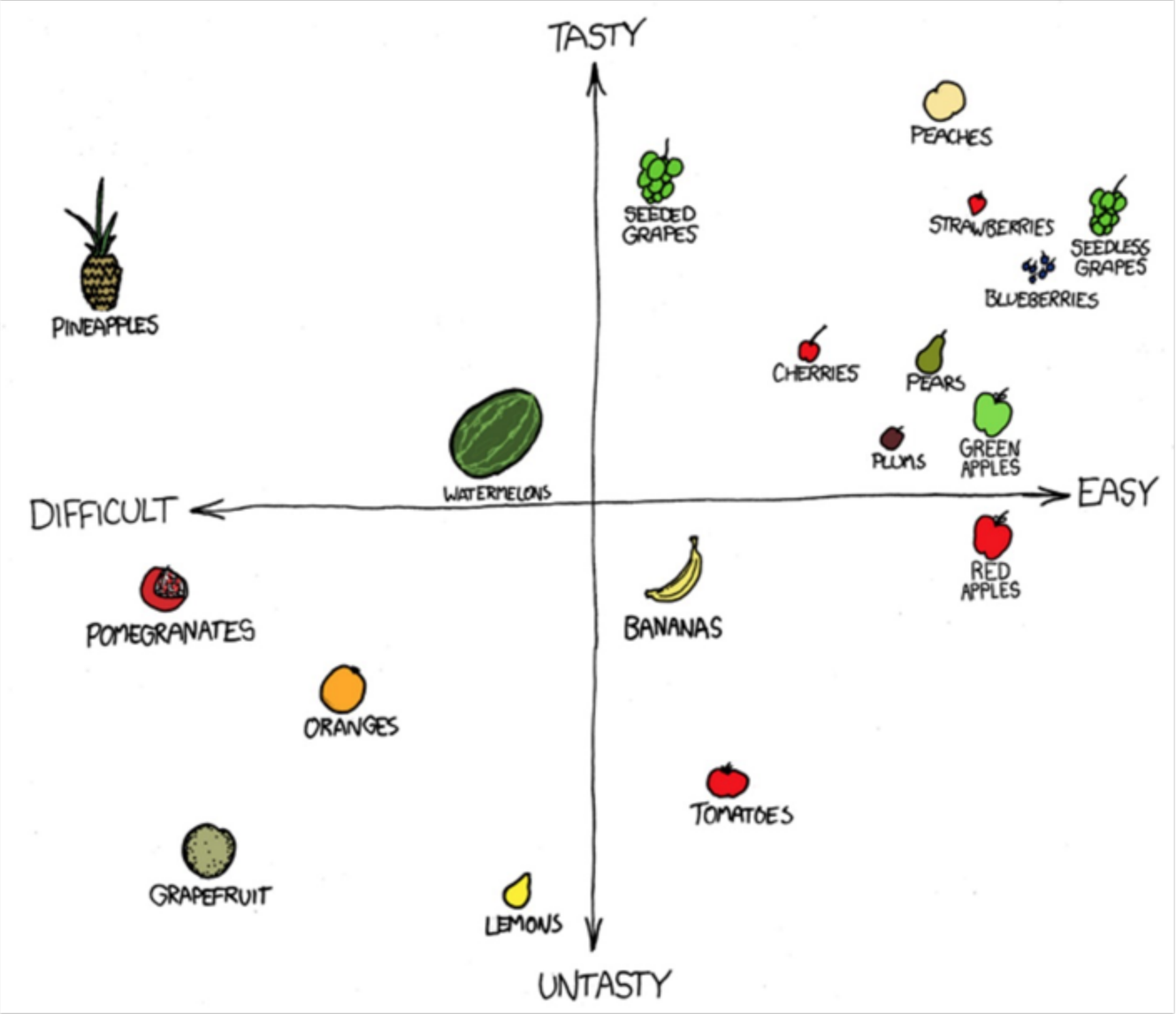


http://lenagroeger.s3.amazonaws.com/papers/BarsLines.pdf

Proper Display

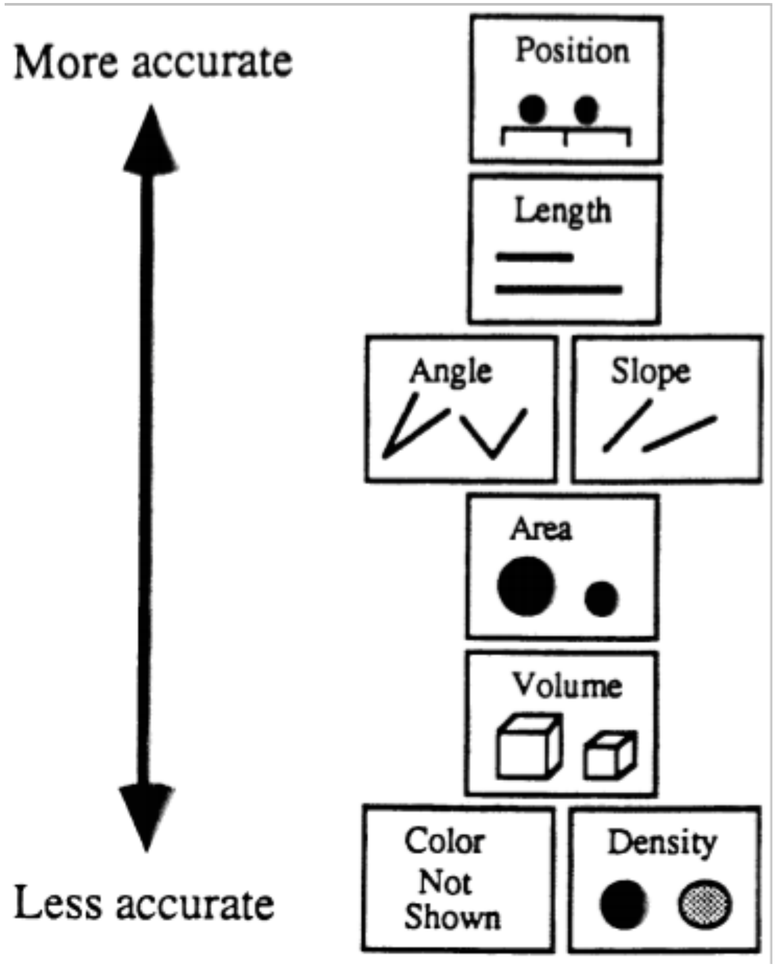
Google Finance

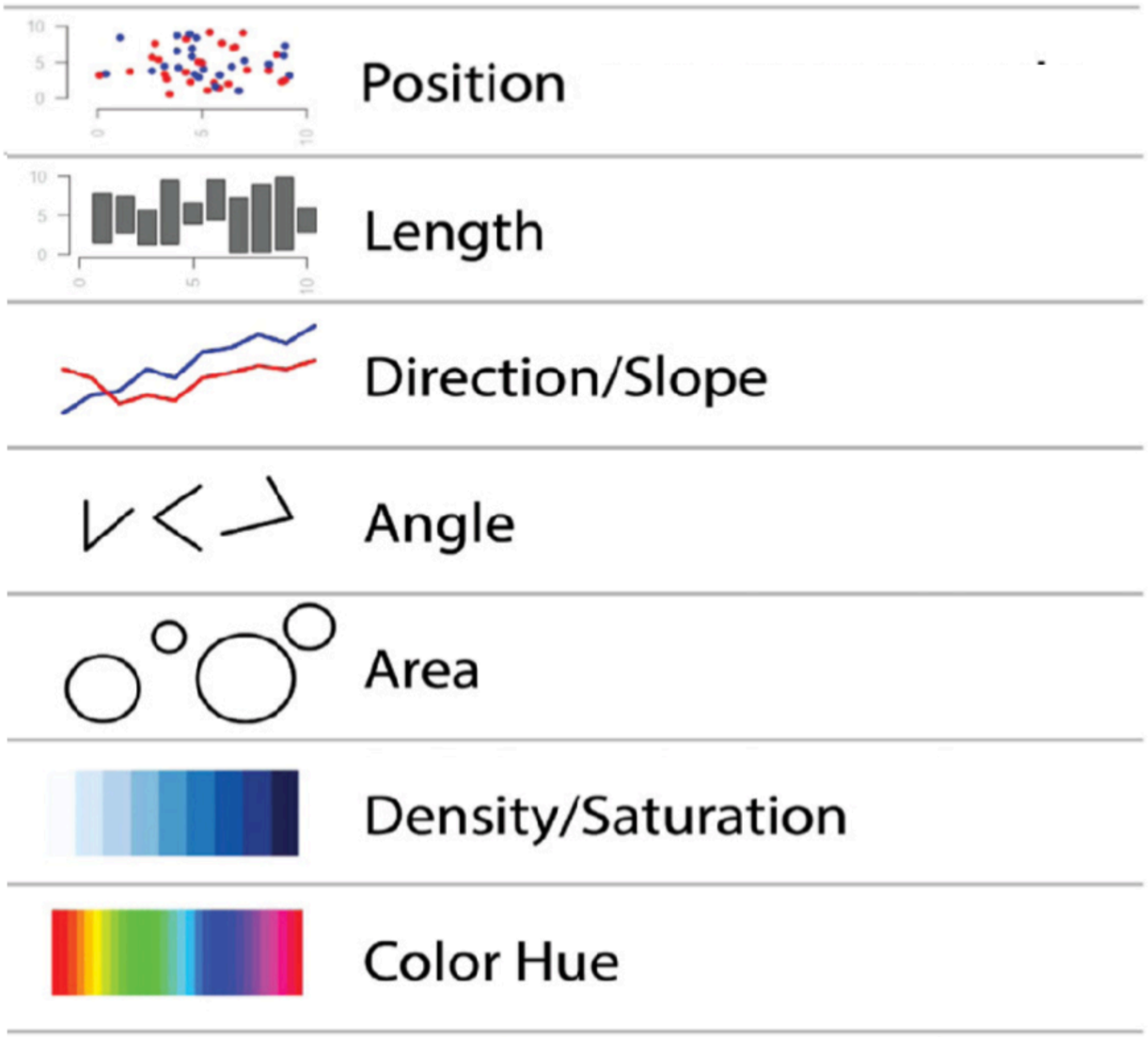
Proper Display



xkcd.com/388

Visual Encoding


More accurate
Position
Length
Angle
Slope
Area
Volume
Color Not Shown
Density
Less accurate


Position
Length
Direction/Slope
Angle
Area
Density/Saturation
Color Hue

Gestalt Laws

Law of Similarity:

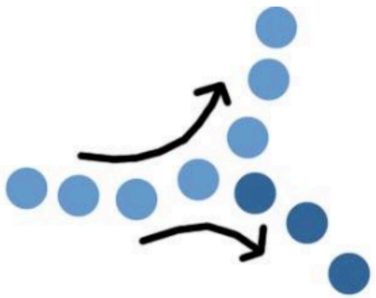
Items that are similar tend to be grouped together.

In the image above, most people see vertical columns of circles and squares.

Law of Pragnanz:

Reality is organized or reduced to the simplest form possible.

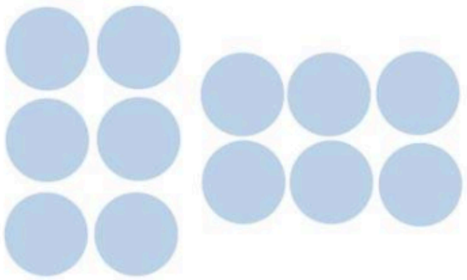
For example, we see the image above as a series of circles rather than as many much more complicated shapes.



Law of Continuity:

Lines are seen as following the smoothest path.

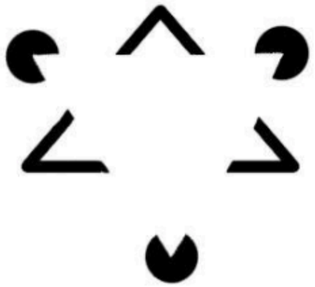
In the image above, the top branch is seen as continuing the first segment of the line. This allows us to see things as flowing smoothly without breaking lines up into multiple parts.



Law of Proximity:

Objects near each other tend to be grouped together.

The circles on the left appear to be grouped in vertical columns, while those on the right appear to be grouped in horizontal rows.



Law of Closure:

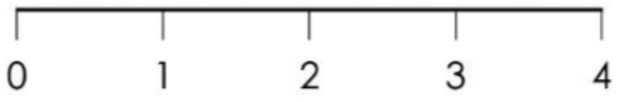
Objects grouped together are seen as a whole.

We tend to ignore gaps and complete contour lines. In the image above, there are no triangles or circles, but our minds fill in the missing information to create familiar shapes and images.

http://psychology.about.com/od/sensationandperception/ss/gestaltlaws.htm

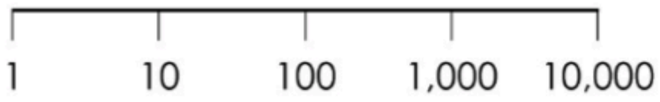
Linear

Values are evenly spaced



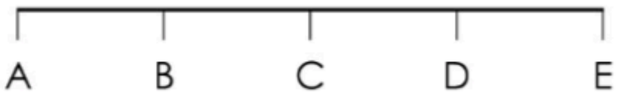
Logarithmic

Focus on percent change



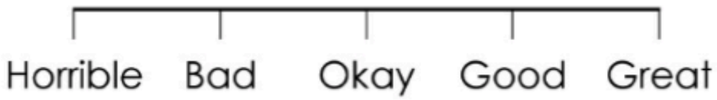
Categorical

Discrete placement in bins



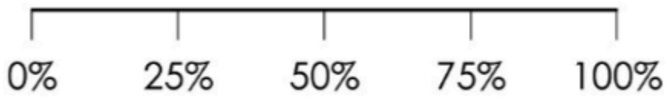
Ordinal

Categories where order matters



Percent

Representing parts of a whole



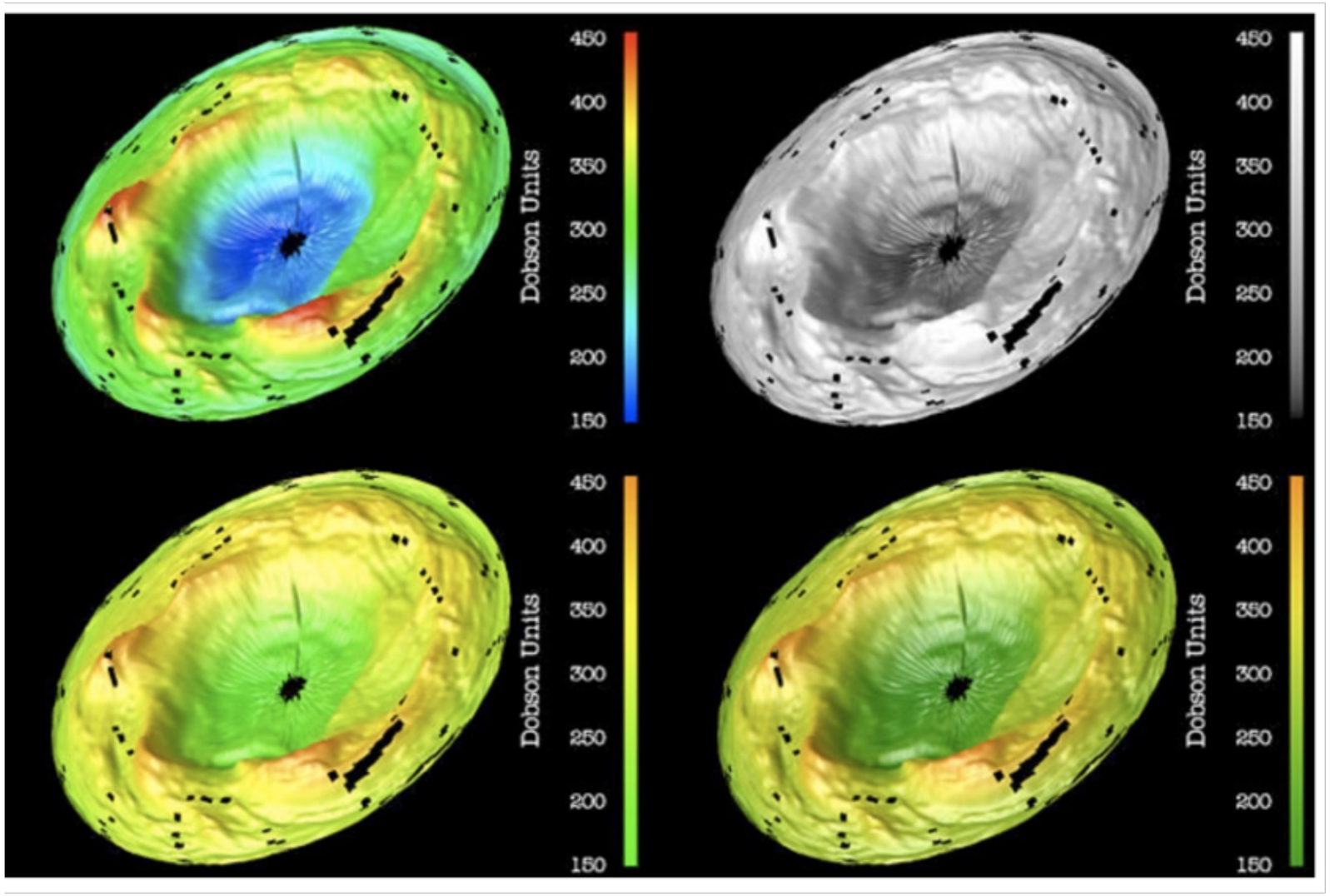
Time

Units of months, days, or hours

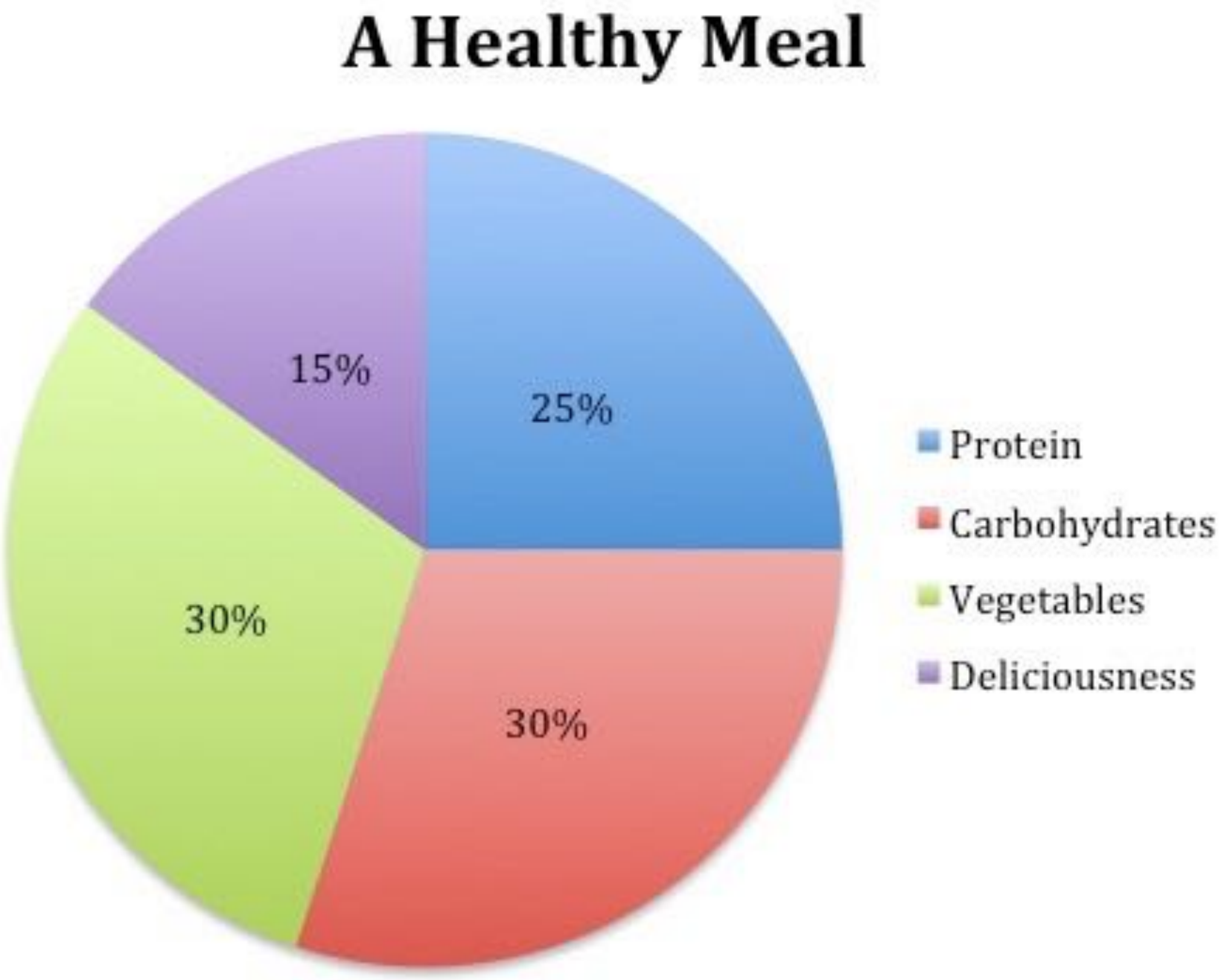
Source: Nathan Yau, Data Points

Color

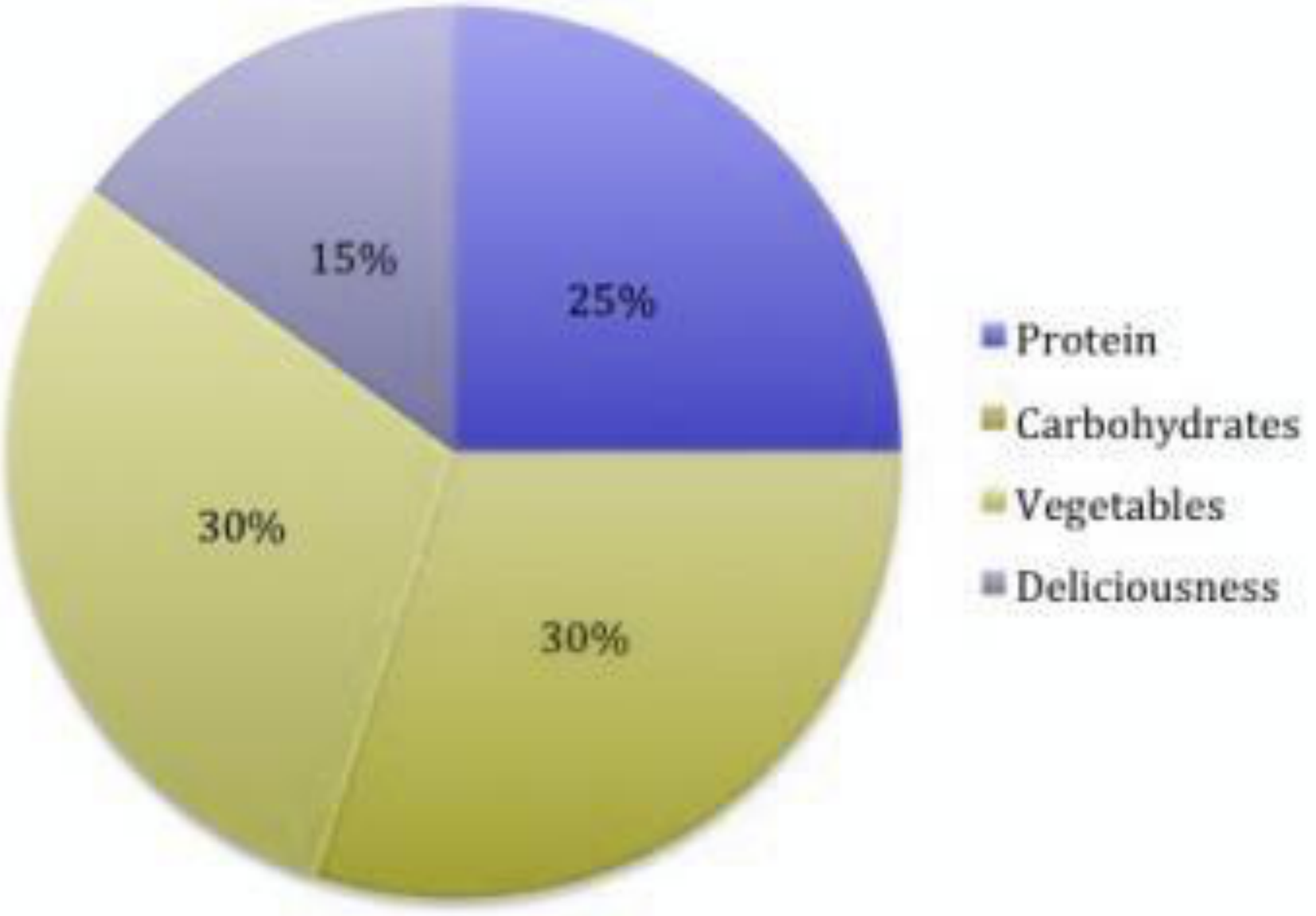
Color



Rogowitz and Treinish


A Healthy Meal
15%
25%
30%
30%
Protein
Carbohydrates
Vegetables
Deliciousness

A Healthy Meal



elastica
Dashboard
Users
Total (192)
28
High Risk
0
Med Risk
4
Blocked
Policies
Total (231)
3
Blocking
142
Alerting
Policy Alerts
Alerting
443
Blocked
13
Rest
0
Threat Alerts
High Risk
2
Med Risk
2
Low Risk
17
Audited Services
by Users
High Risk (736)
Medium Risk (3k)
Low Risk (3k)
3k
494.3 GB
963k
243

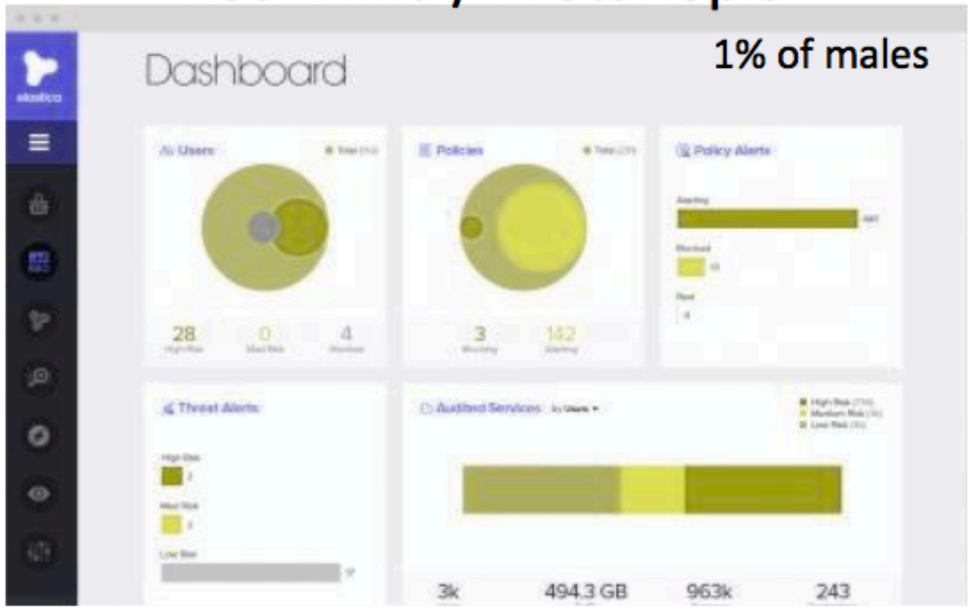
Original

Green-Blind / Deuteranopia

8% of male
0.5% female

Red-Blind / Protanopia

1% of males

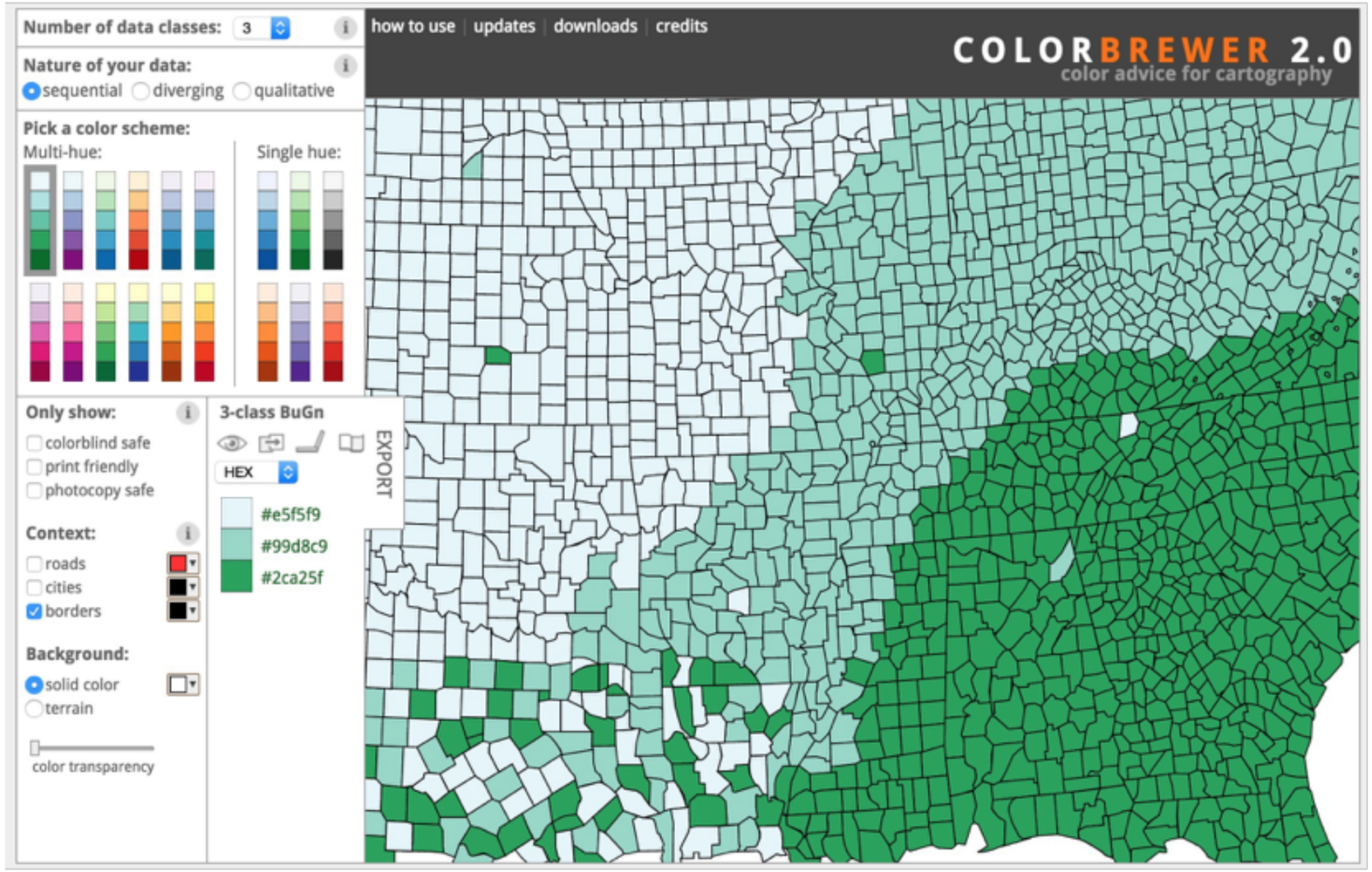


Blue-Blind / Tritanopia

rare

http://www.color-blindness.com/coblis-color-blindness-simulator/

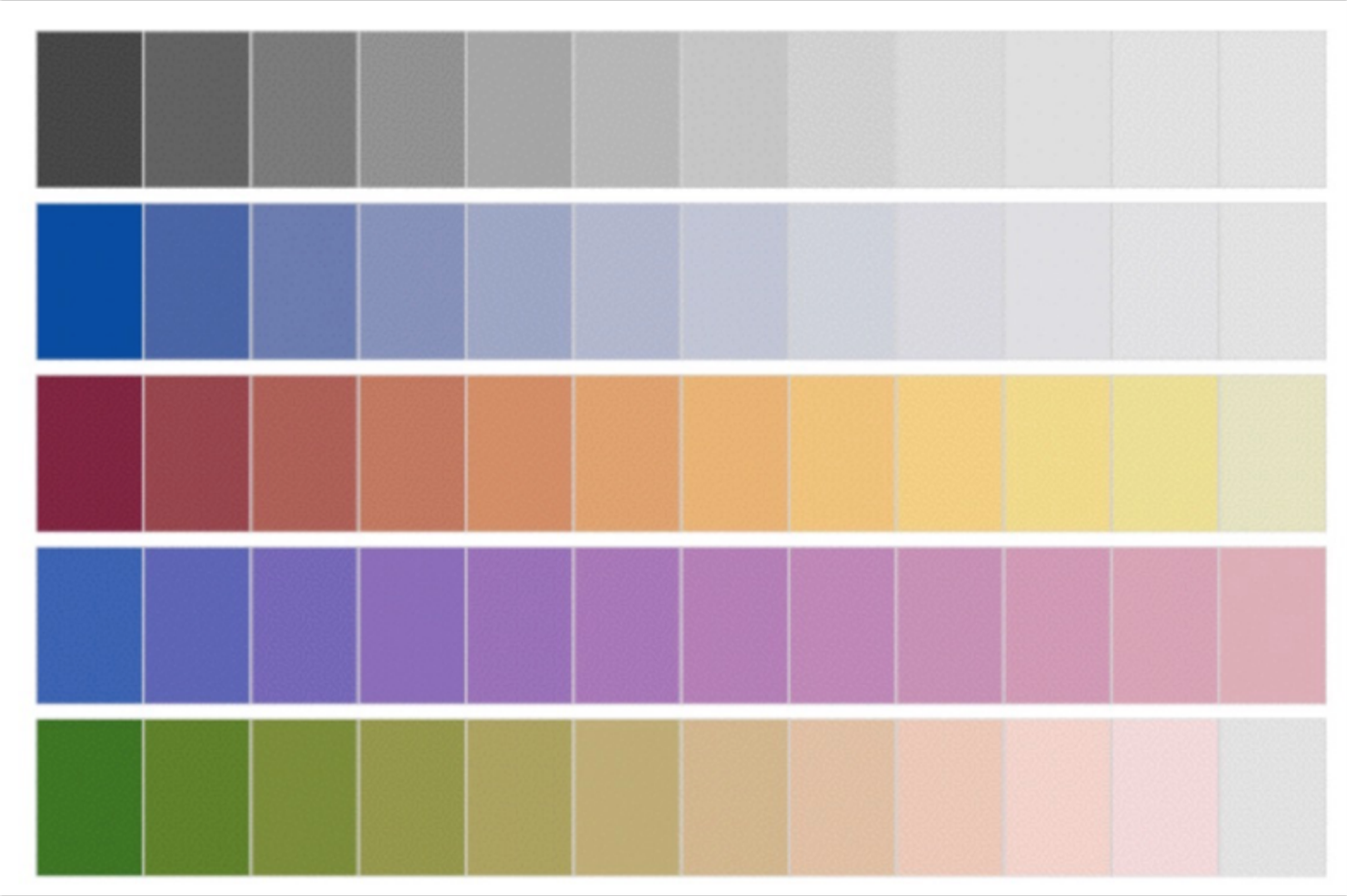
Color



gtkcyber.com

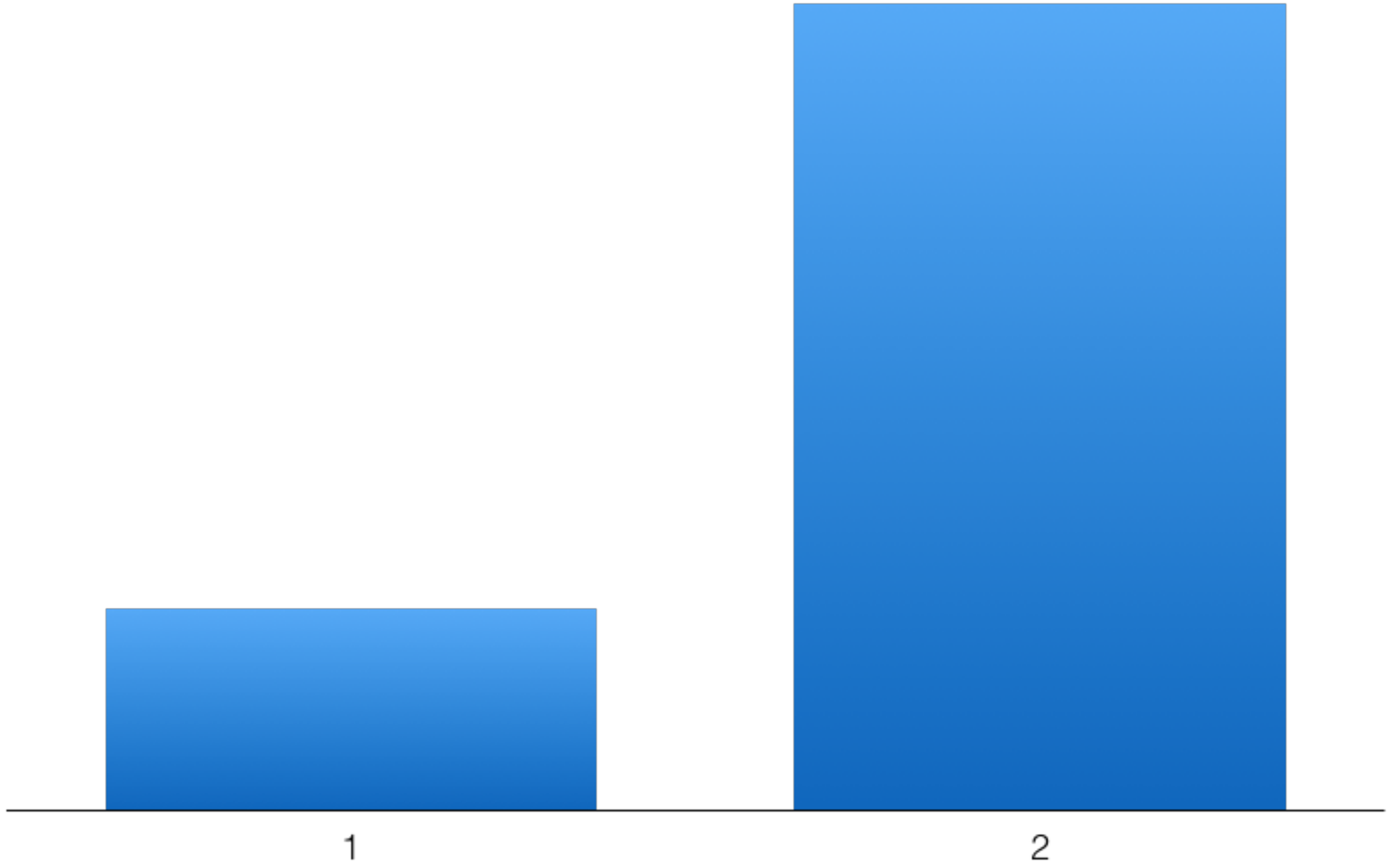
colorbrewer2.org

Color



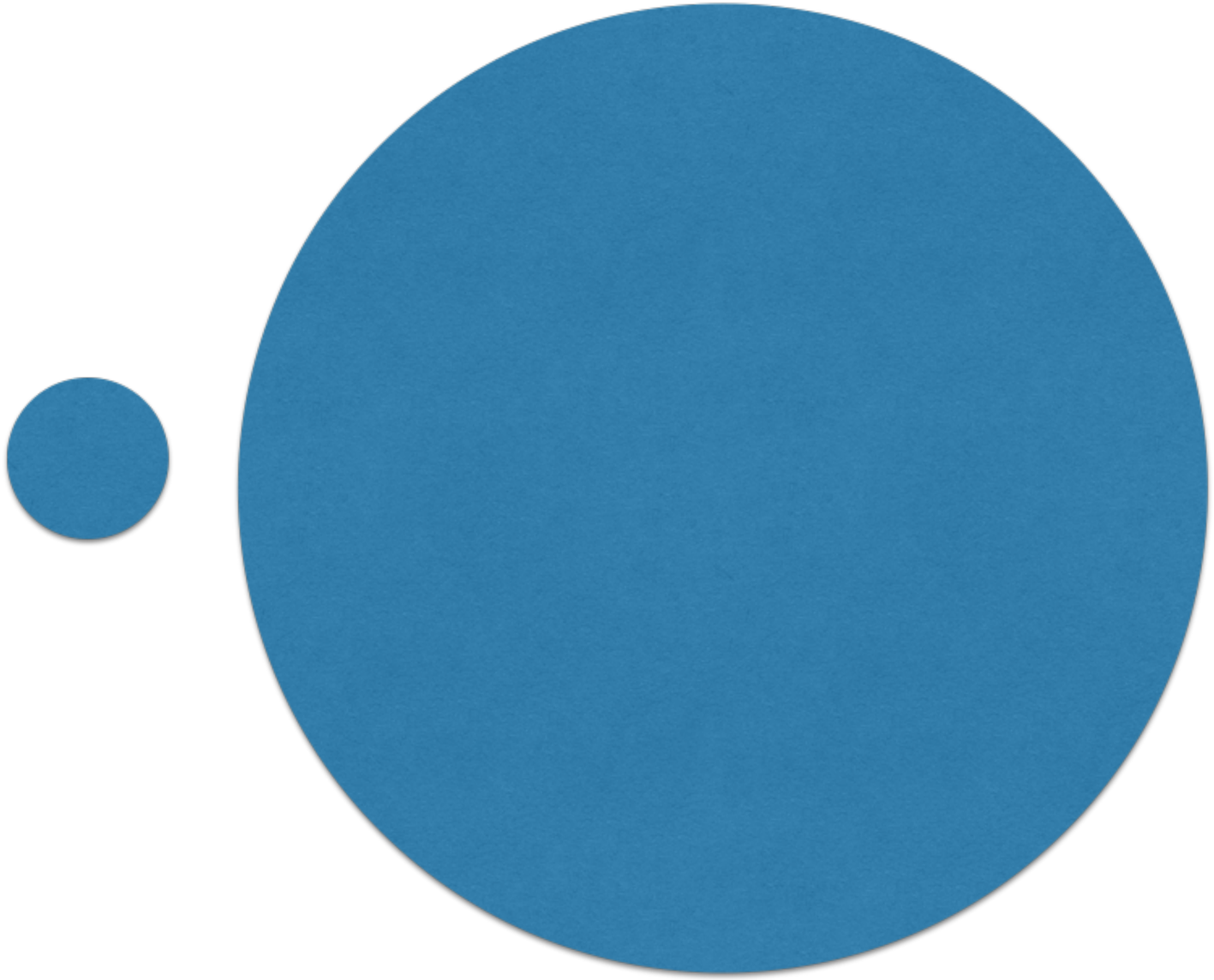
colorbrewer2.org

Height Difference?



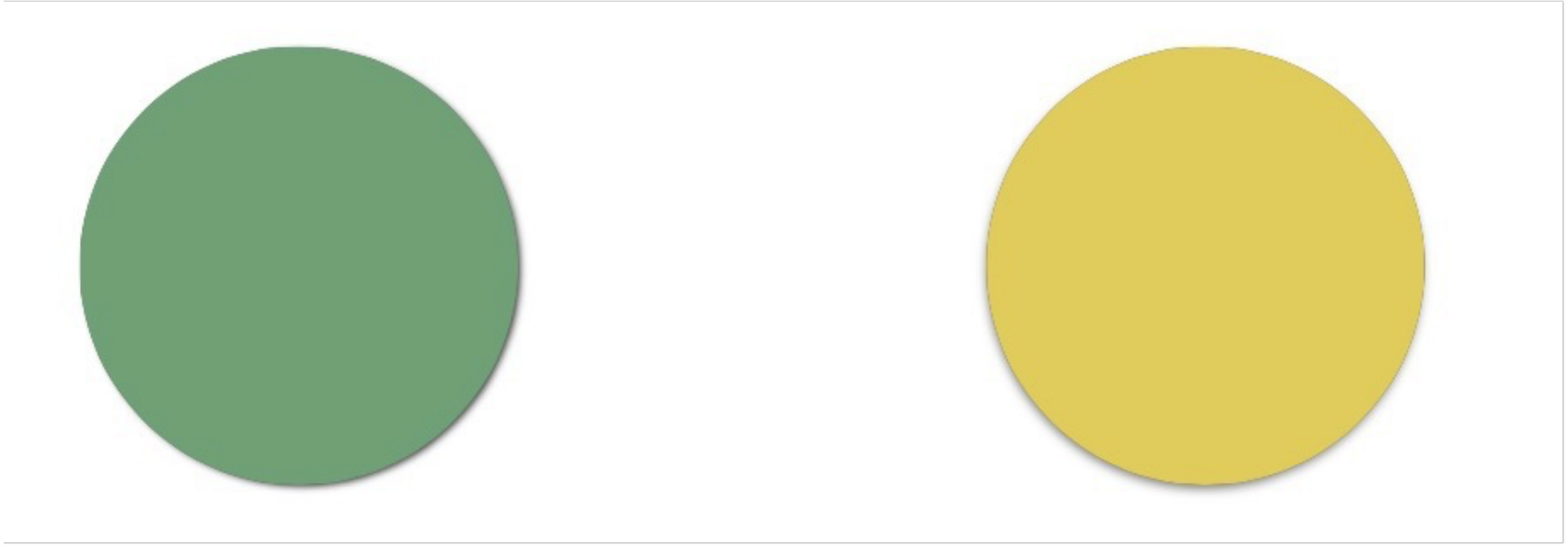
Slope Difference?

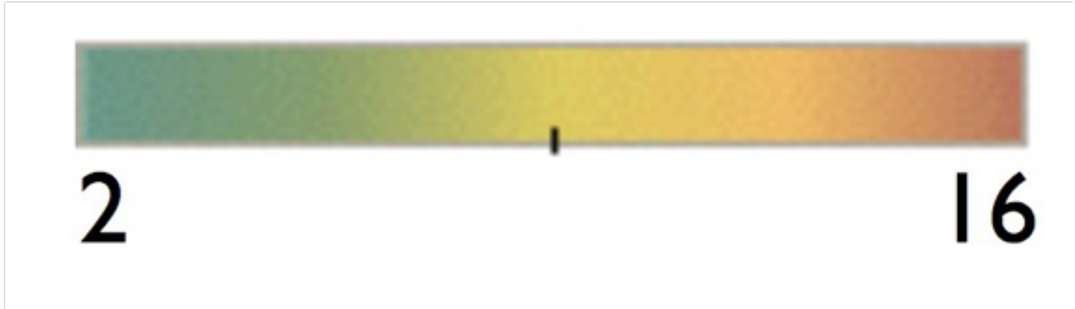
Area Difference?

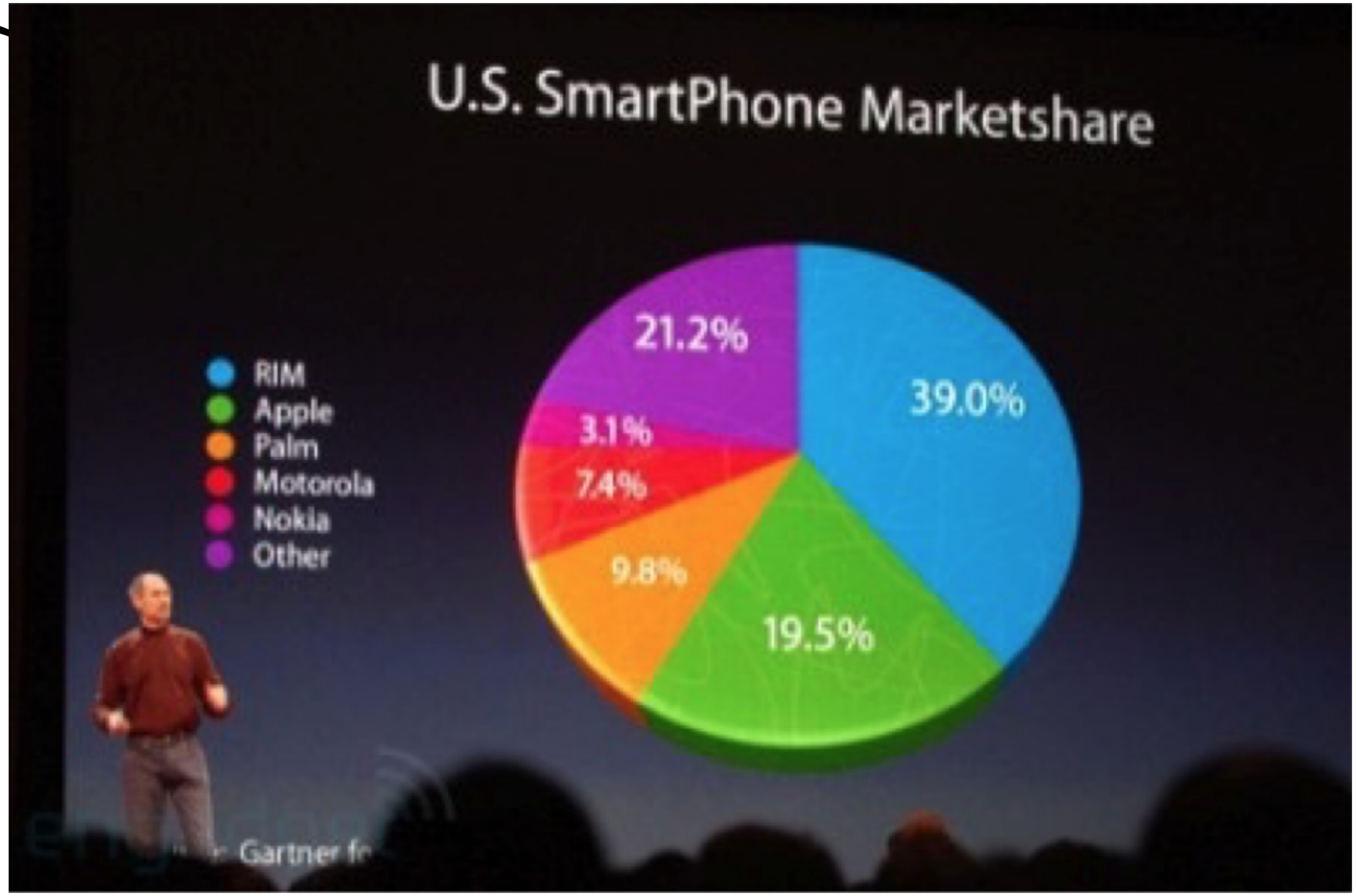


Saturation Difference?

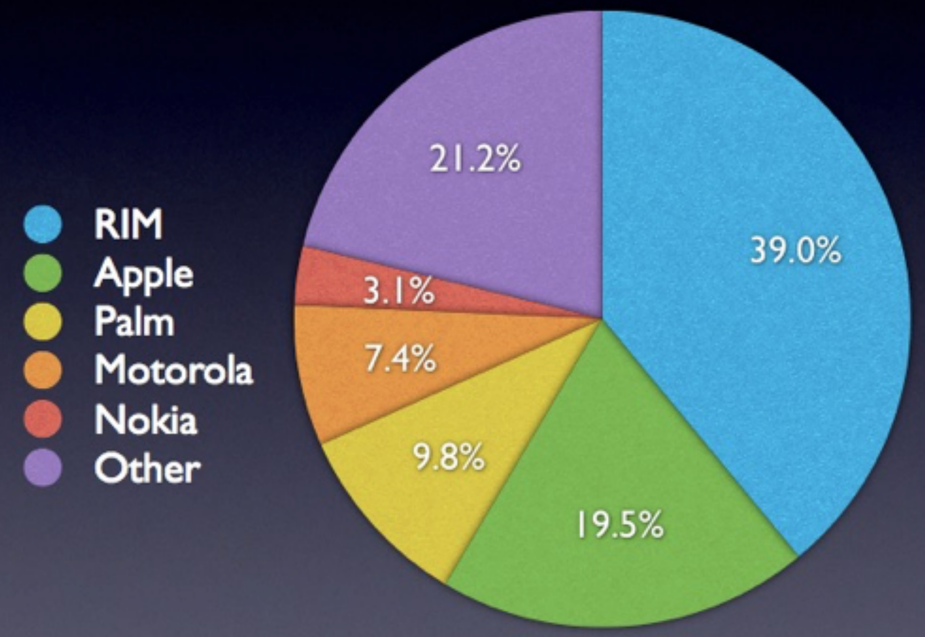
Mapping Difference?



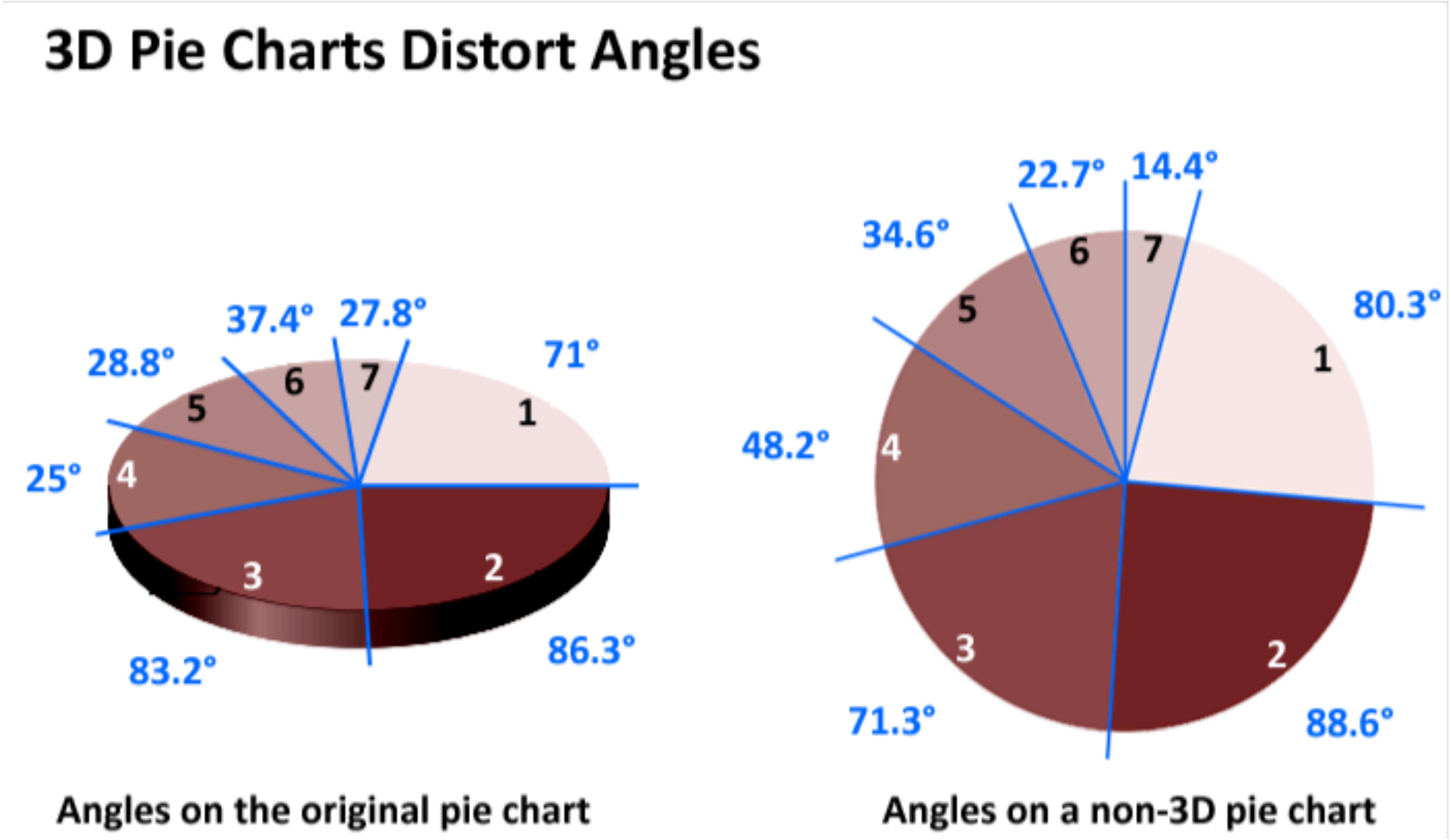



U.S. SmartPhone Marketshare
RIM
Apple
Palm
Motorola
Nokia
Other
39.0%
19.5%
9.8%
7.4%
3.1%
21.2%
Gartner fe

U.S. SmartPhone Marketshare

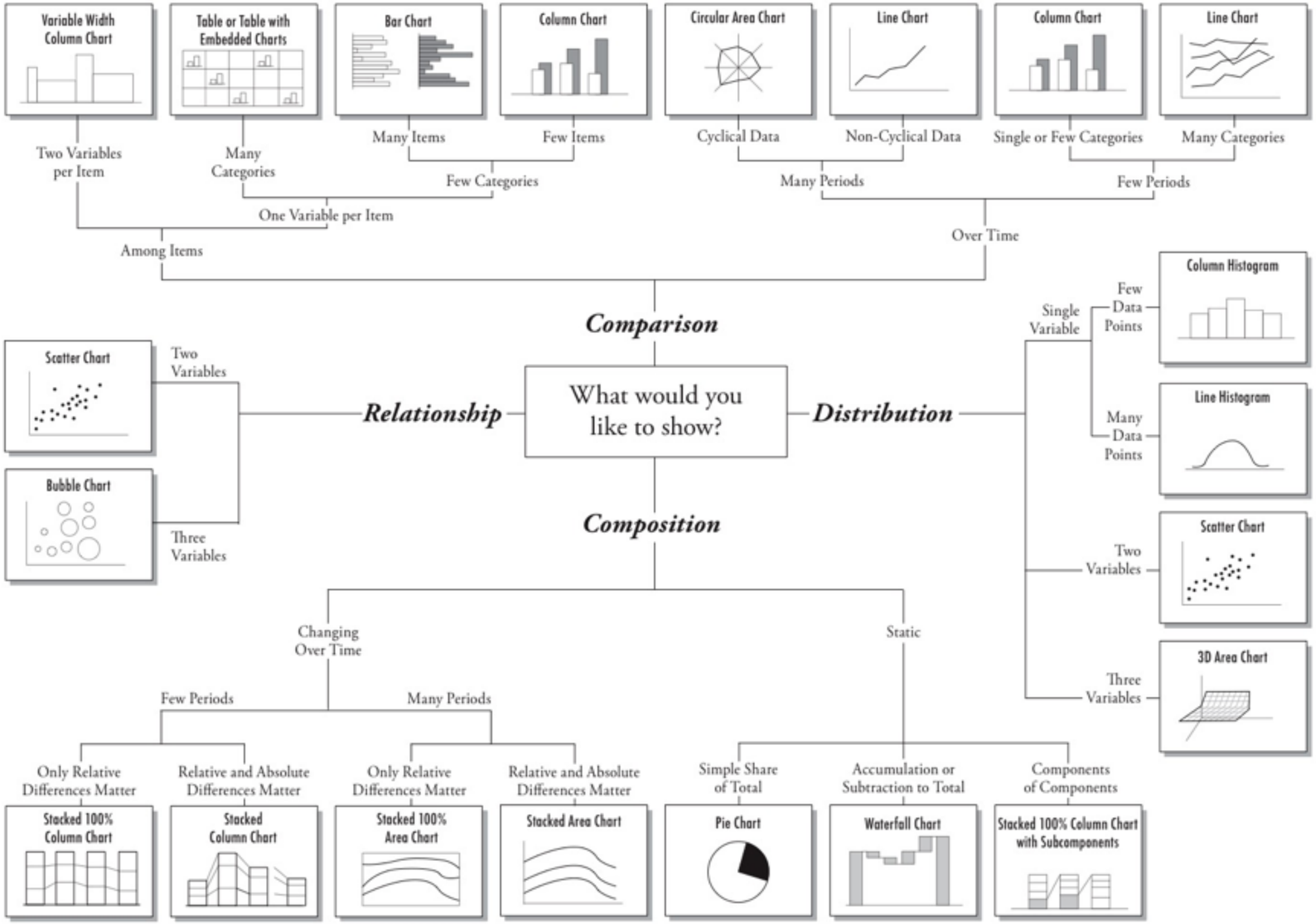


3D Pie Charts are BAD!

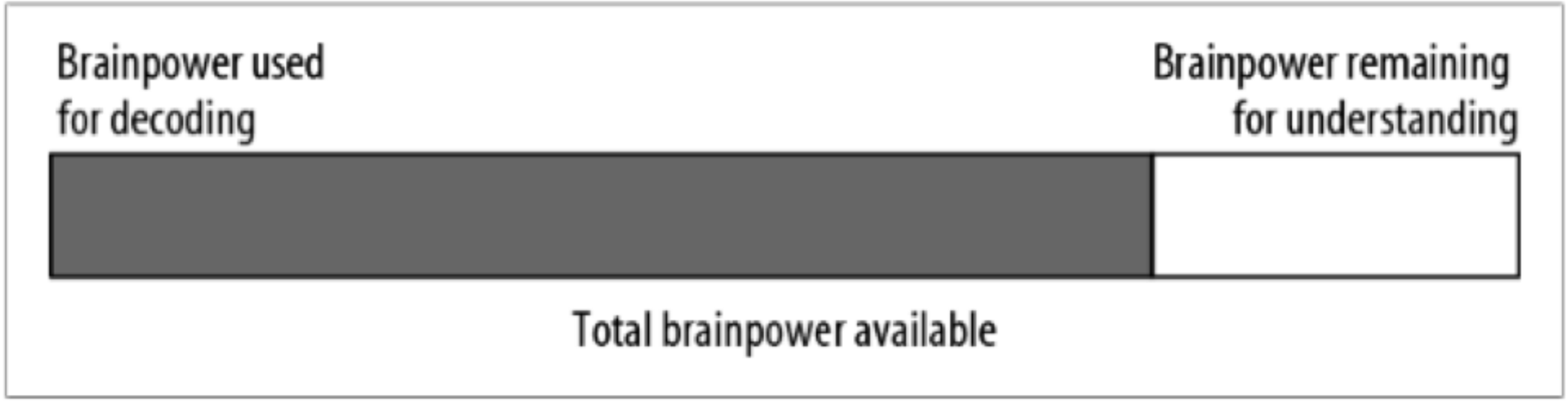


h8p://themendozaline.org/post/113828947026/3d-pie-charts-are-lie-charts

Chart Suggestions—A Thought-Starter

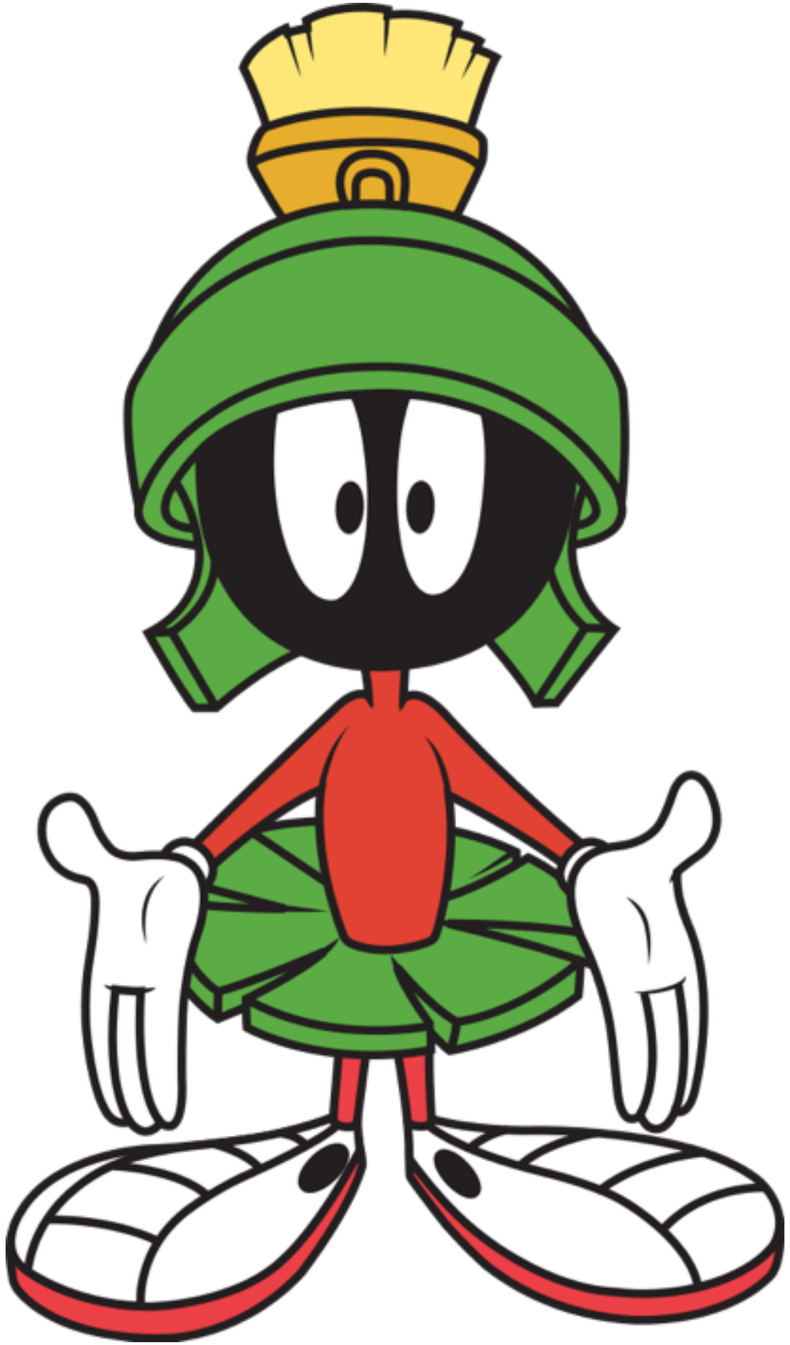


© 2006 A. Abela — a.v.abela@gmail.com

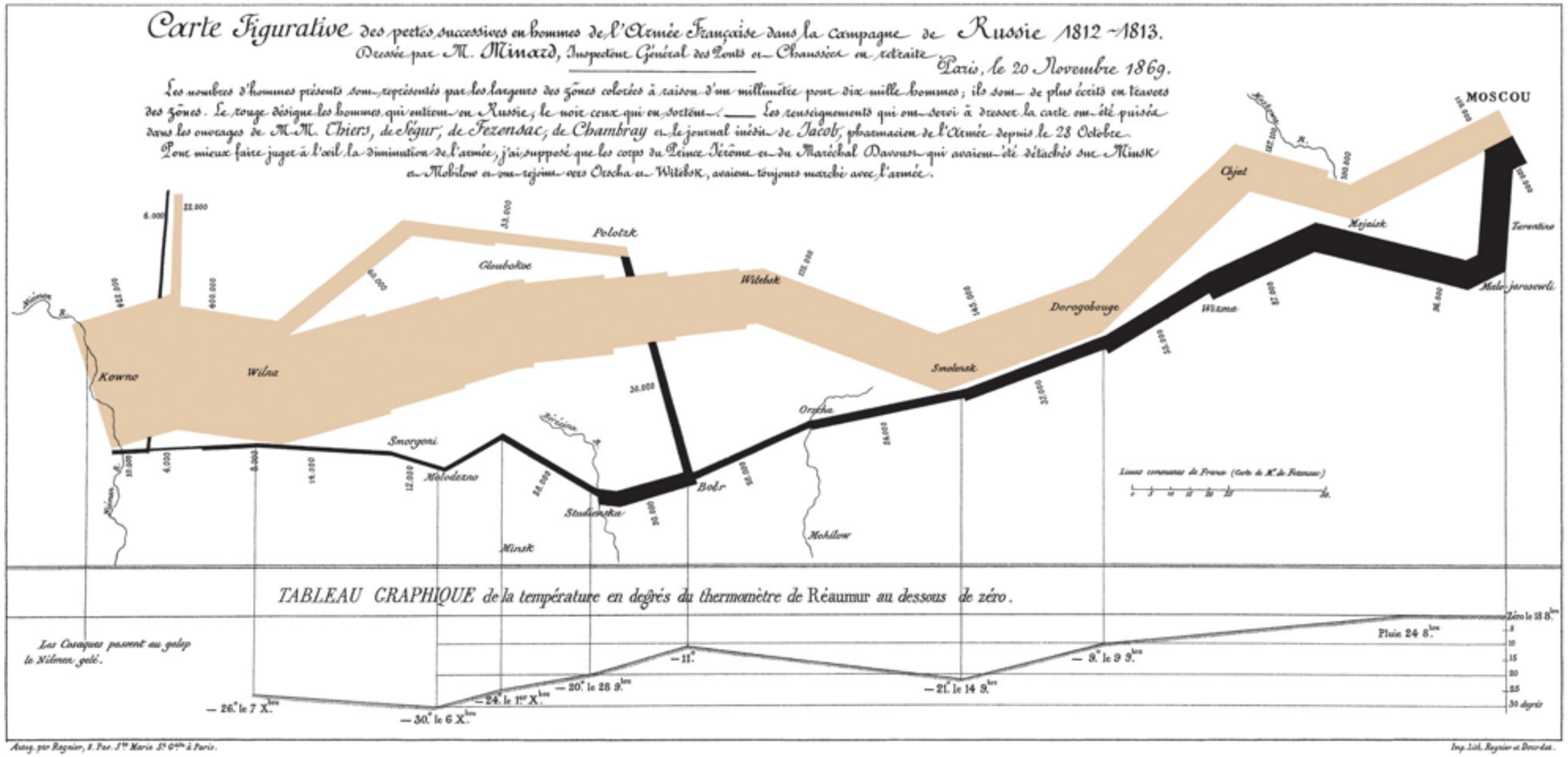

Brainpower used
for decoding
Brainpower remaining
for understanding
Total brainpower available

GTK CYBER

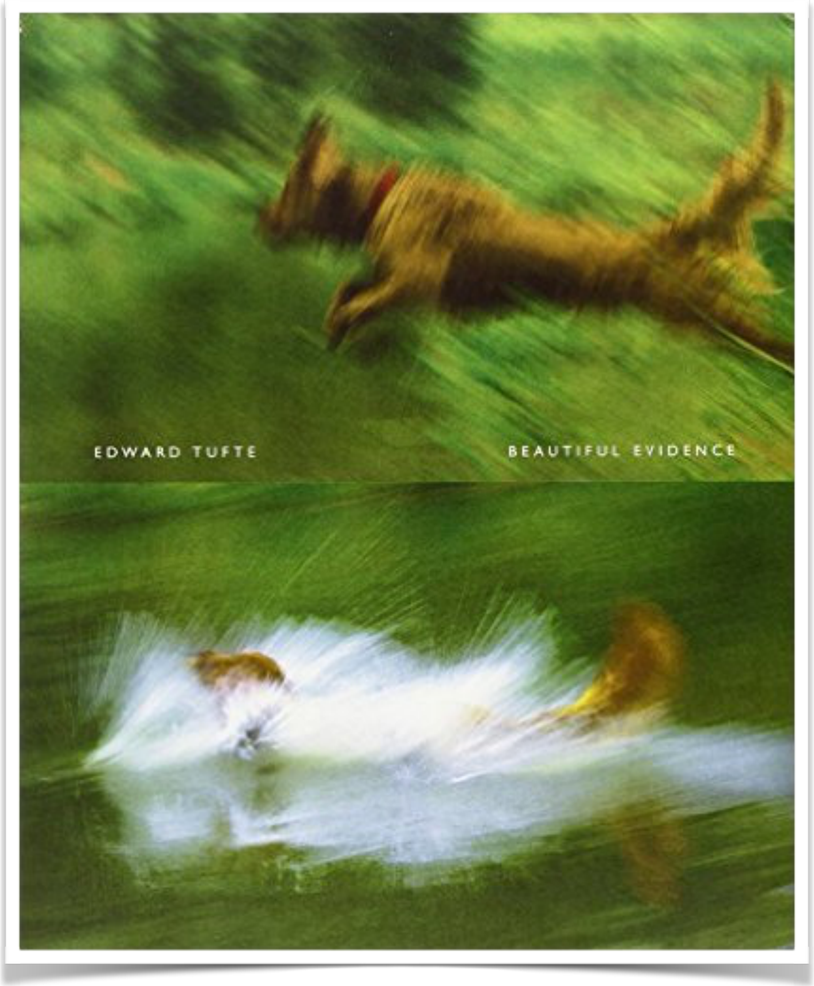
GET ON THE FAST TRACK



gtkcyber.com


Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et vers Orcha et Witebsk, avaient toujours marché avec l'armée.
MOSCOU
Kowno
Wilna
Gloubokoe
Polotzk
Witebsk
Smolensk
Dorogobouge
Wizma
Mojaisk
Tarentino
Malo-jarosewli
Ojat
Smorgoni
Molodezno
Studienska
Minsk
Bobr
Orscha
Mohilow
Niémen R.
Wilia R.
Dniepr R.
Moskwa R.
Lieues communes de France (Carte de M. de Fezensac)
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niémen gelé.
— 26.° le 7 X.bre
— 30.° le 6 X.bre
— 24.° le 1er X.bre
— 20.° le 28 9.bre
— 11.°
— 21.° le 14 9.bre
— 9.° le 9 9.bre
Pluie 24 8.bre
Zéro le 13 8.bre
30 degrés
Imp. Lith. Regnier et Dourdet.

Recommended Reading



Storytelling

Storytelling

What is the Question?

Who Cares?

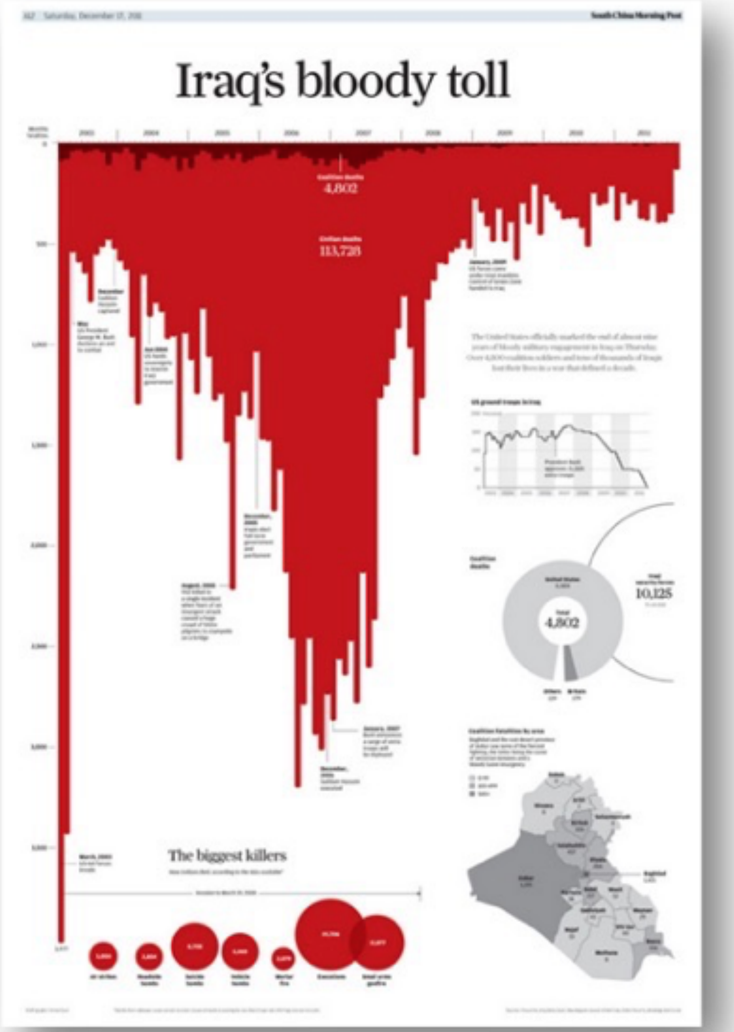
Storytelling

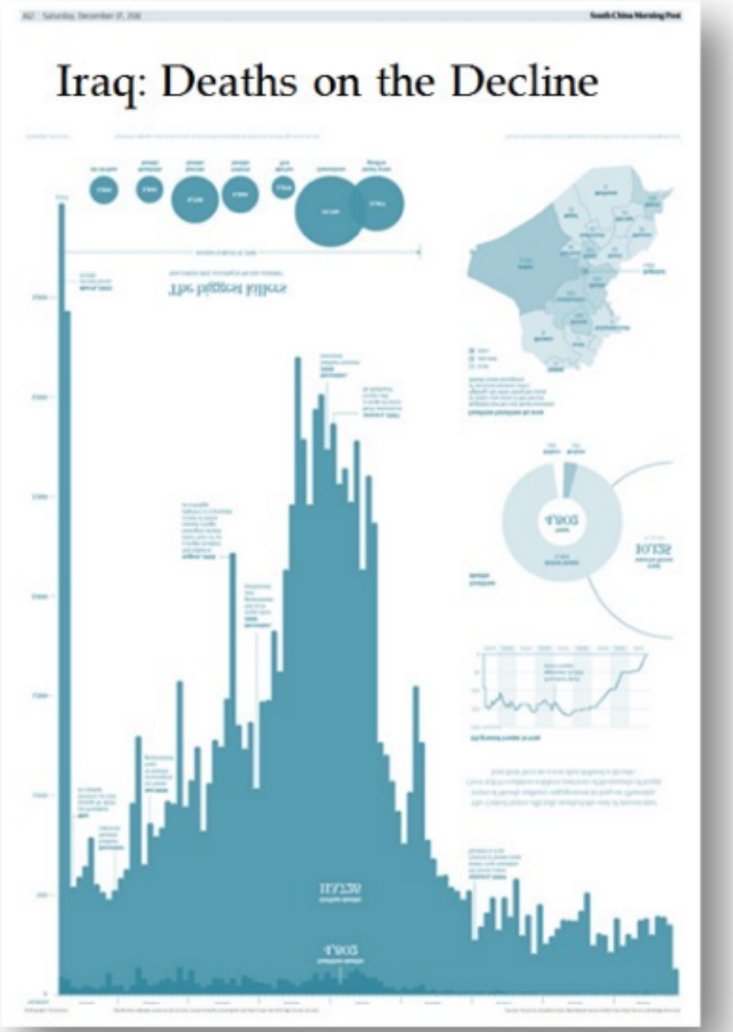
1. Know Your Audience
2. Tell a Story
3. Don’t Confuse
4. Go with the Flow

Elements of a Story

1. Characters
2. Conflict
3. Resolution
4. Sequel?

Influential Storytelling





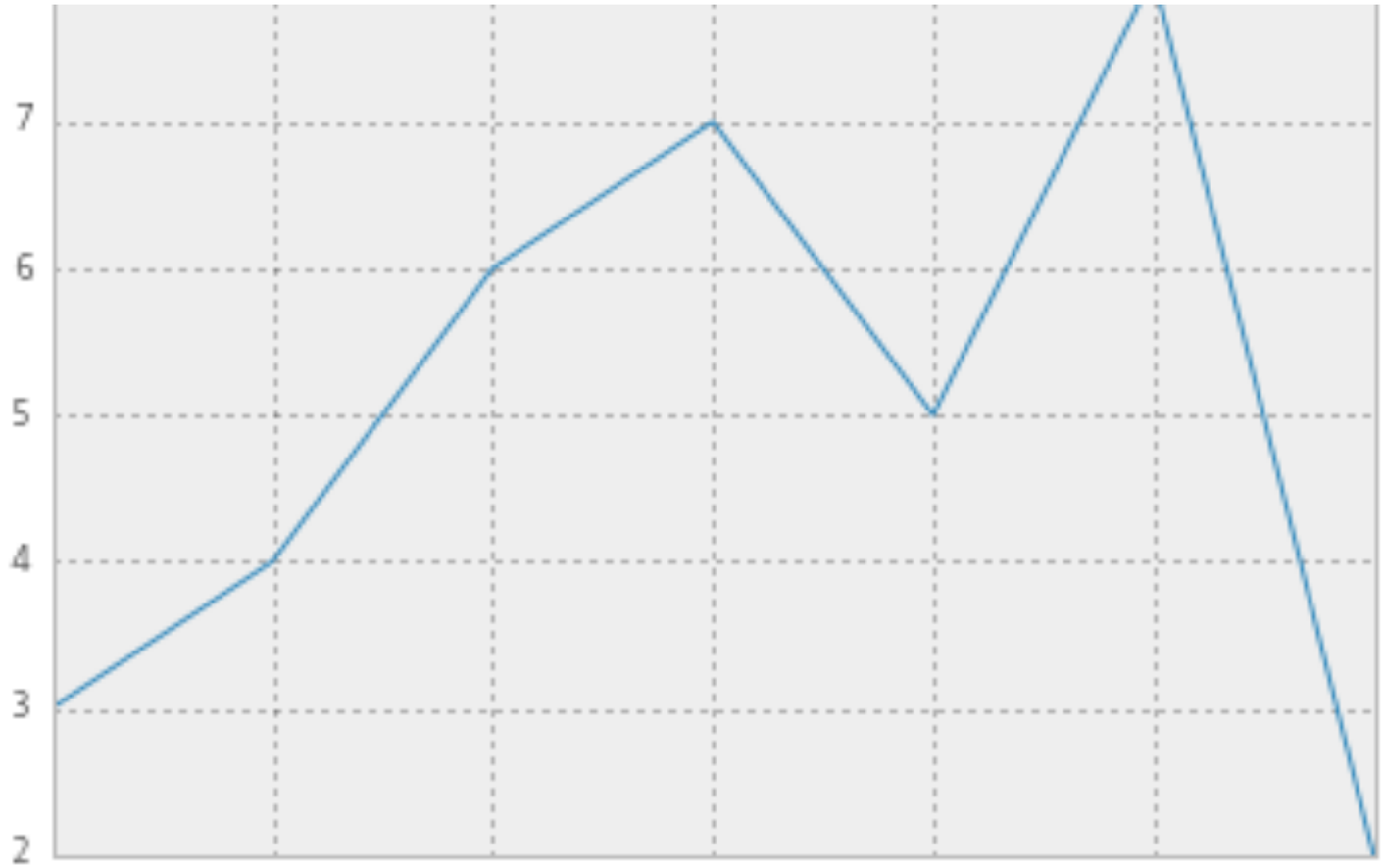
Andy Cotgreave

Data Visualization in Python

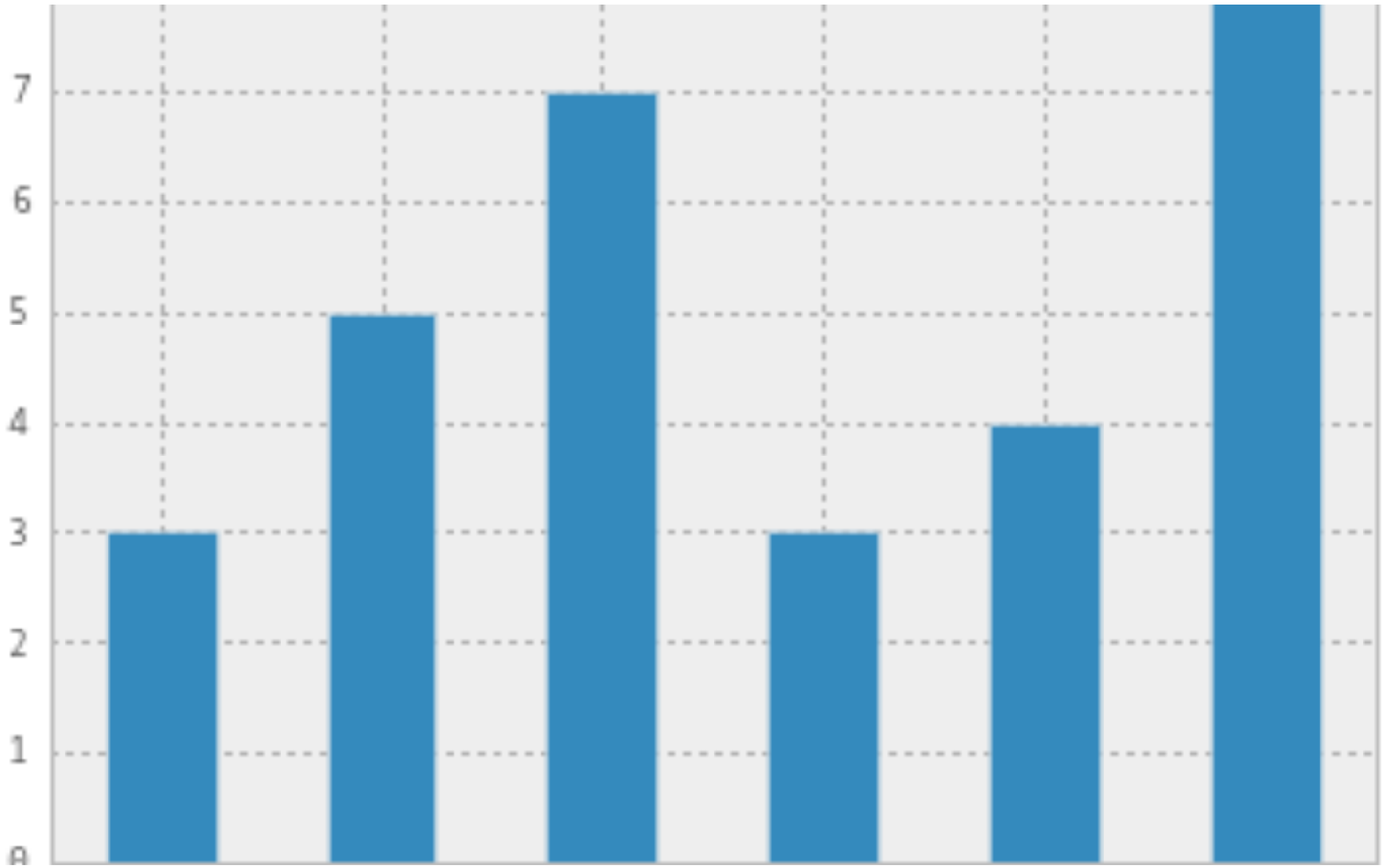
Python Visualization Libraries

- **Matplotlib**: “Lowest” level visualization library. Very powerful, but little abstraction
- **Pandas/Pyplot**: Easier, but requires transformation for complex visualizations
- **Seaborn:** Good for advanced statistical visualizations
- **Altair:** New library: declarative statistical visualization library

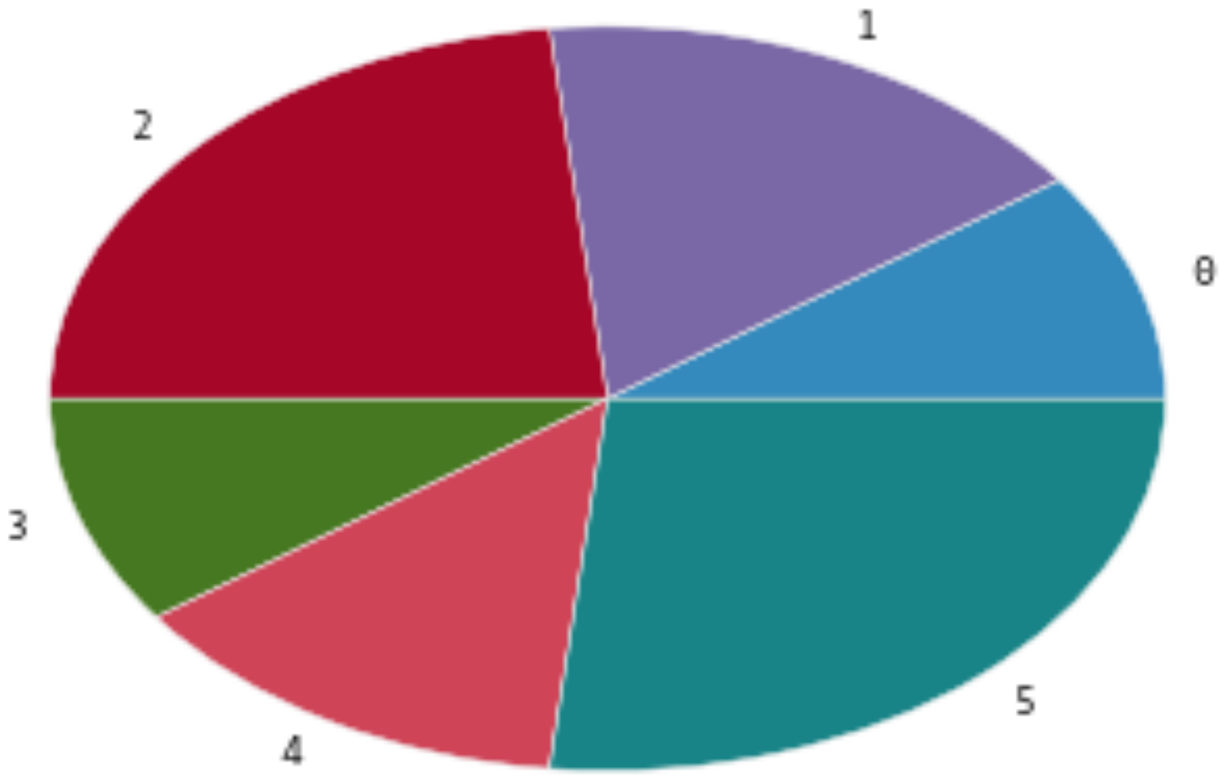
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
%matplotlib inline
pd.options.display.mpl_style = 'default'
```



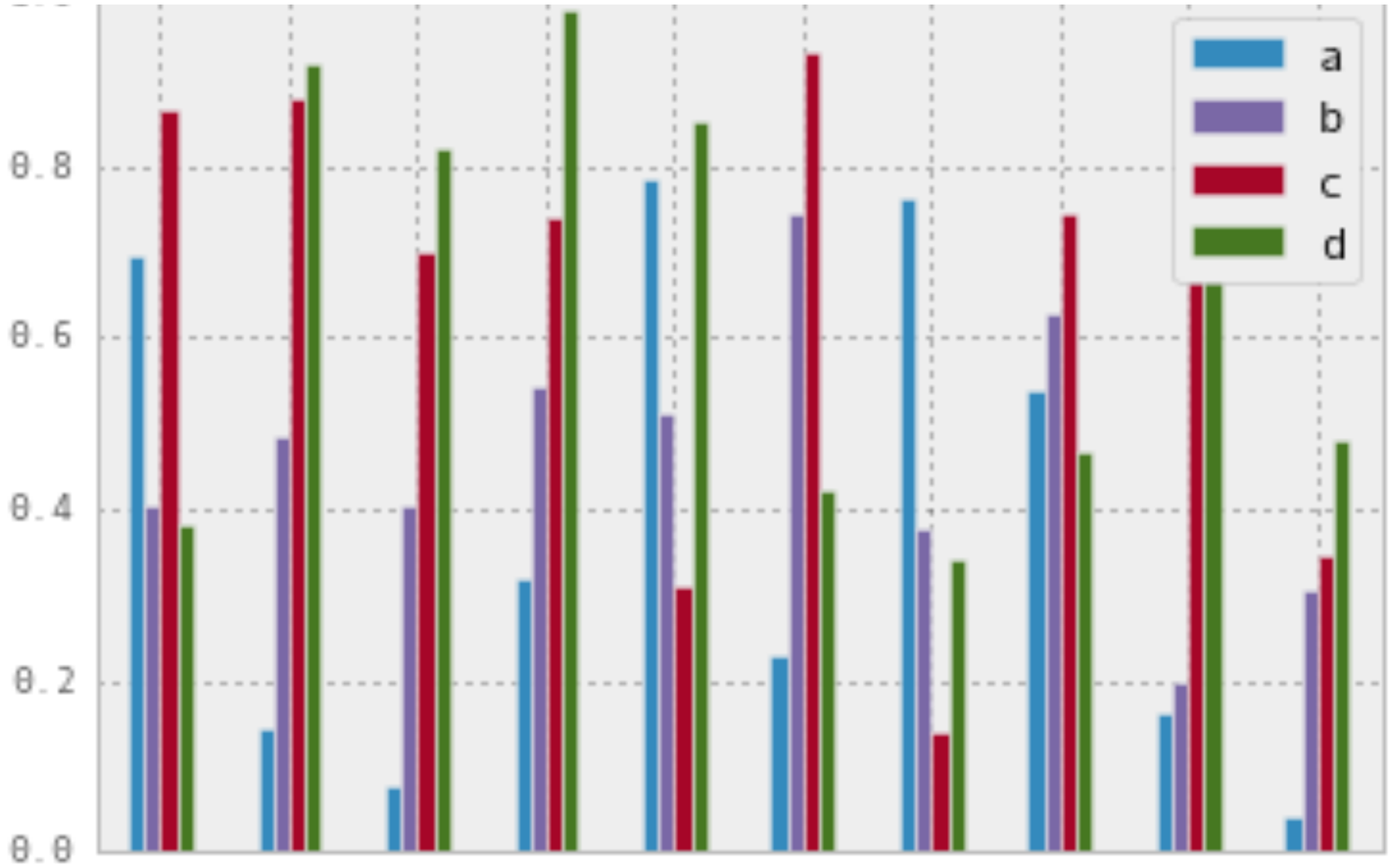

```
data = pd.Series( [3,4,6,7,5,8,2 ])
graph = data.plot()
```



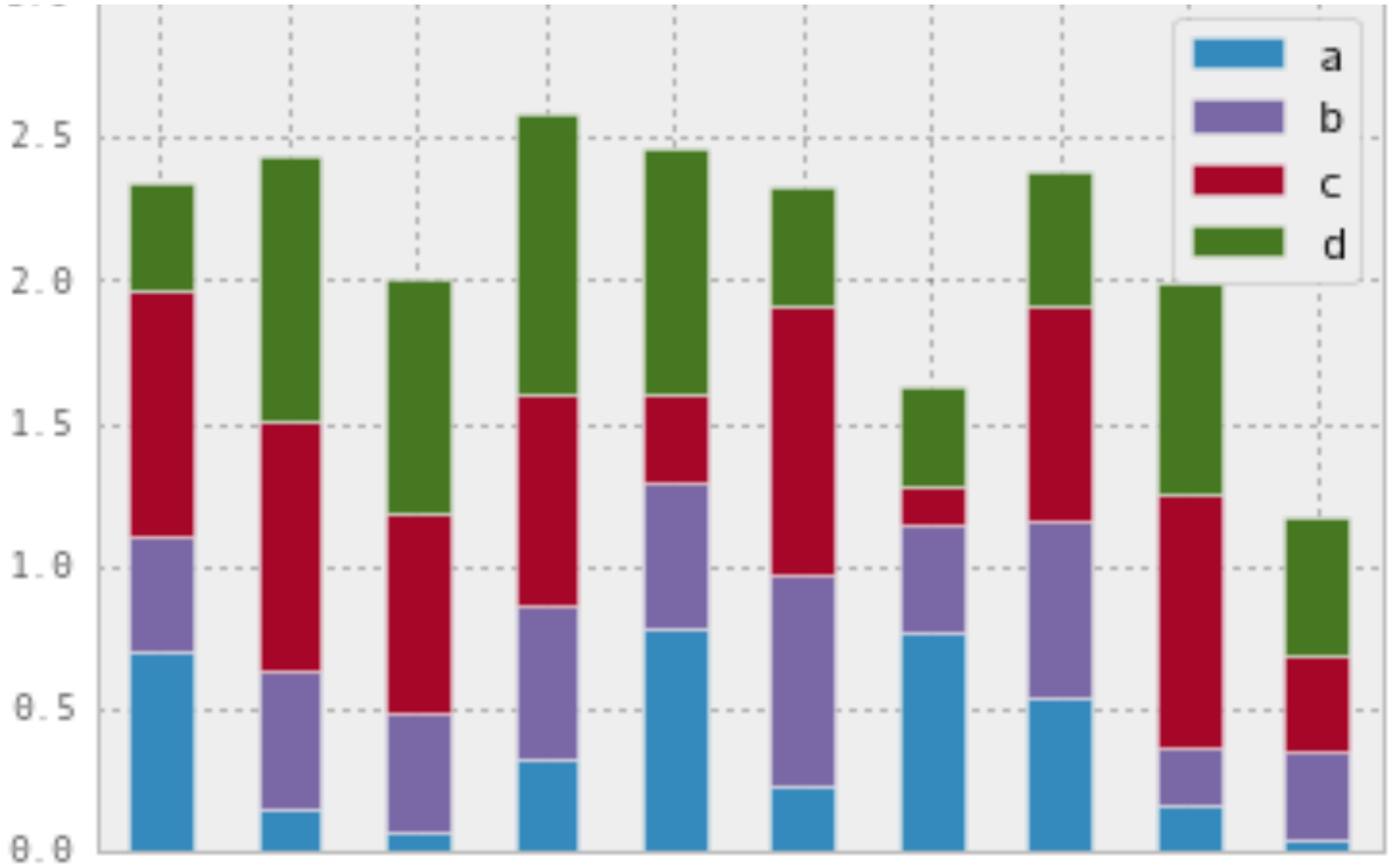

```
barchart = data.plot( kind="bar" )
```



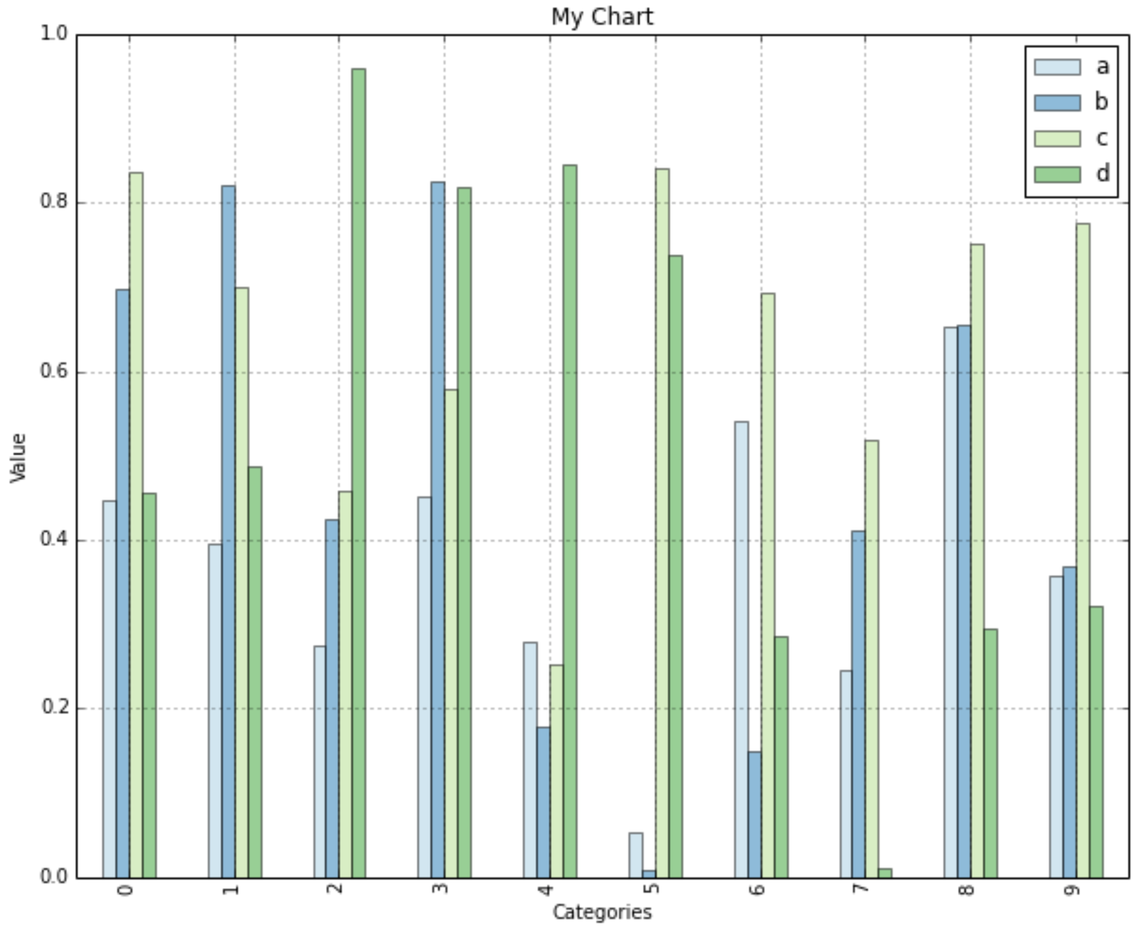

```
piechart = data.plot( kind="pie" )
```



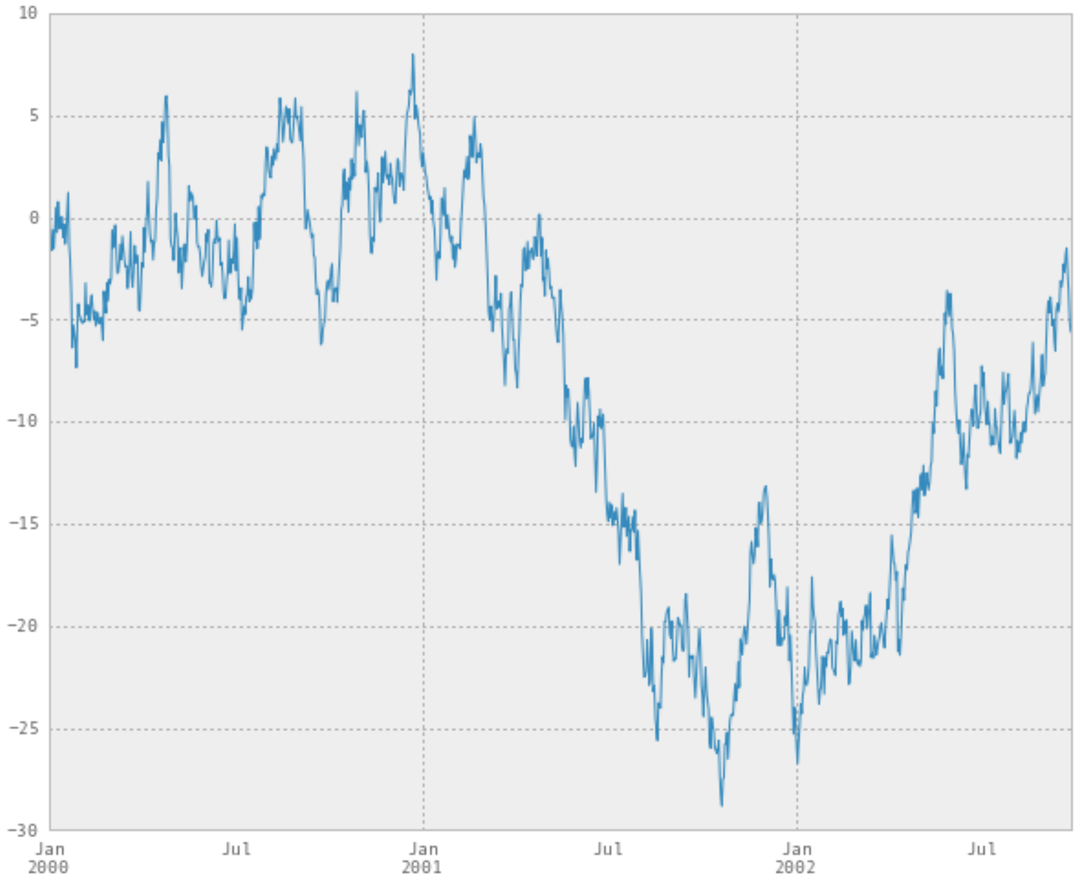

```
df2 = pd.DataFrame(np.random.rand(10, 4),
columns=['a', 'b', 'c', 'd'])
df2.plot( kind='bar' )
```



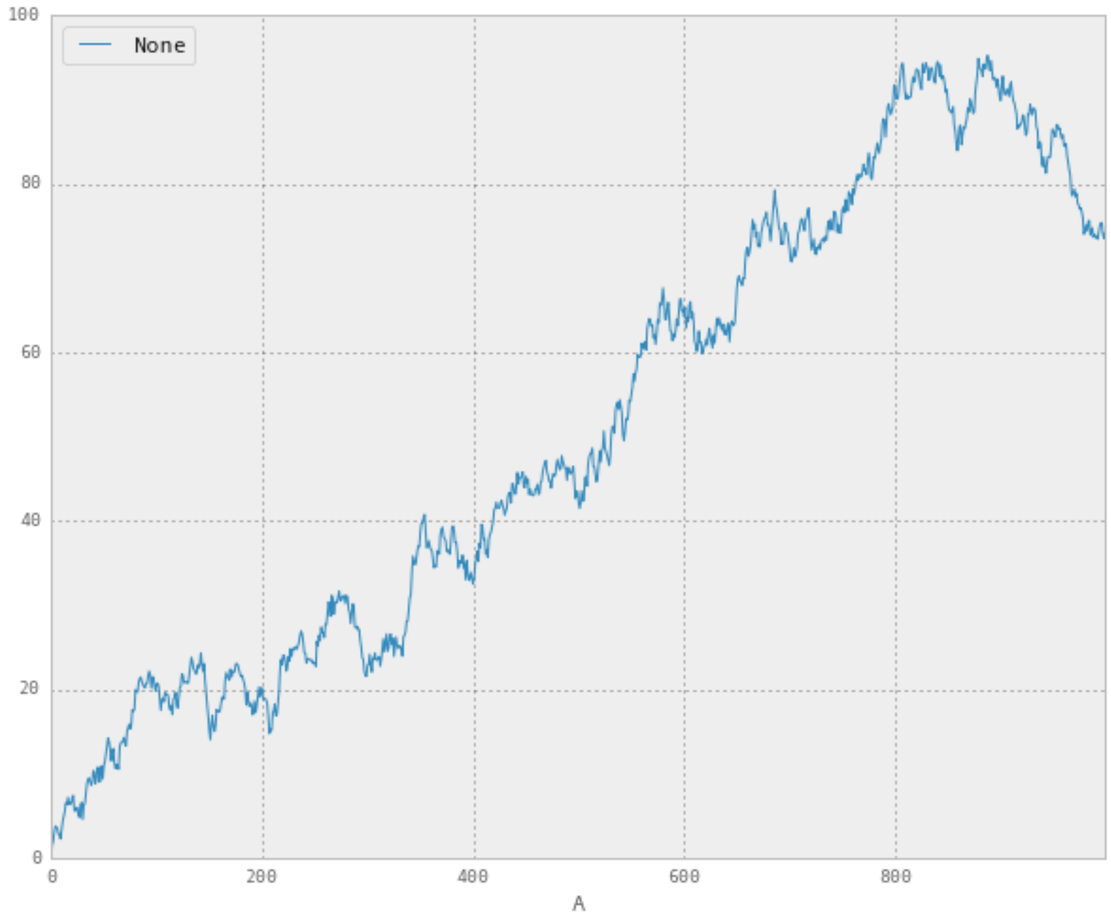

```
df2.plot( kind='bar', stacked=True )
```



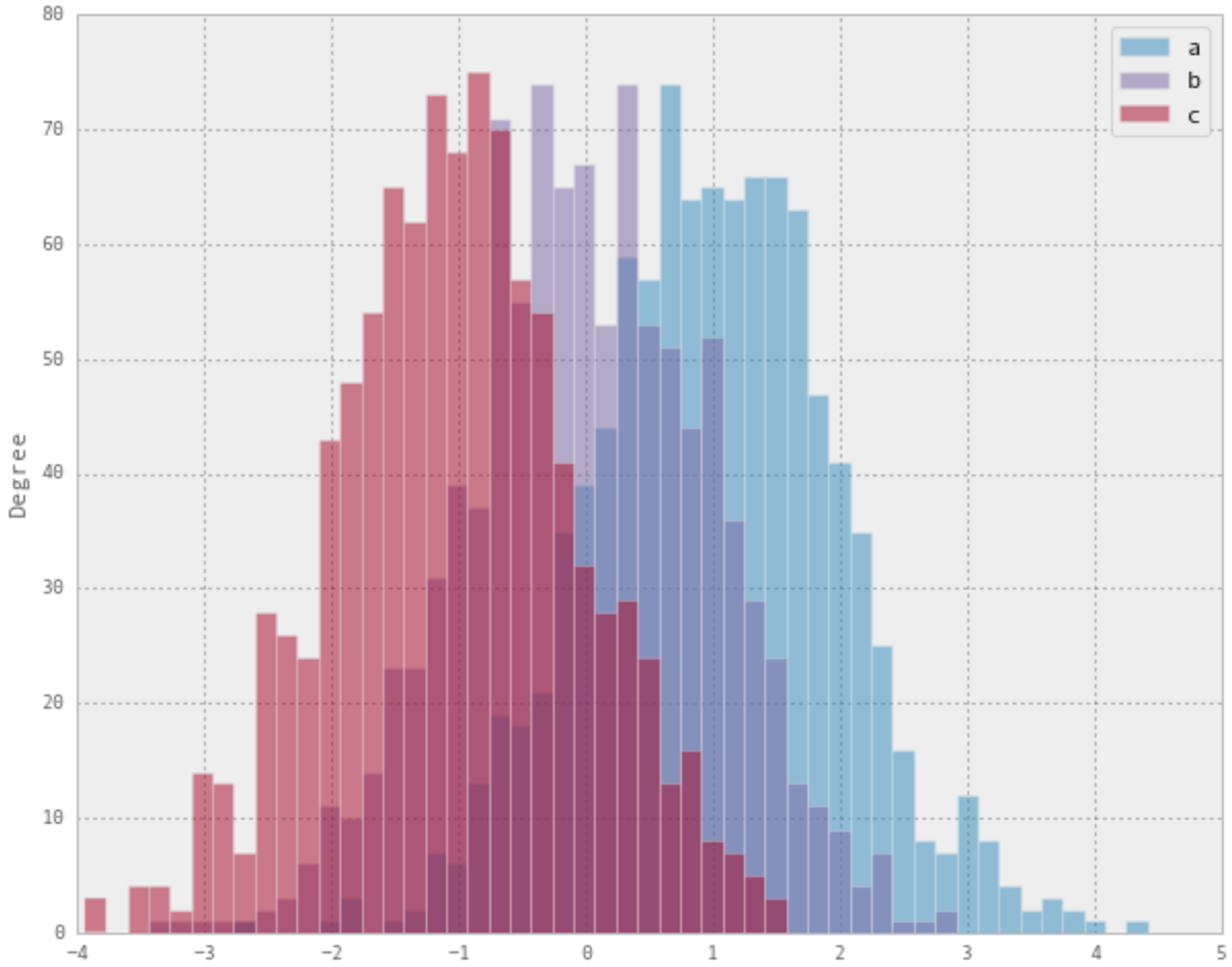

```
df2.plot( kind='bar',
         color=('#a6cee3','#1f78b4','#b2df8a','#33a02c' ),
          alpha=0.5,
            width=0.5,
            figsize=(10, 8))
plt.title( "My Chart" )
plt.xlabel( "Categories" )
plt.ylabel( "Value" )
```



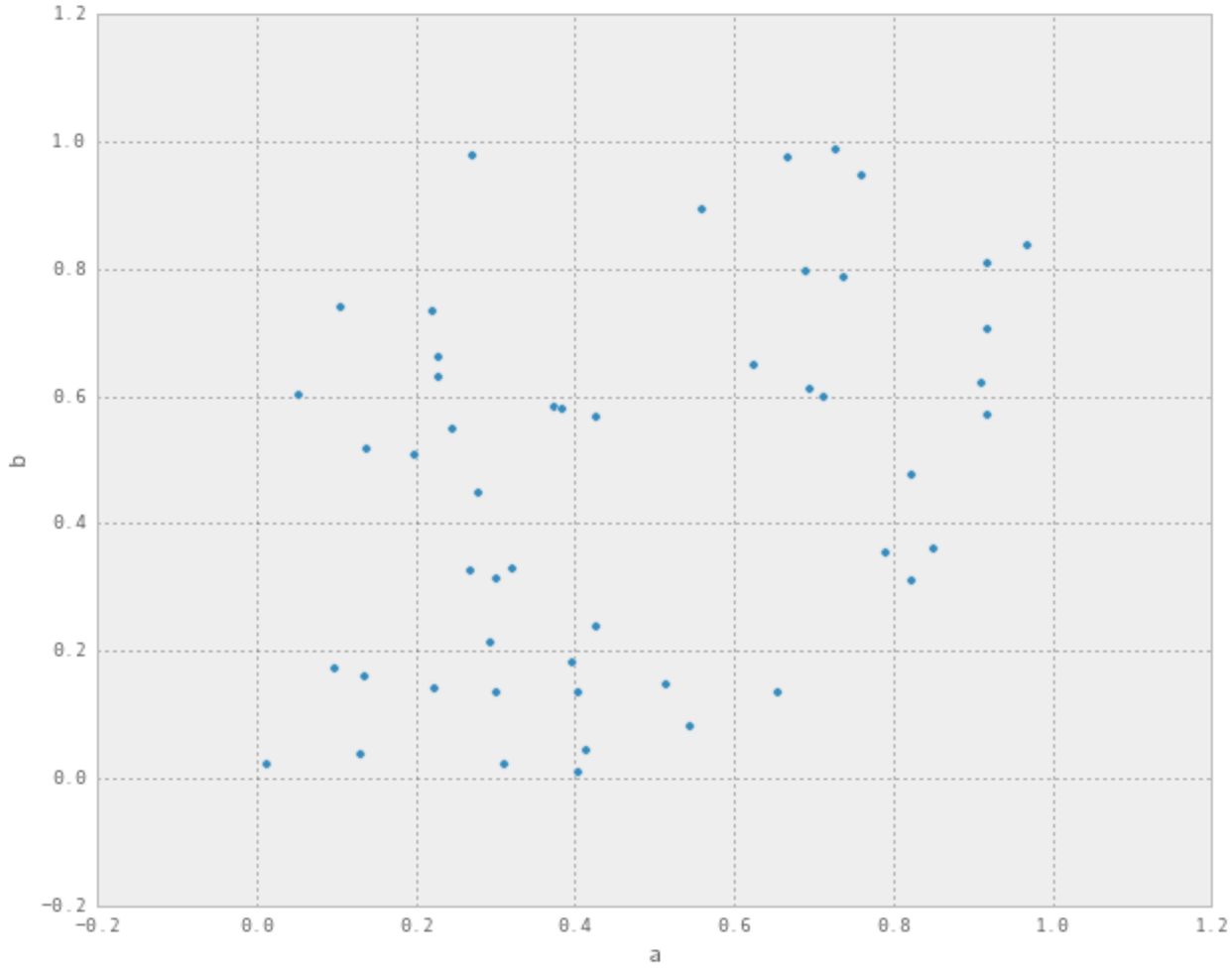

```
ts = pd.Series(np.random.randn( 1000),
index=pd.date_range('1/1/2000', periods=1000))
ts = ts.cumsum()
timeseriesChart = ts.plot( figsize=(10, 8) )
```



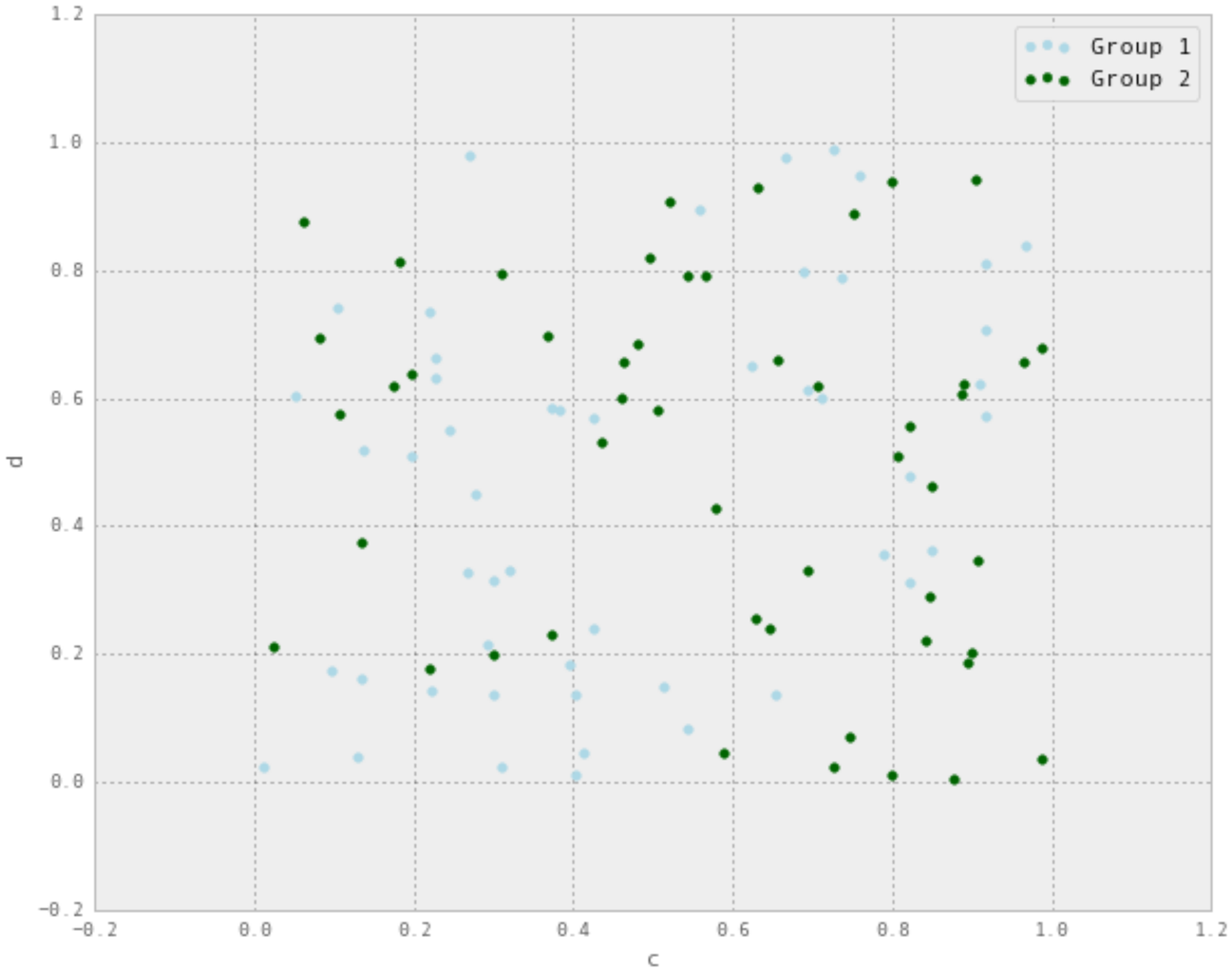

```
df3 = pd.DataFrame(np.random.randn(1000, 2),
columns=['B', 'C']).cumsum()
df3['A'] = pd.Series(list(range(len(df3))))

df3.plot( x='A', y='B' )
```



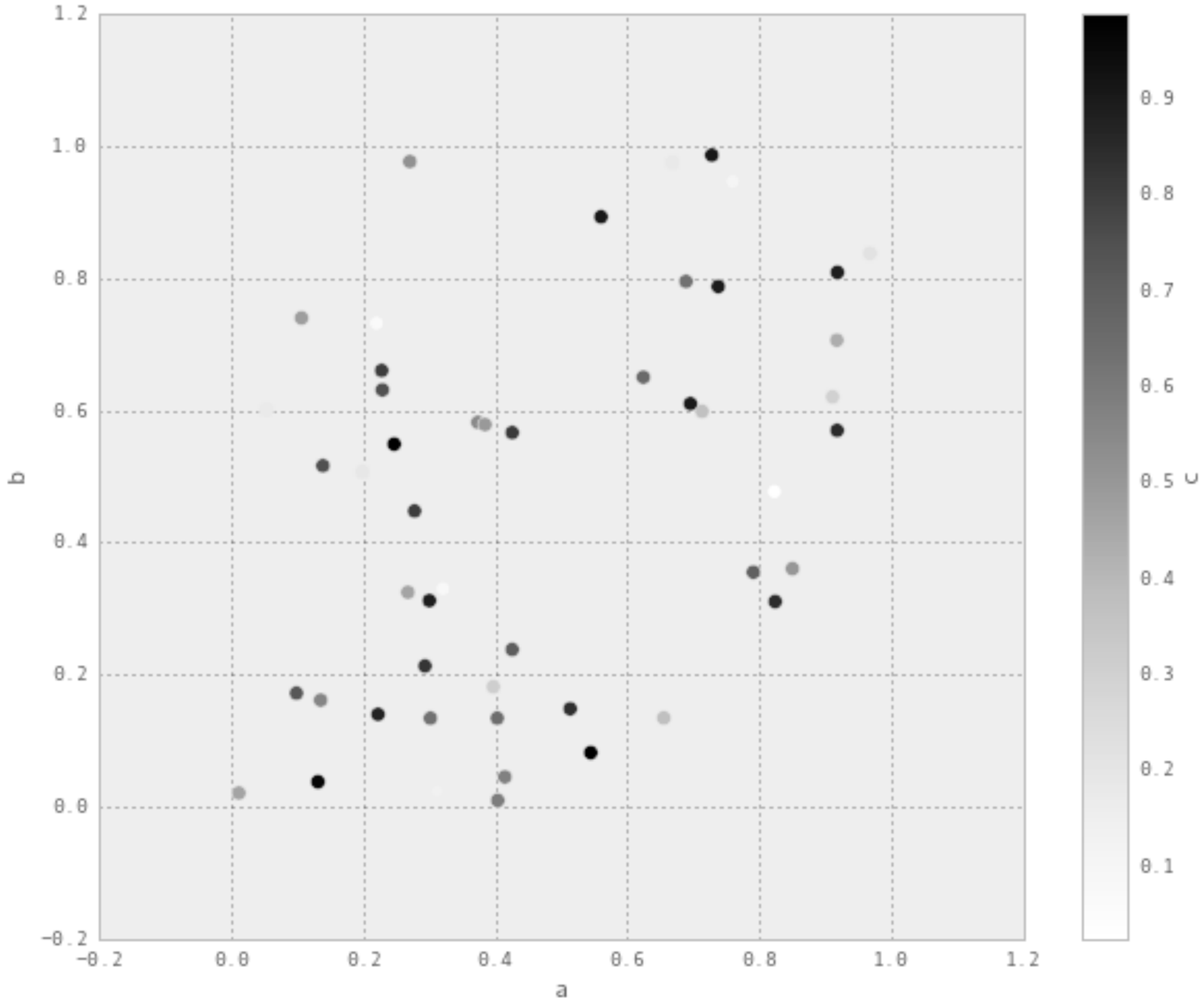

```
df4.plot(kind='hist',
         alpha=0.5,
         bins=50 )
```



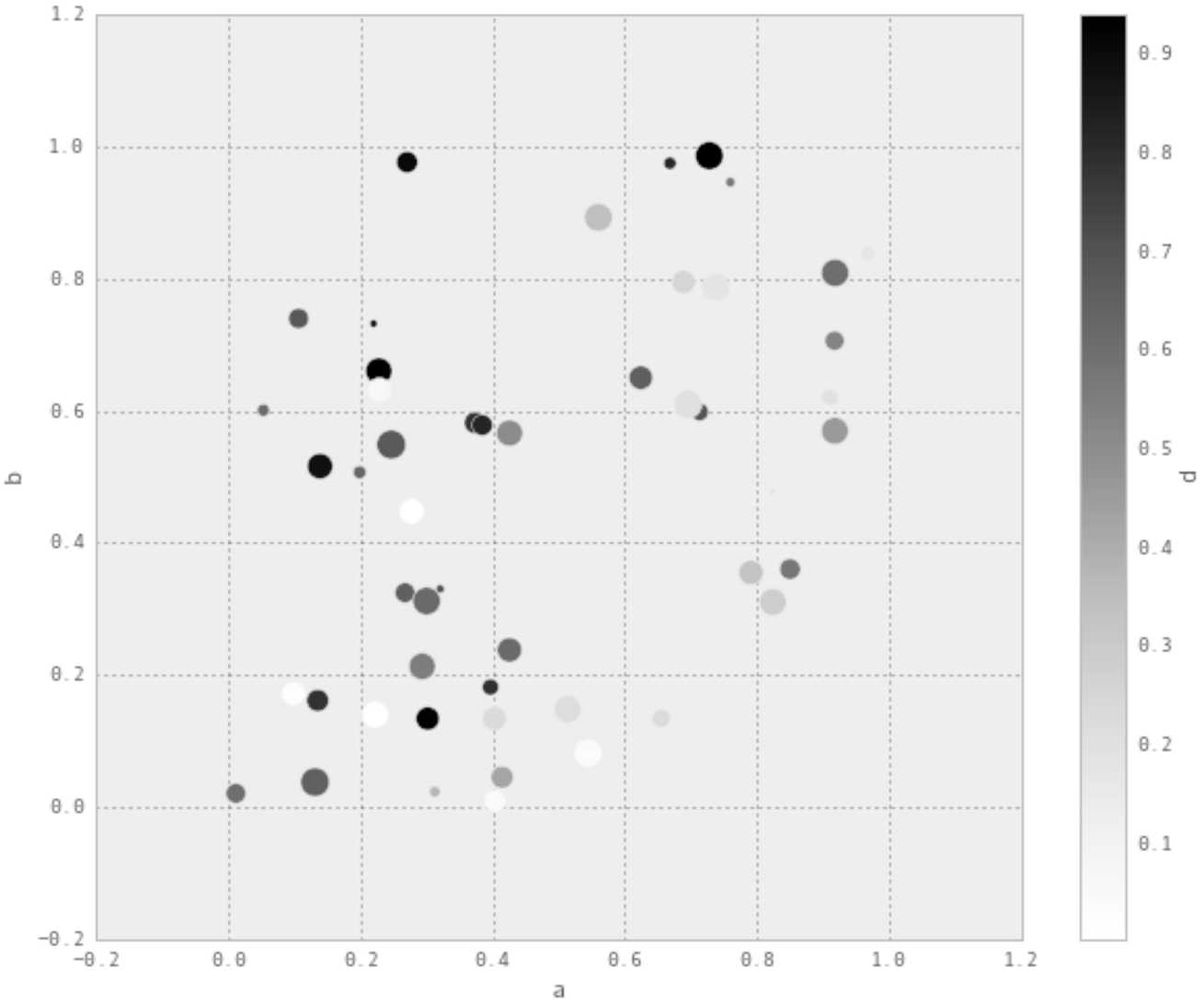

```
df5 = pd.DataFrame(np.random.rand(50, 4),
columns=['a', 'b', 'c', 'd'] )
df5.plot(kind='scatter', x='a', y='b' )
```



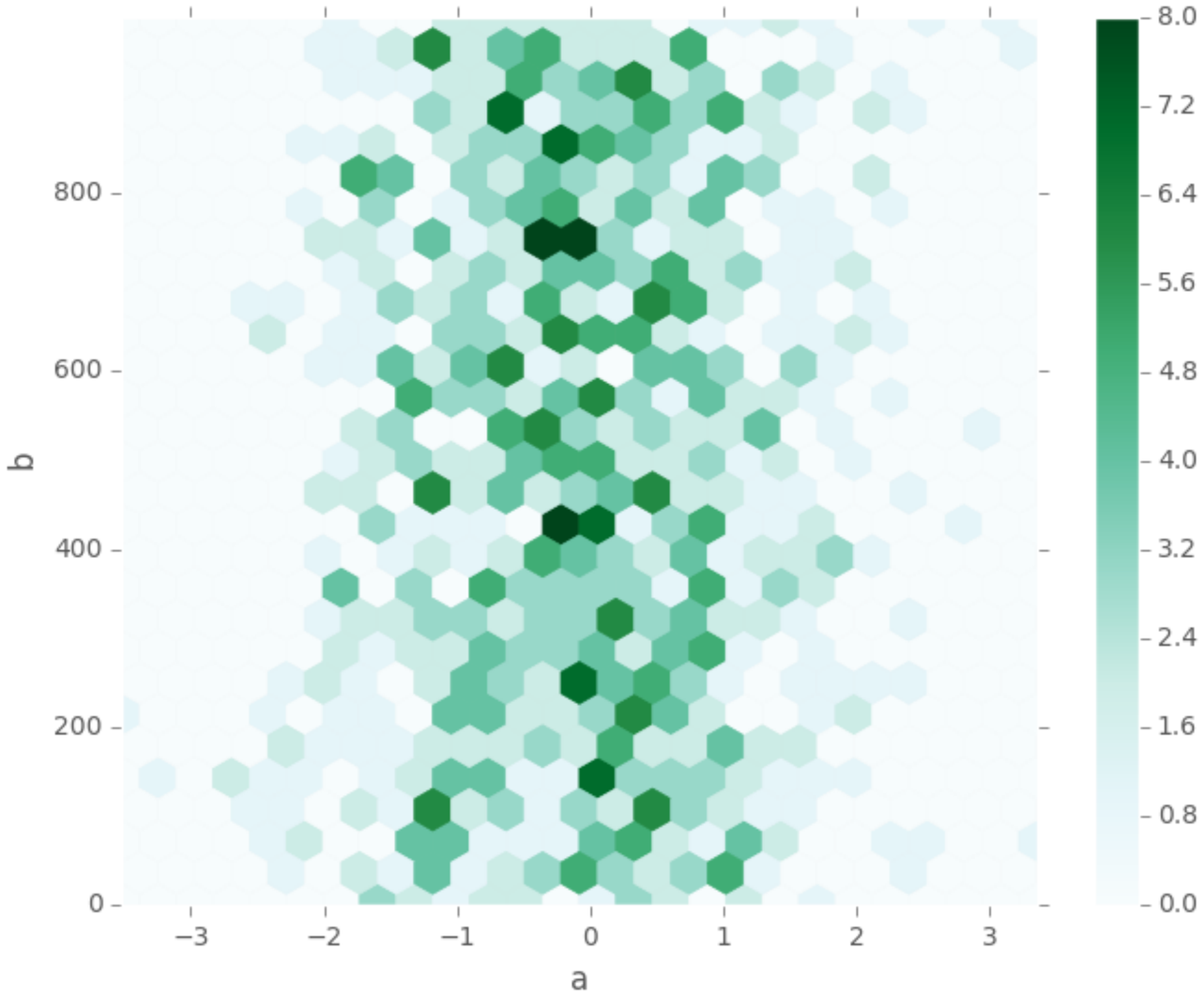

```
ax = df5.plot(kind='scatter', x='a', y='b',
              color='LightBlue',
              label='Group 1',
              figsize=(10, 8) )

df5.plot(kind='scatter', x='c', y='d',
             color='DarkGreen',
            label='Group 2',
            ax=ax)
```




```
df5.plot(kind='scatter', x='a', y='b',
c='c', s=50 )
```




```
scatter = df5.plot(kind='scatter',
                   x='a',
                   y='b',
                   s=df5['c']*200,
                   c='d')
```

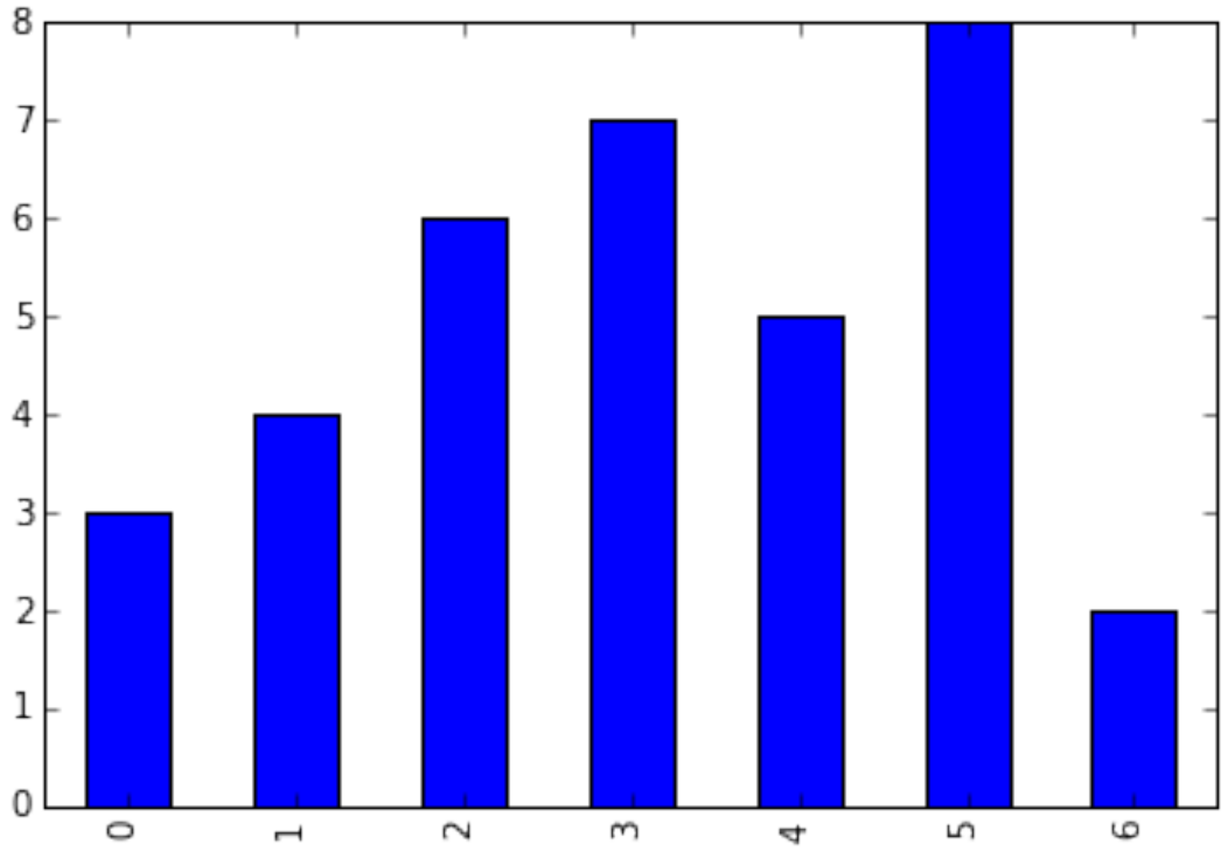



```
df.plot(kind='hexbin', x='a', y='b',
gridsize=25)
```

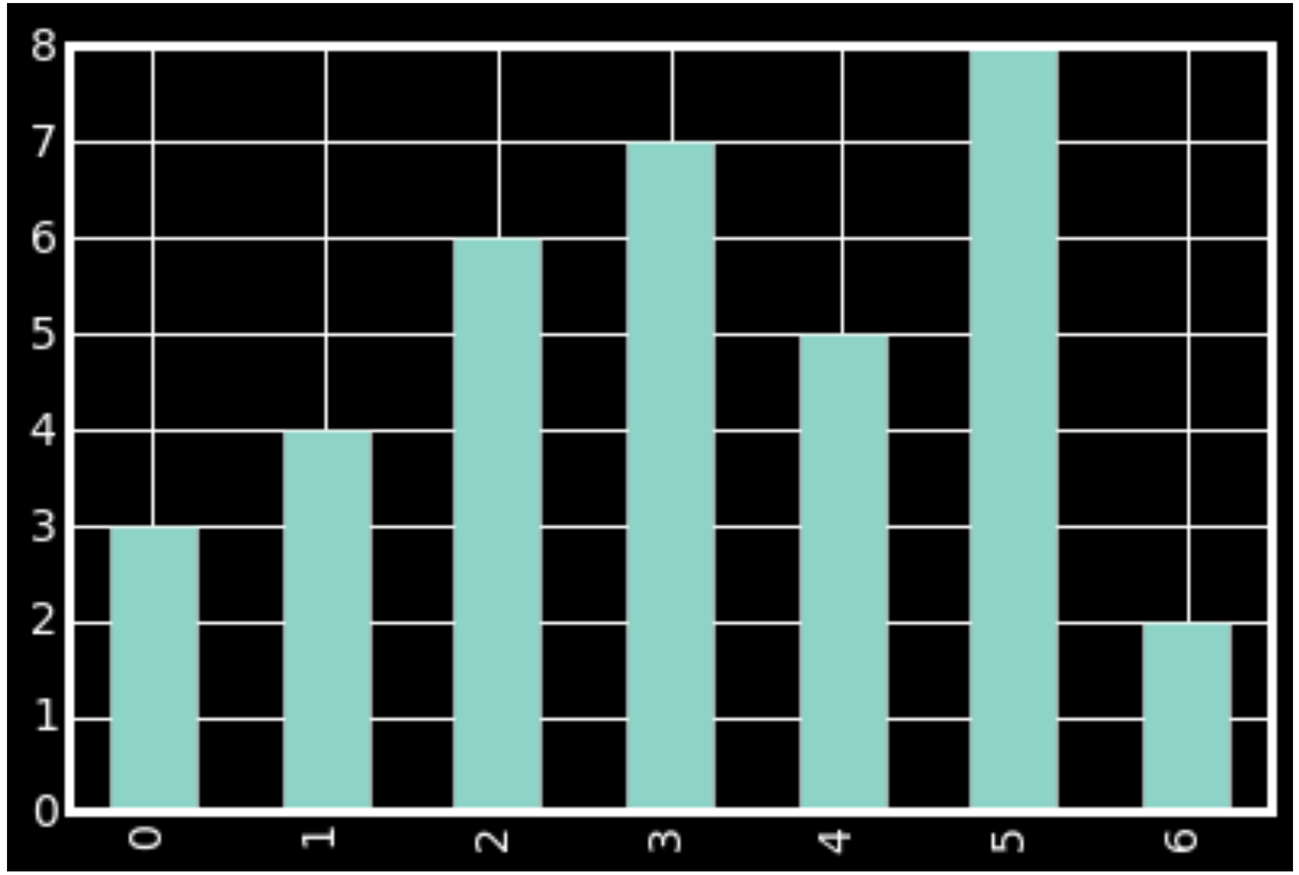
```
scatterPlot.get_figure().savefig( "scatterPlot.png" )
```

```
print(plt.style.available)

['dark_background', 'grayscale', 'ggplot',
'bmh', 'fivethirtyeight']
```

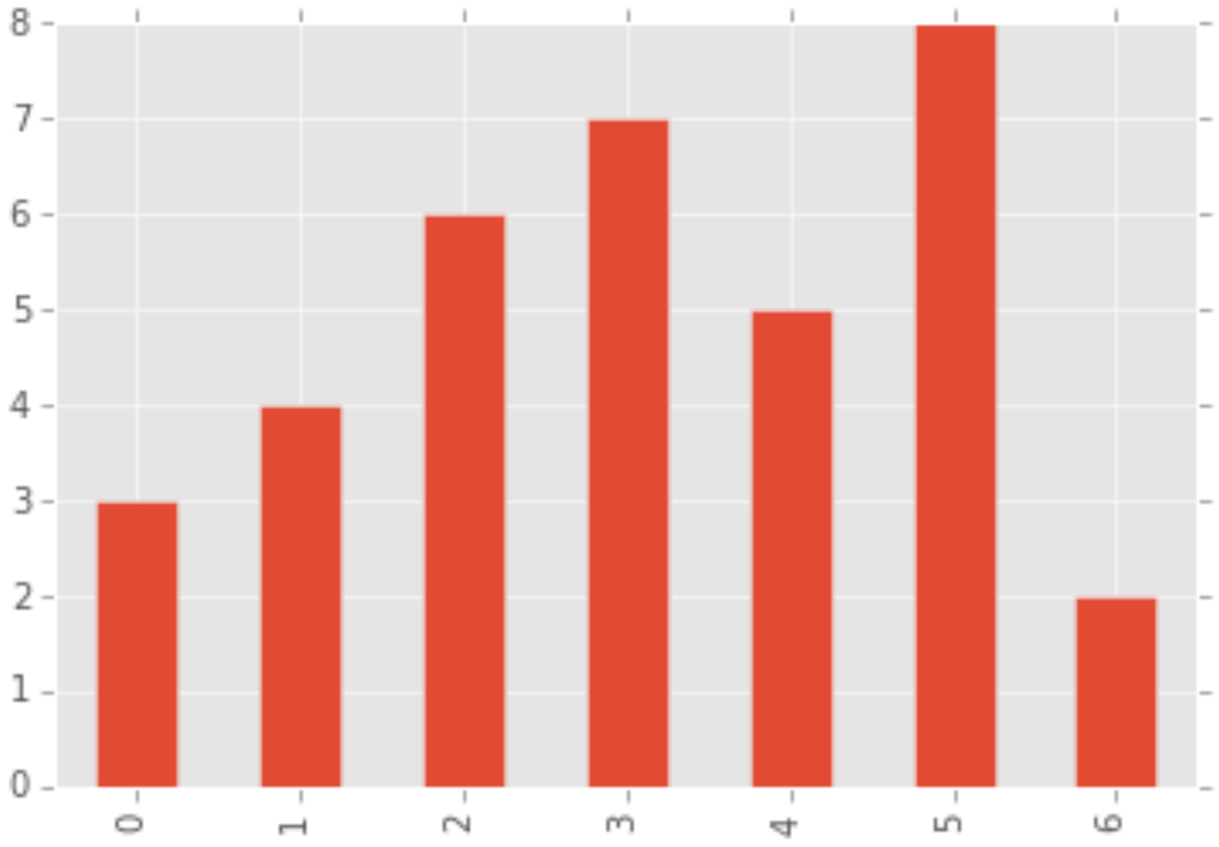
UGLY



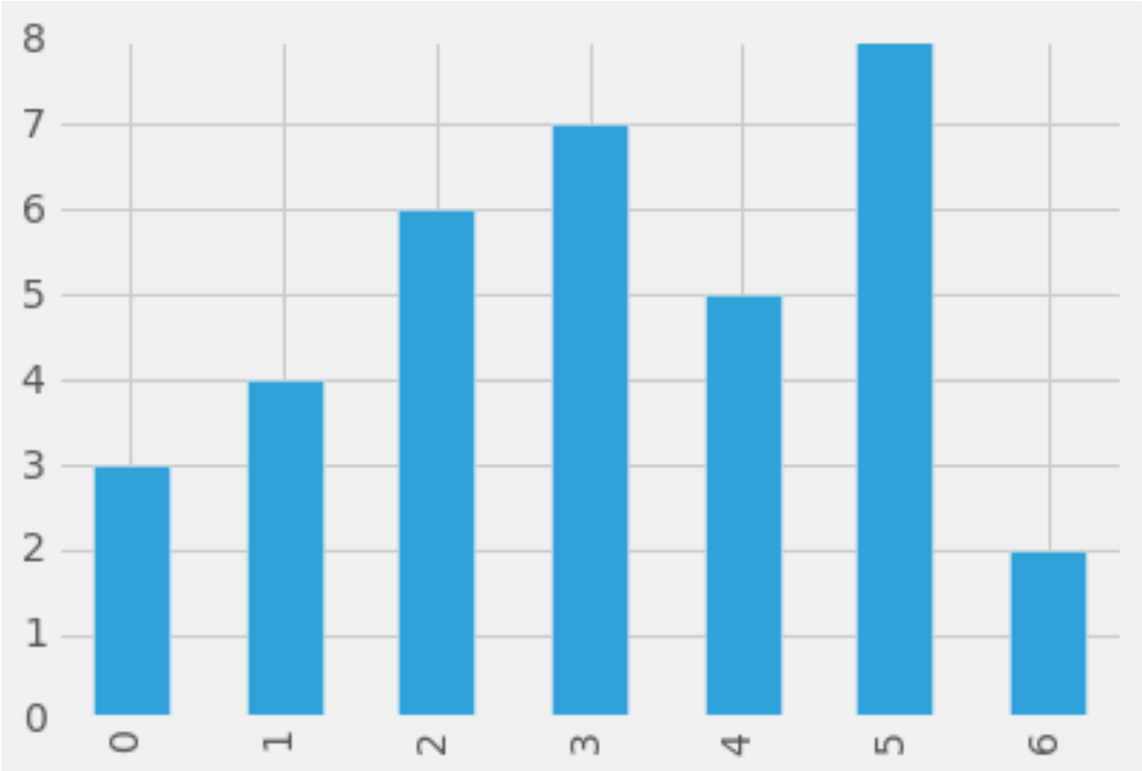

```
data = pd.Series( [3,4,6,7,5,8,2 ])
barchart = data.plot( kind="bar" )
```



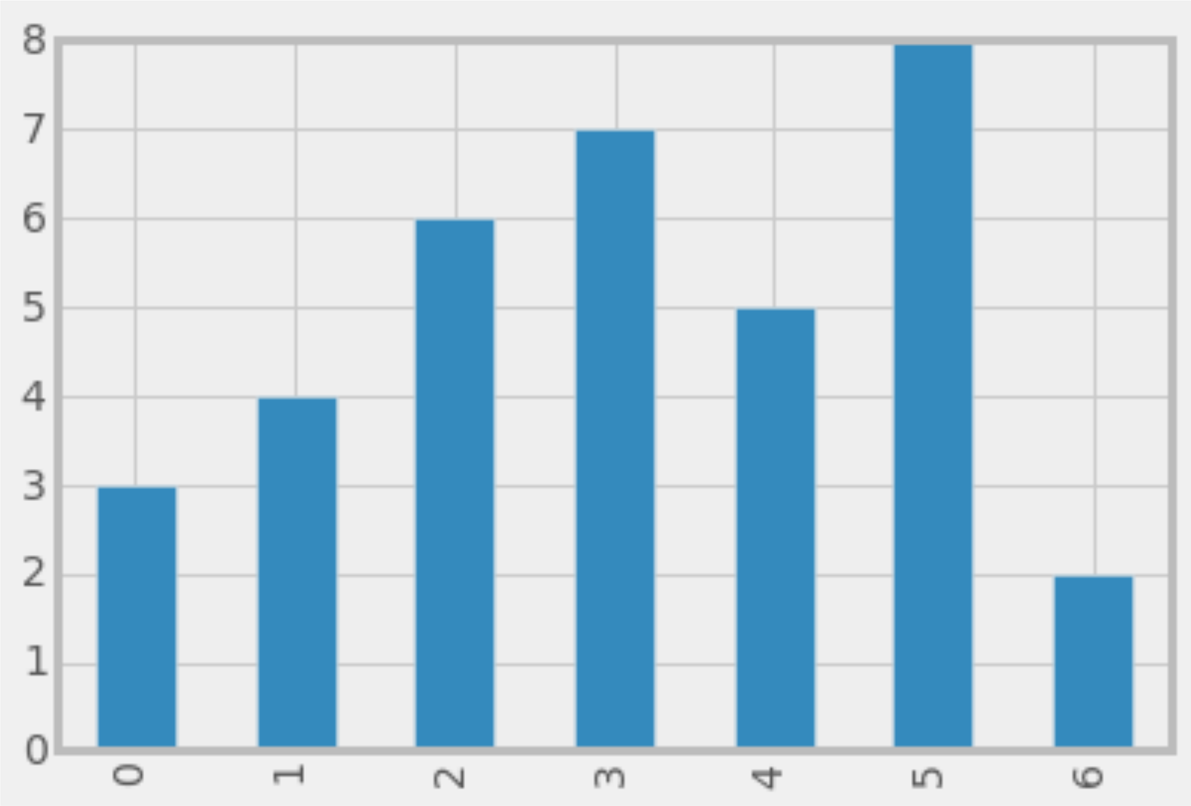

```
plt.style.use('dark_background')
barchart = data.plot( kind="bar" )
```




```
plt.style.use('ggplot')
barchart = data.plot( kind="bar" )
```




```
plt.style.use('fivethirtyeight')
barchart = data.plot( kind="bar" )
```

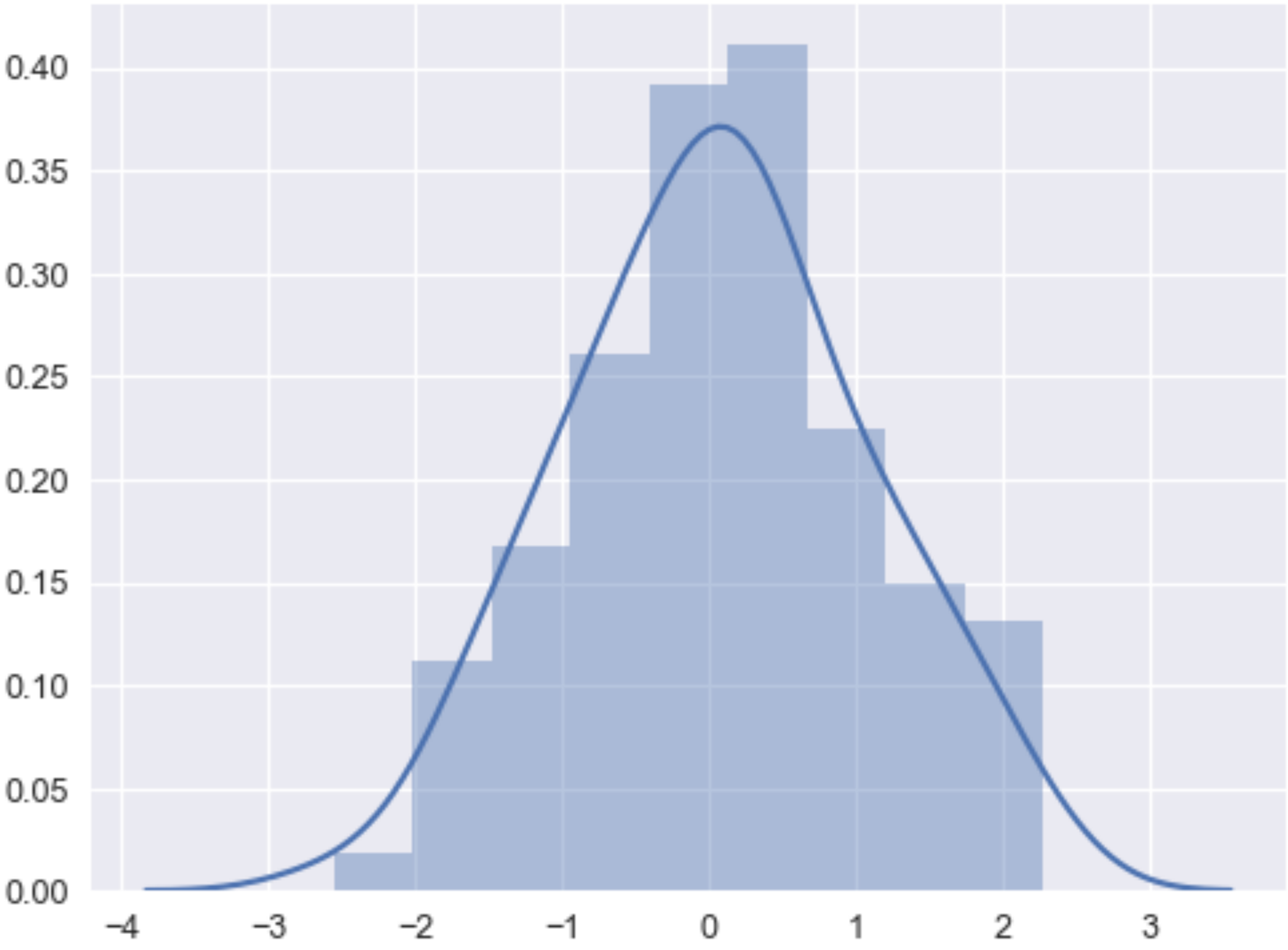



```
plt.style.use('bmh')
barchart = data.plot( kind="bar" )
```

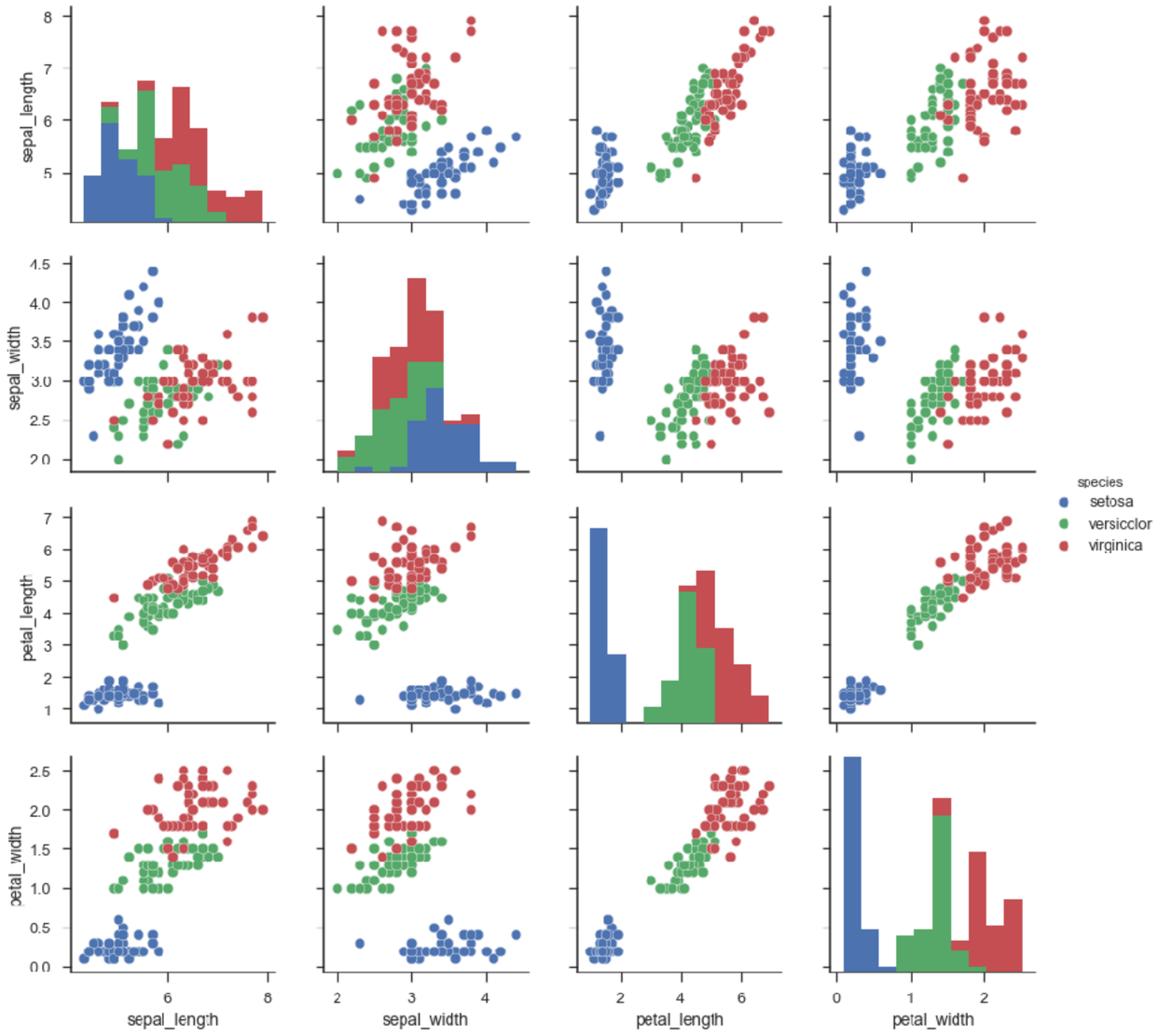
More Advanced Visualizations

```
import seaborn as sns
```

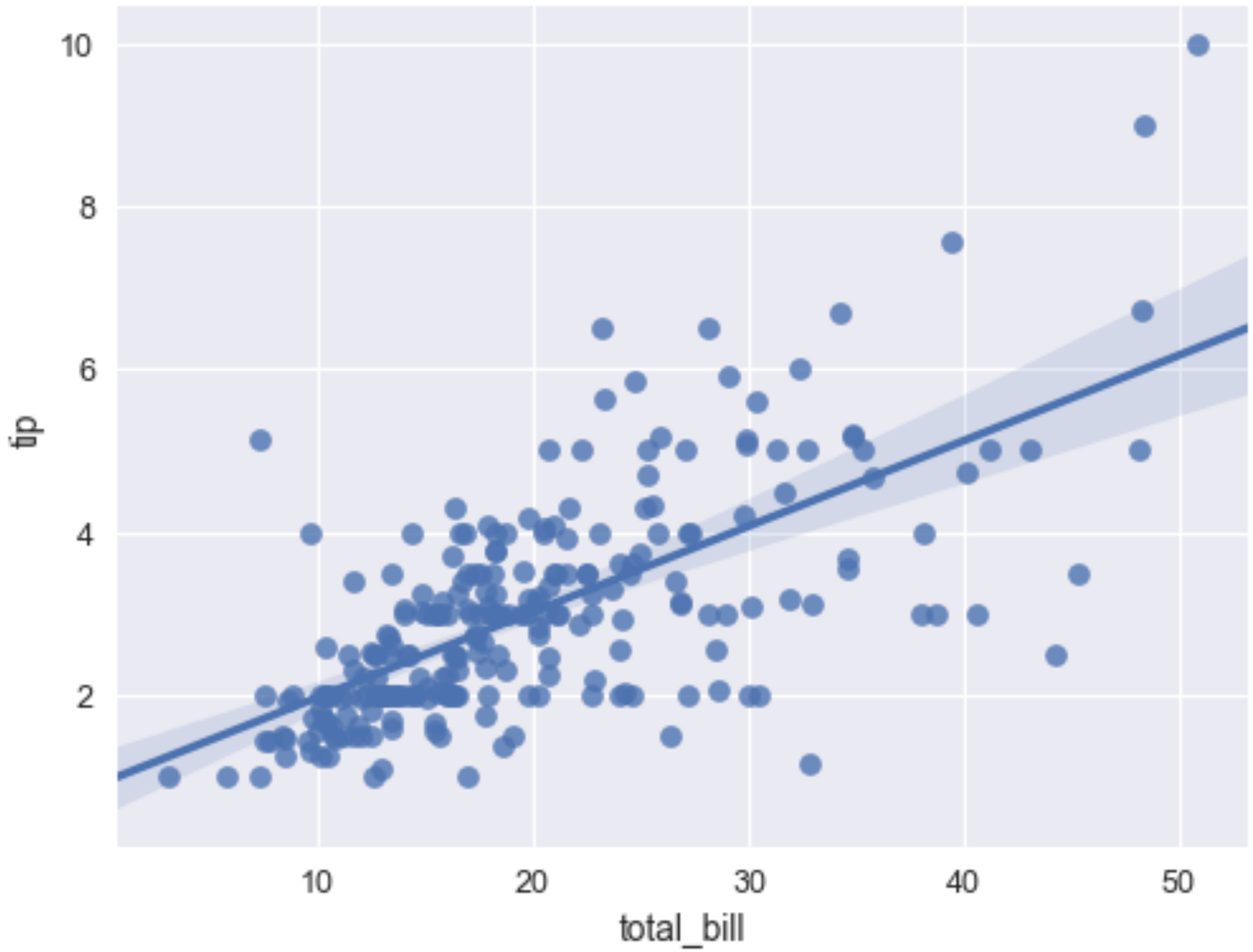
```
ax = sns.distplot(<data>)
```



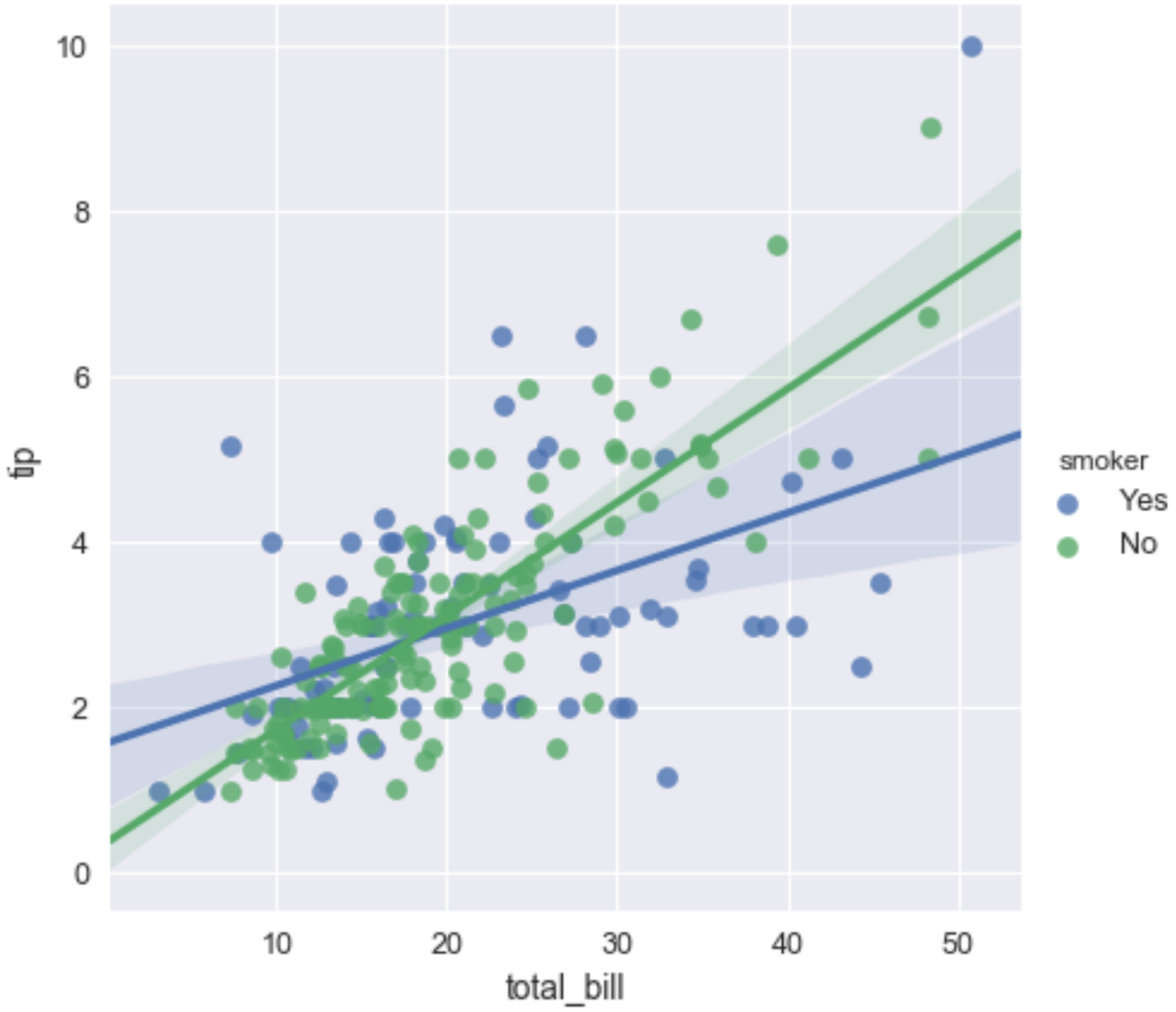
```
g = sns.pairplot(<data>,
hue="<target>")
```



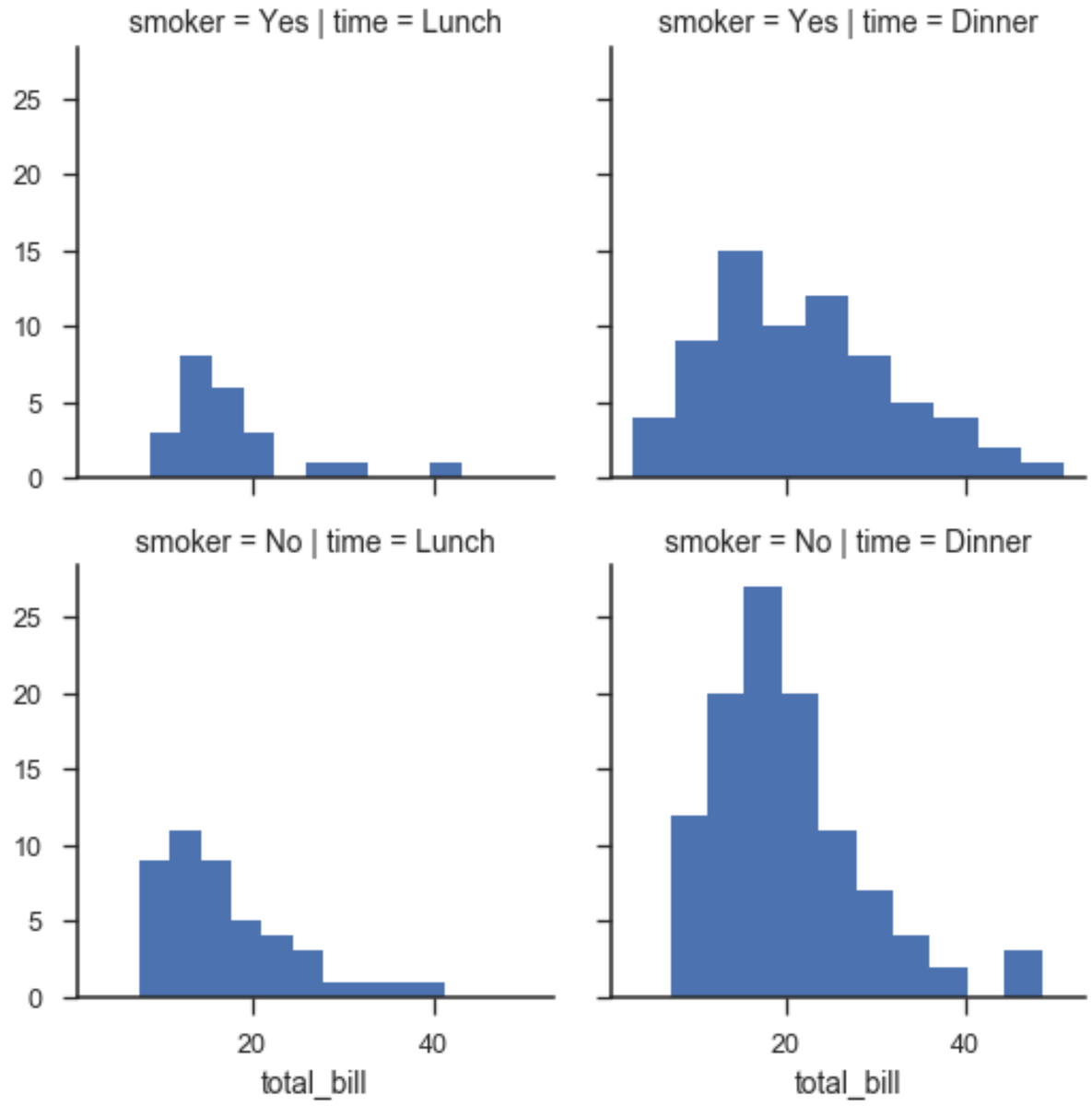
```
ax = sns.regplot(x="total_bill",
y="tip", data=tips)
```



```
g = sns.lmplot(x="total_bill",
y="tip", hue="smoker", data=tips)
```



```
g = sns.FacetGrid(tips, col="time",  row="smoker")
>>> g = g.map(plt.hist, "total_bill")
```



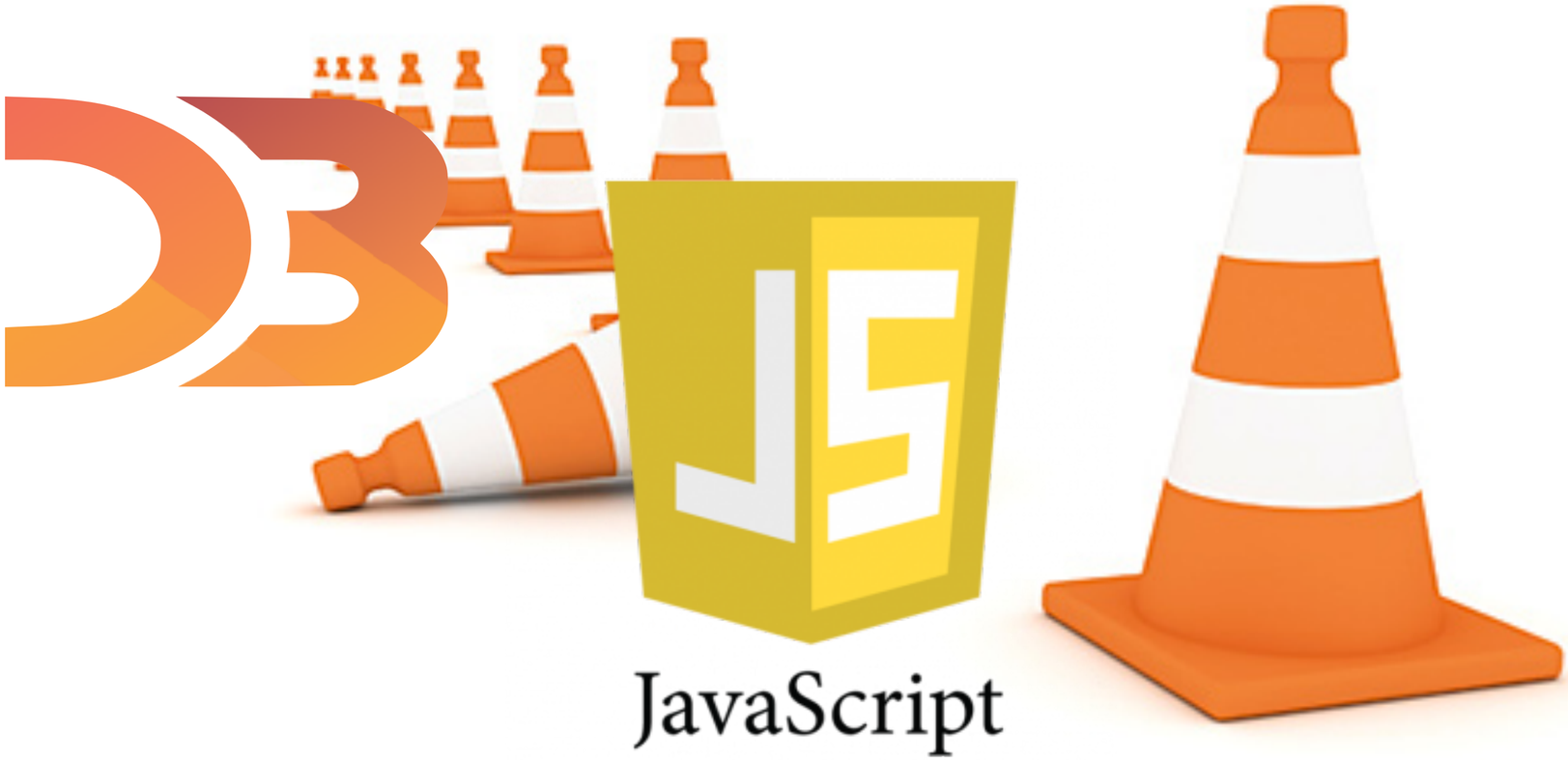
Interactive Visualizations

GTK CYBER
GET ON THE FAST TRACK

tableau
SOFTWARE

QlikView

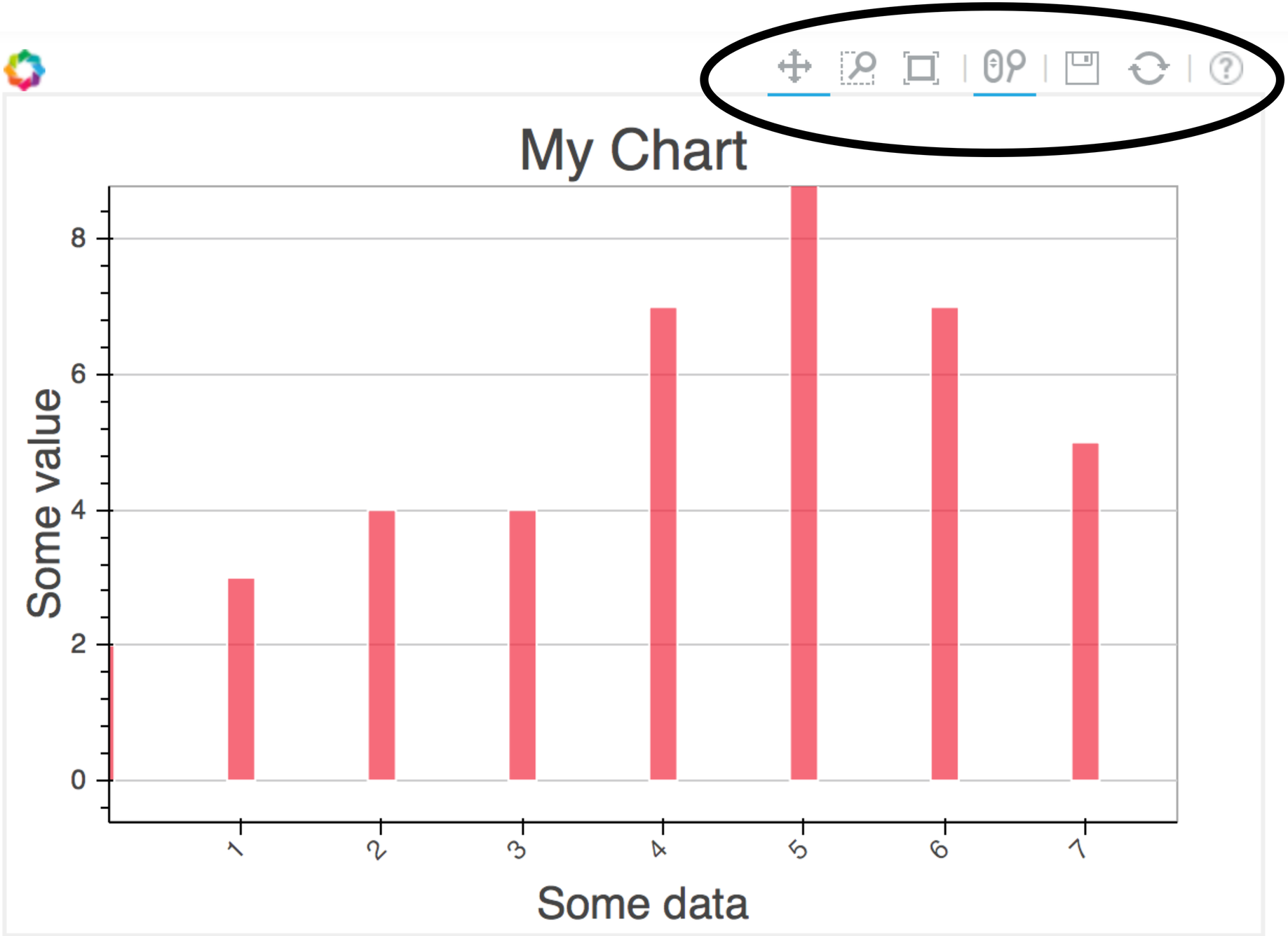
FEDERAL RESERVE NOTE
THE UNITED STATES OF AMERICA
AE 63456225 C
ONE HUNDRED DOLLARS


JavaScript

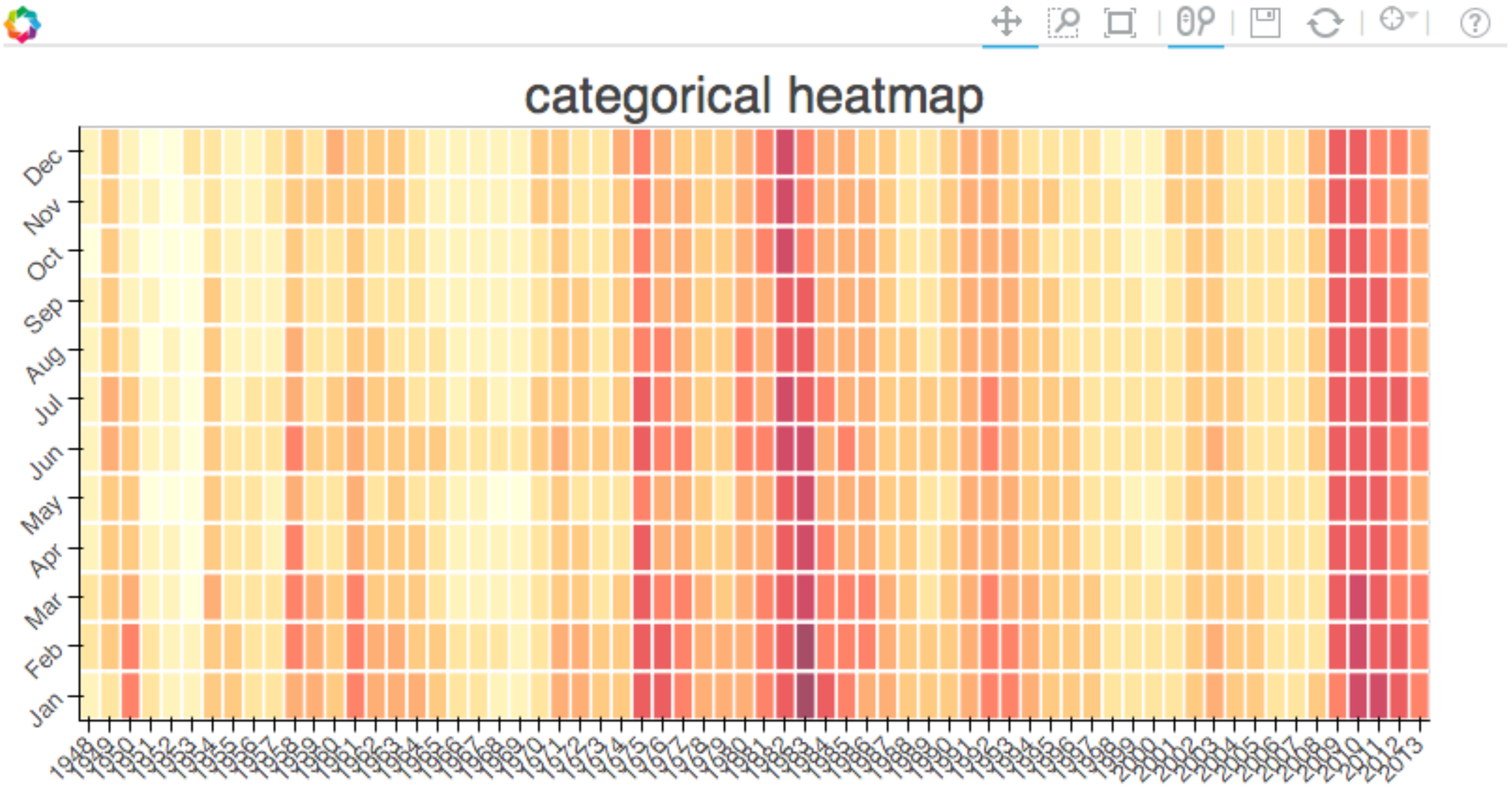
Shiny

Easy to use… if you know R

Introducing Bokeh



gtkcyber.com

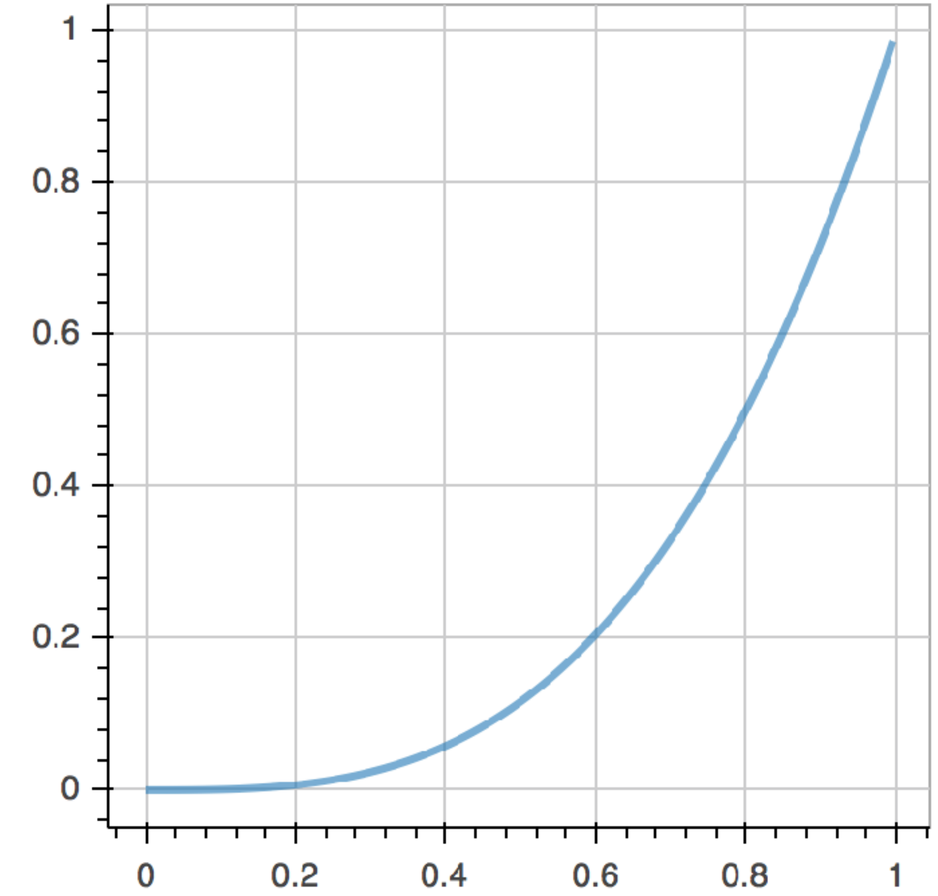


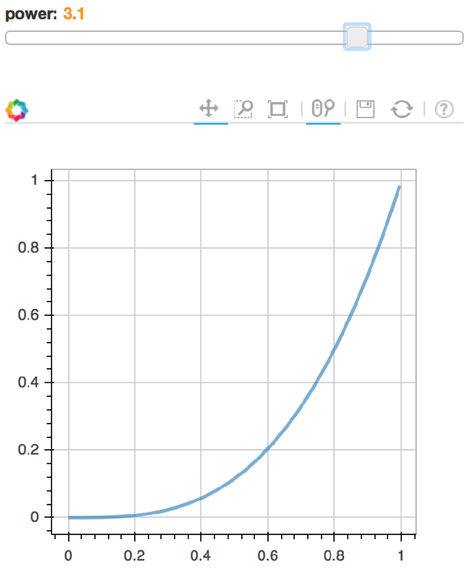
http://bokeh.pydata.org/en/latest/docs/gallery/cat_heatmap_chart.html

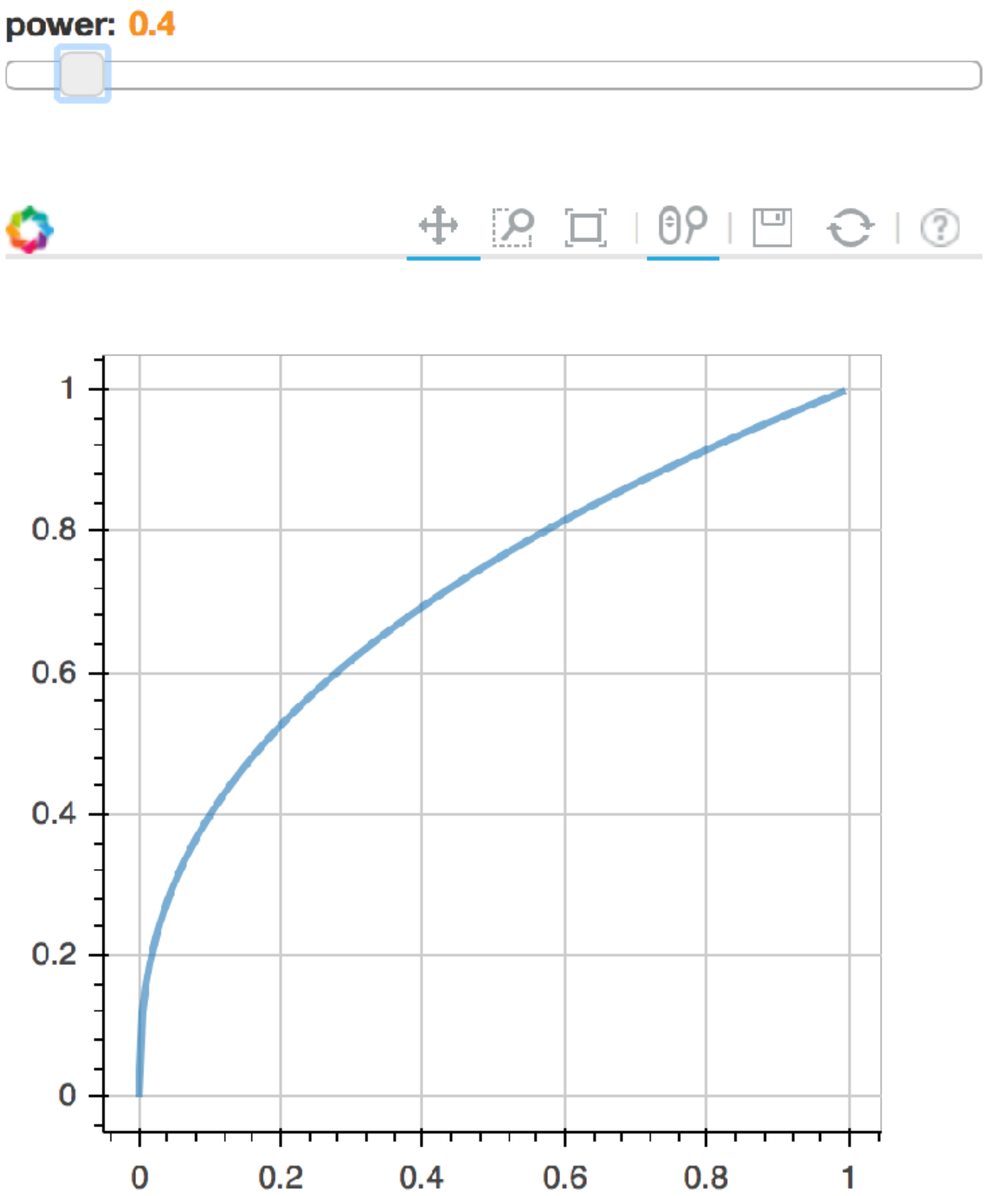
```
from bokeh.charts import HeatMap, output_file, show
from bokeh.palettes import YlOrRd9 as palette
from bokeh.sampledata.unemployment1948 import data

# pandas magic
df = data[data.columns[:-1]]
df2 = df.set_index(df[df.columns[0]].astype(str))
df2.drop(df.columns[0], axis=1, inplace=True)
df3 = df2.transpose()

output_file("cat_heatmap.html")

palette = palette[::-1]  # Reverse the color order so dark red is highest un
hm = HeatMap(df3, title="categorical heatmap", width=800, palette=palette)

show(hm)
```

http://bokeh.pydata.org/en/latest/docs/gallery/cat_heatmap_chart.html

power: 3.1




power: 3.1
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1


power: 0.4
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

```python
from bokeh.io import vform
from bokeh.models import Callback, ColumnDataSource, Slider
from bokeh.plotting import figure, output_file, show

x = [x*0.005 for x in range(0, 200)]
y = x

source = ColumnDataSource(data=dict(x=x, y=y))

plot = figure(plot_width=400, plot_height=400)
plot.line('x', 'y', source=source, line_width=3, line_alpha=0.6)

callback = Callback(args=dict(source=source), code="""
        var data = source.get('data');
        var f = cb_obj.get('value')
        x = data['x']
        y = data['y']
        for (i = 0; i < x.length; i++) {
            y[i] = Math.pow(x[i], f)
        }
        source.trigger('change');
    """)

slider = Slider(start=0.1, end=4, value=1, step=.1, title="power", callback=callback)

layout = vform(slider, plot)

show(layout)
```

Using Bokeh

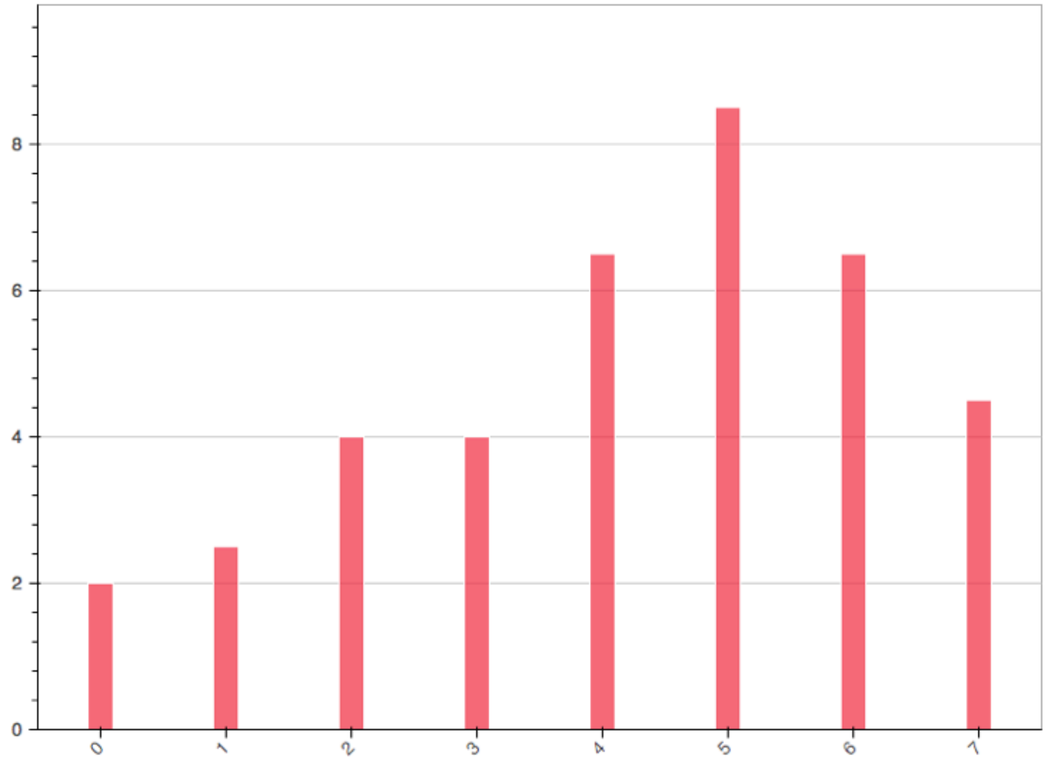
```
from bokeh.plotting import output_notebook
output_notebook()
```

- Area (Overlapped & Stacked)
- Bar (Grouped & Stacked)
- BoxPlot
- Donut (Ick!)
- HeatMap
- Line
- Scatter
- Step
- Timeseries

```
from bokeh.charts import <chartname>
from bokeh.io import show
```

```
from bokeh.charts import Bar, show
data = [2,3,4,4,7,9,7,5]

barchart = Bar( data, notebook=True )
show()
```

```
from bokeh.charts import Bar
from bokeh.io import show
data = [2,3,4,4,7,9,7,5]

barchart = Bar( data, notebook=True )
show( bar chart )
```



Method Chaining

```
barchart =
   Bar( data, title="My Chart", xlabel="Categories", ylabel="Value",
notebook=True )

show( barchart )
```

```
barchart =
   Bar( data, title="My Chart", xlabel="Categories", ylabel="Value",
notebook=True )

barchart.show()
```

```
barchart3 = Bar(data)
   .title( "My Chained Chart" )
   .notebook( True )
   .legend( True )
   .xlabel( "Categories" )
   .ylabel( "Value" )

show( barchart3 )
```

Bokeh Config Options

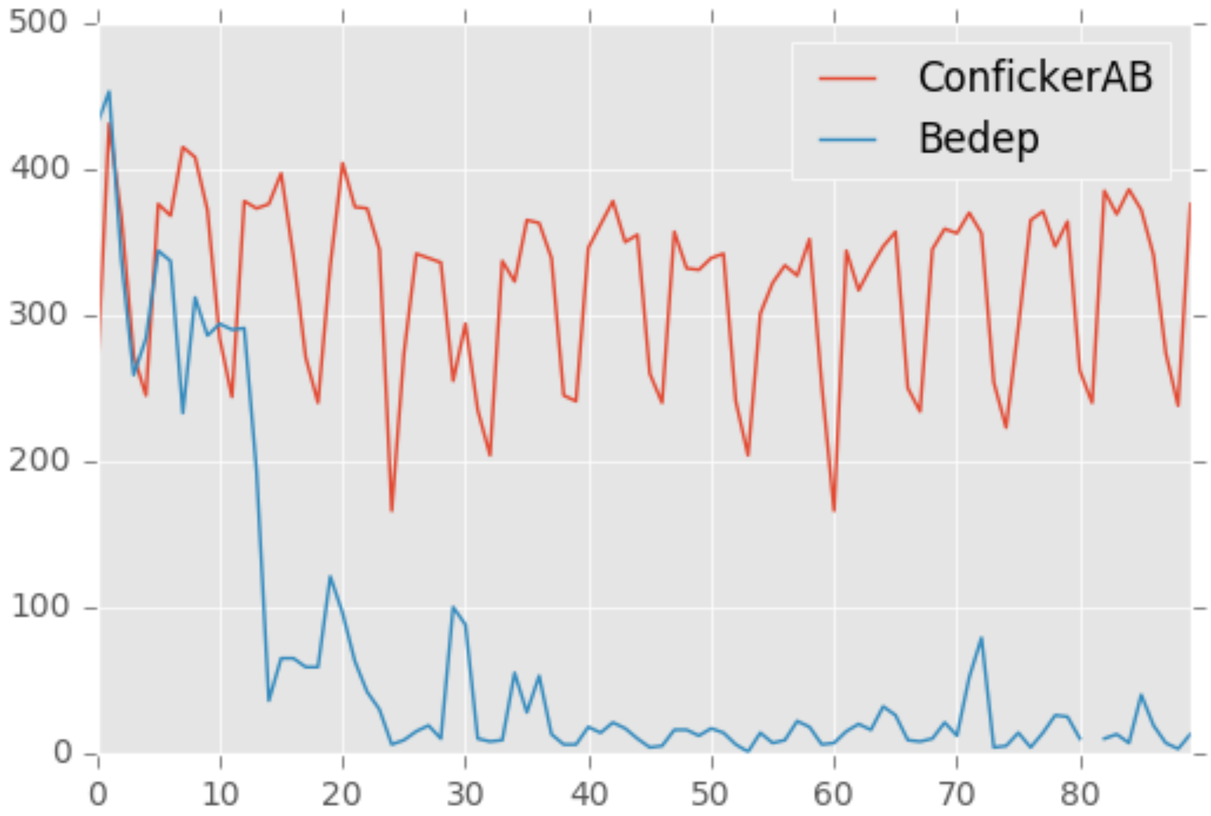
- **title** (str): the title of your plot.
- **xlabel** (str): the x-axis label of your plot.
- **ylabel** (str): the y-axis label of your plot.
- **legend** (str, bool): the legend of your plot.
- **xscale** (str): the x-axis type scale of your plot.
- **yscale** (str): the y-axis type scale of your plot.
- **width** (int): the width of your plot in pixels.
- **height** (int): the height of you plot in pixels.
- **tools** (bool): to enable or disable the tools in your plot.
- **filename** (str or bool): the name of the file where your plot will be written.
- **server** (str or bool): the name of your plot in the server.
- **notebook** (bool):if you want to output (or not) your plot into the IPython notebook.

In Class Exercise

Please complete Worksheet 5: Data Visualization

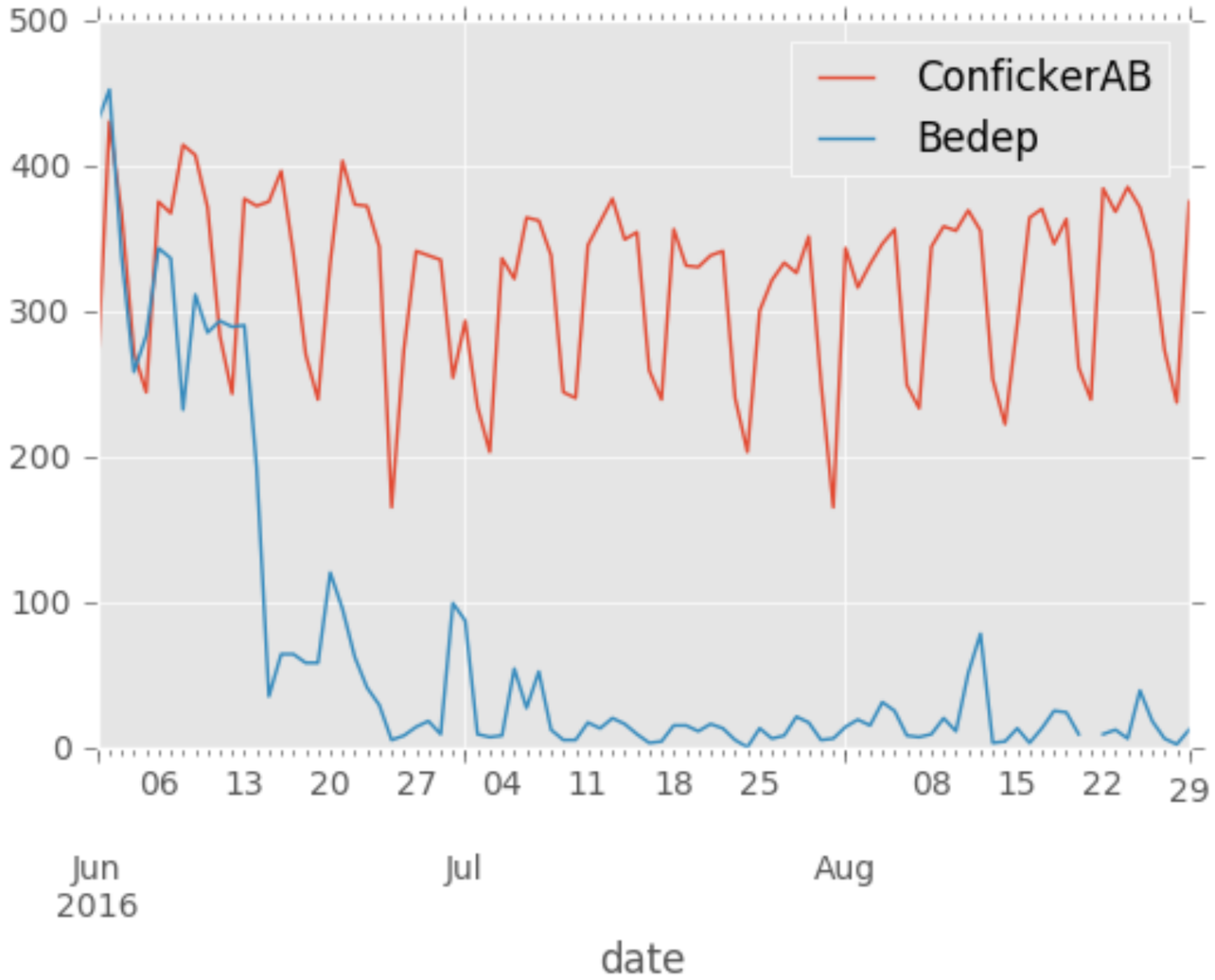
Exercise 1

```
data = pd.read_csv('../data/dailybots.csv')
filteredData = data[data['industry'] == “Government/Politics"]
filteredData2 = filteredData[filteredData['botfam']== 'ConfickerAB' ][['date','hosts']]
filteredData2.columns = ['date', 'ConfickerAB']
filteredData2.date = pd.to_datetime( filteredData2.date )
finalData.plot(kind='line' )
```



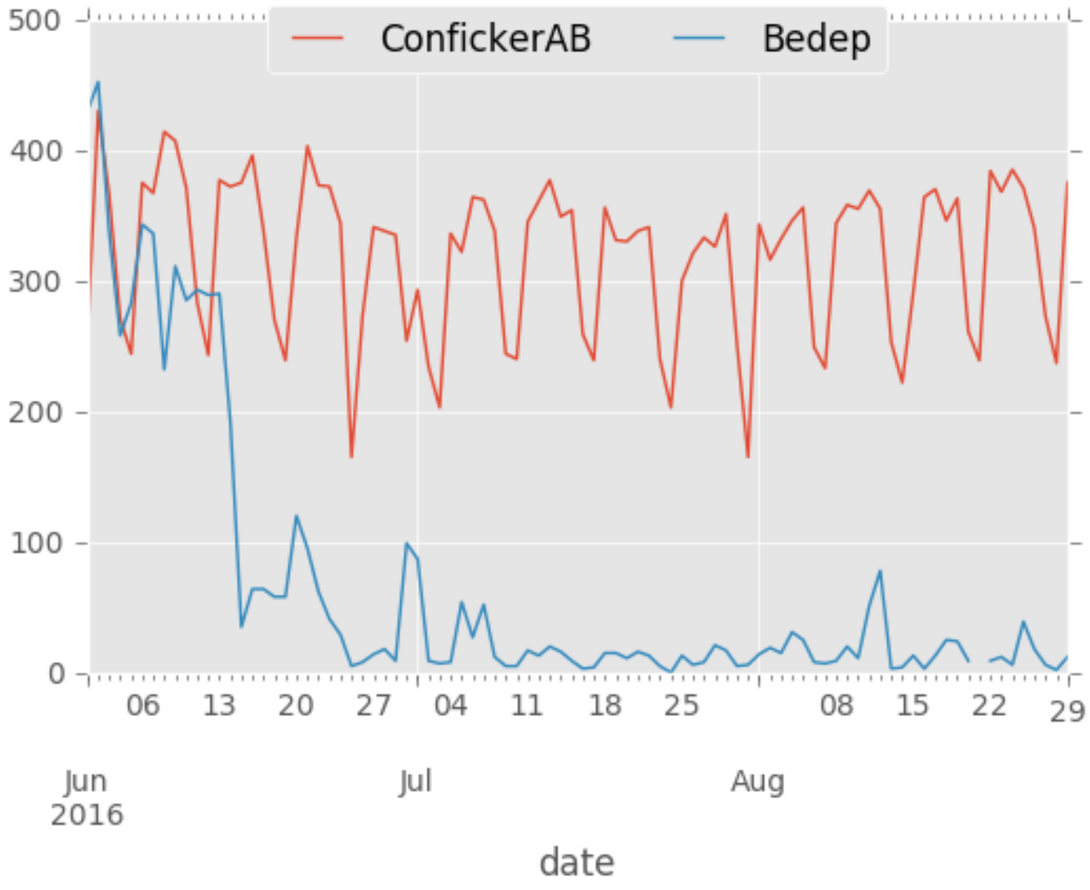
Exercise 1

```
finalData.set_index('date', inplace=True)
finalData.plot(kind='line' )
```



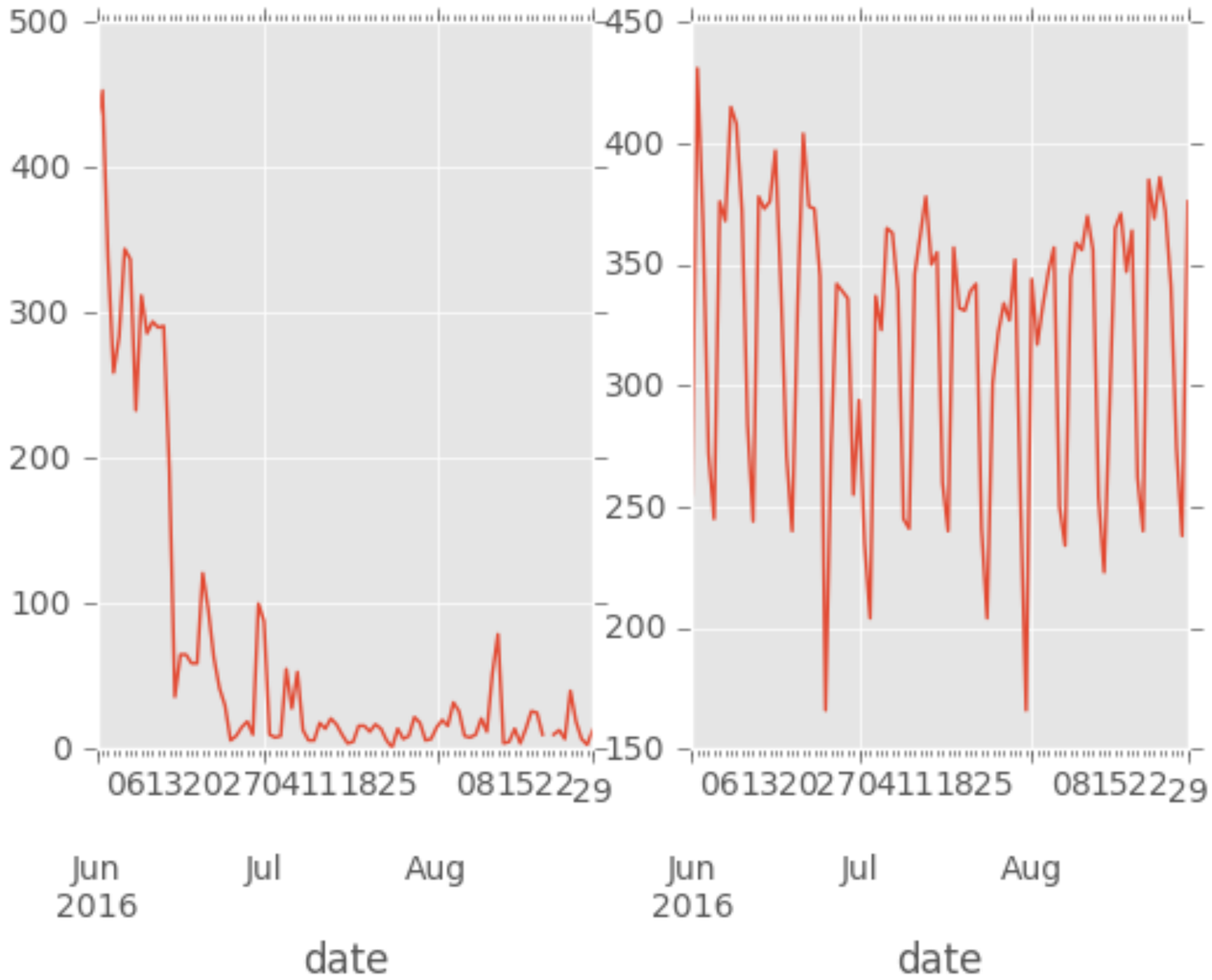
Exercise 1

```
nicePlot = finalData.plot( kind="line")
nicePlot.legend(loc='upper center', bbox_to_anchor=(0.5, 1.05),
          ncol=3, fancybox=True, shadow=False)
```



Exercise 1

```
fig, axes = plt.subplots(nrows=1, ncols=2)

finalData['ConfickerAB'].plot()
finalData['Bedep'].plot(ax=axes[0])
```



Exercise 2

```
from bokeh.plotting import output_notebook
output_notebook()
from bokeh.charts import TimeSeries
from bokeh.io import show

linechart = TimeSeries( data=finalData,
                        title="ConfickerAB Hosts",
                        legend="top_left" )
show( linechart )
```

Exercise 2

gtkcyber.com

Questions?