Chapter 01 Importing Data for Analysis

A Header

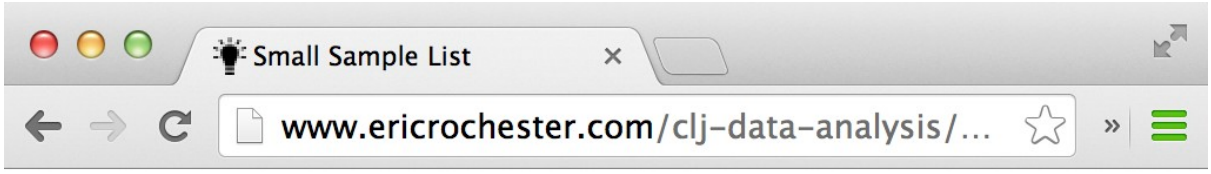
This contains a bunch of people.

Tonight we're going to code HTML like it's 1999!

Menu Here

Given Name	Surname	Relation
Gomez	Addams	father
Morticia	Addams	mother
Pugsley	Addams	brother
Wednesday	Addams	sister
Pubert	Addams	brother
Fester	Addams	uncle
Grandmama		grandmother
Thing		hand
Lurch		butler
Itt		cousin
Cackle		cousin

Something here. More links here.



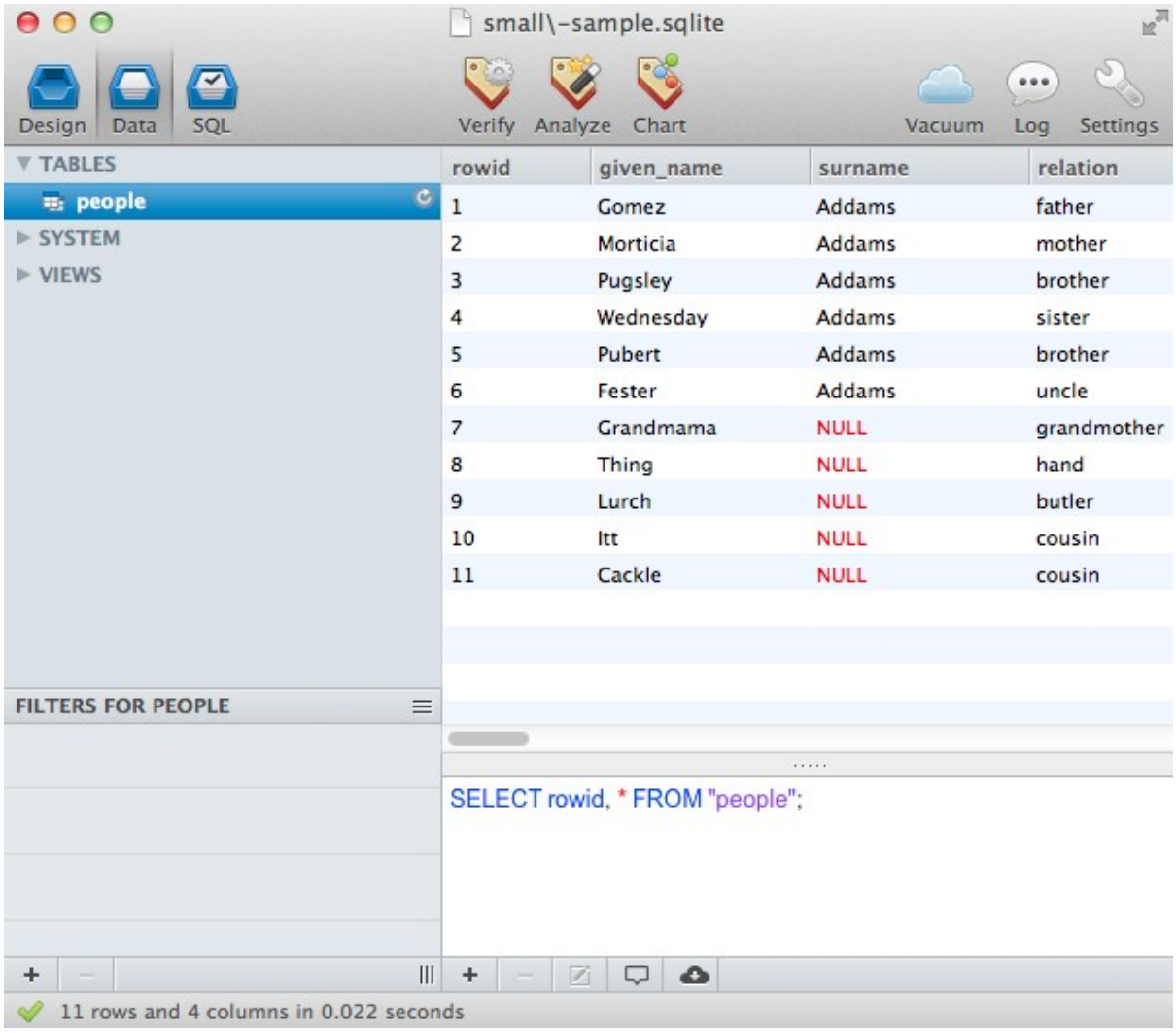
Small Sample List

- Section 1
- Section 2
- Section 3

Addam's Family

Here's some information about the Addam's Family.

- *Gomez Addams* — father
- *Morticia Addams* — mother
- *Pugsley Addams* — brother
- *Wednesday Addams* — sister
- *Pubert Addams* — brother
- *Fester Addams* — uncle
- *Grandmama* — grandmother
- *Thing* — hand
- *Lurch* — butler
- *Cousin Itt* — cousin
- *Cousin Cackle* — cousin




```
<td><em>Menu Here</em></td>
<td>
  <!-- Here's the data. -->
  <table id="data" border="0">
    <tr><th>Given Name</th> <th>Surname</th> <th>Relation</th></tr>
    <tr><td>Gomez</td> <td>Addams</td> <td>father</td></tr>
    <tr><td>Morticia</td> <td>Addams</td> <td>mother</td></tr>
    <tr><td>Pugsley</td> <td>Addams</td> <td>brother</td></tr>
    <tr><td>Wednesday</td> <td>Addams</td> <td>sister</td></tr>
    <tr><td>Pubert</td> <td>Addams</td> <td>brother</td></tr>
```

table [:table#data]

headers [:tr :th]

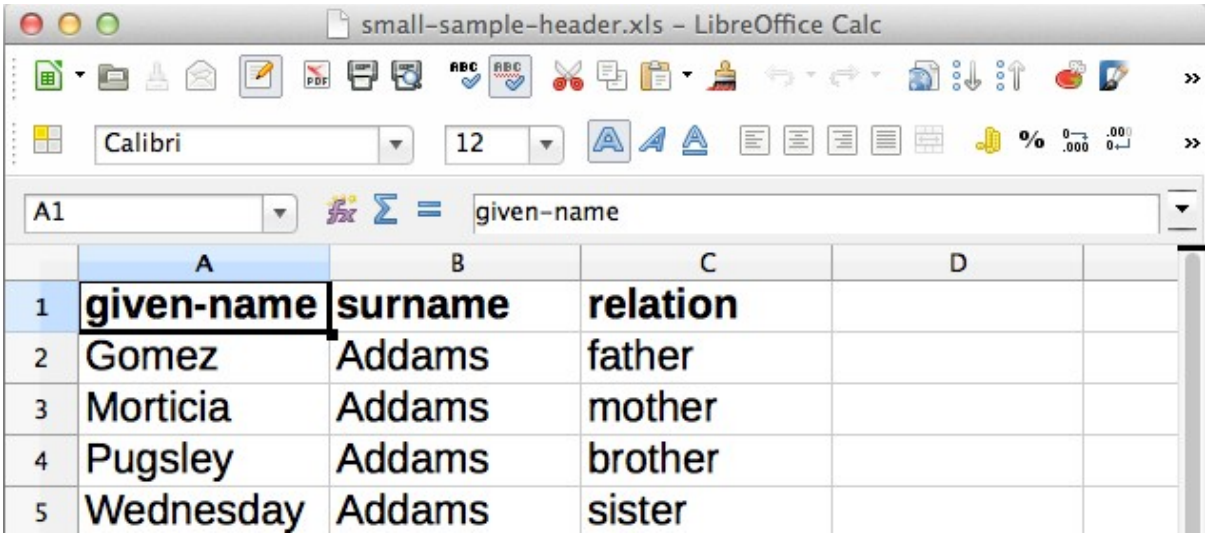
rows [:tr] > [:td]

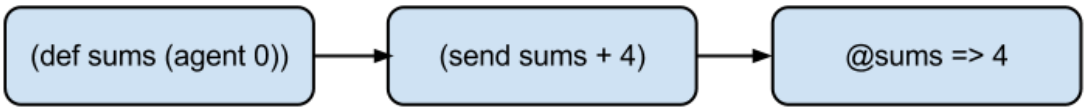
```
<article>
  <header>
    <h2 id='addams'>Addam's Family</h2>
  </header>
  <p>Here's some information about the Addam's Family.</p>
  <ul>
    <li><em>Gomez Addams</em> &mdash; father</li>
    <li><em>Morticia Addams</em> &mdash; mother</li>
    <li><em>Pugsley Addams</em> &mdash; brother</li>
    <li><em>Wednesday Addams</em> &mdash; sister</li>
    <li><em>Pubert Addams</em> &mdash; brother</li>
    <li><em>Fester Addams</em> &mdash; uncle</li>
    <li><em>Grandmama</em> &mdash; grandmother</li>
    <li><em>Thing</em> &mdash; hand</li>
    <li><em>Lurch</em> &mdash; butler</li>
    <li><em>Cousin Itt</em> &mdash; cousin</li>
    <li><em>Cousin Cackle</em> &mdash; cousin</li>
  </ul>
</article>
```

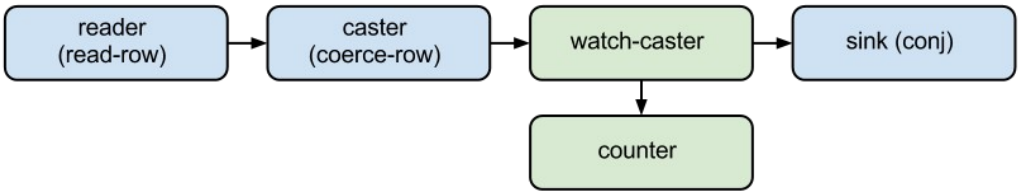
get-family

get-rows

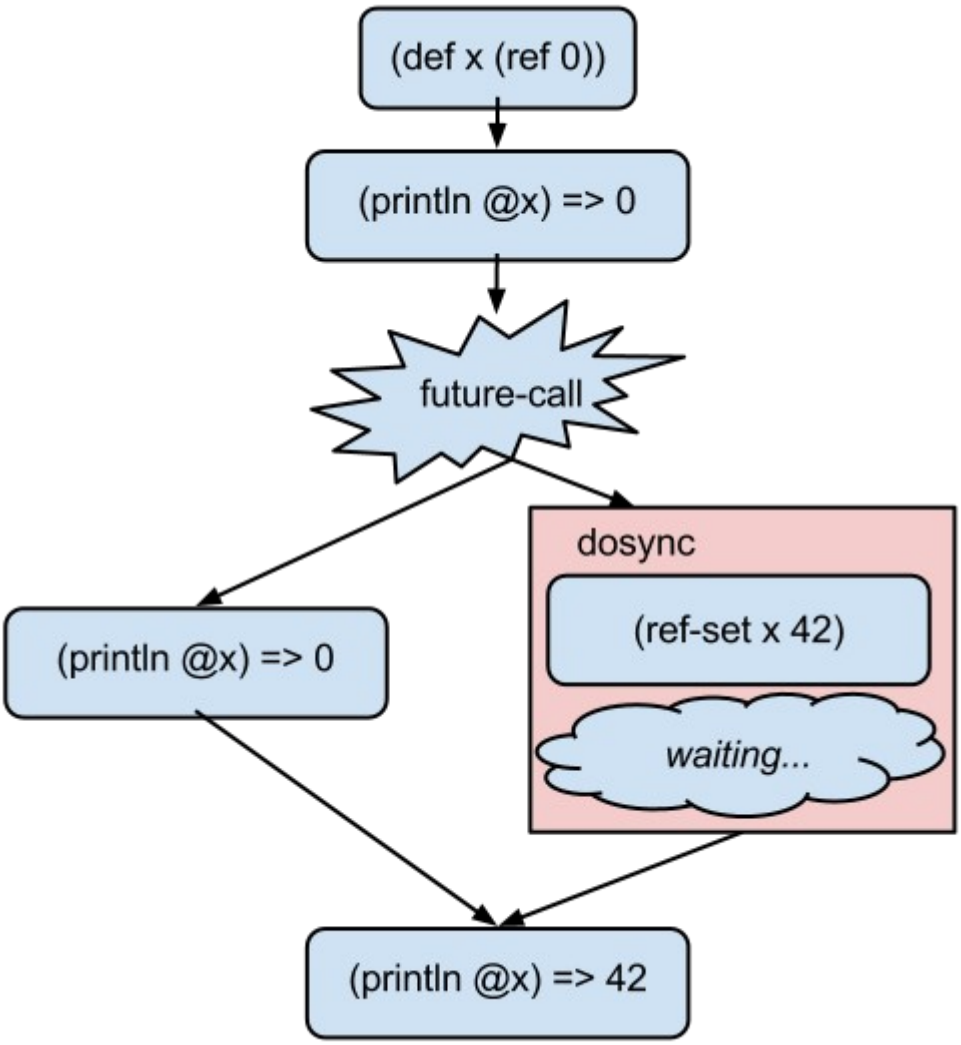
get-person

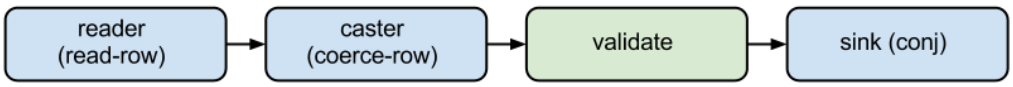


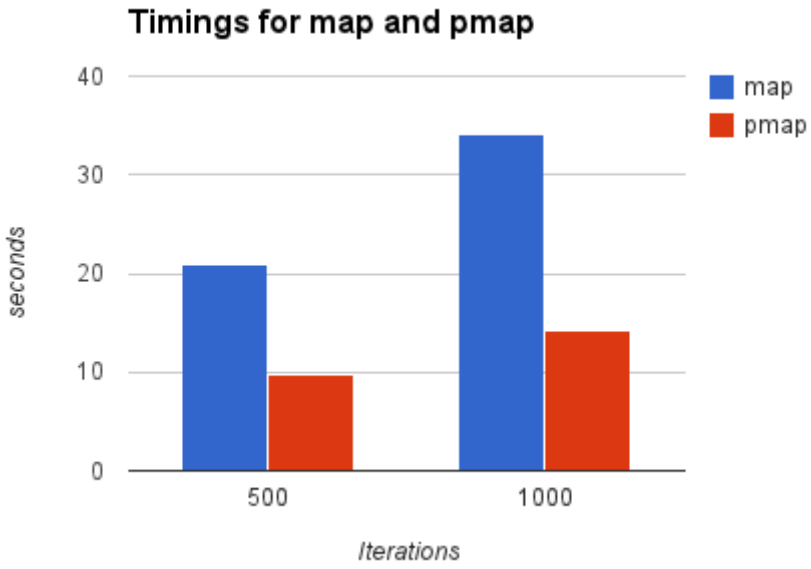


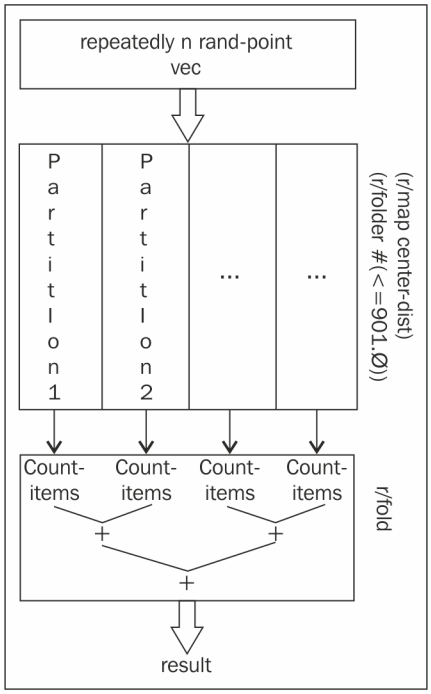


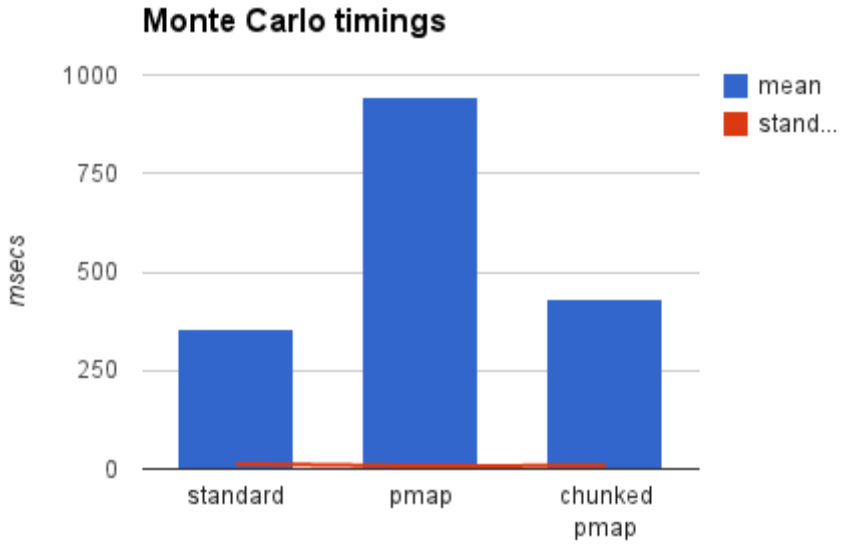
	A	B	C	D	E	F	G	H	I	J	K
1	GEOID	SUMLEV	STATE	COUNTY	CBSA	CSA	NECTA	CNECTA	NAME	POP100	HU100
2	5100148	160	51						Abingdon town	8191	4271
3	5100180	160	51						Accomac town	519	229
4	5100724	160	51						Alberta town	298	163
5	5101000	160	51						Alexandria city	139966	72376
6	5101256	160	51						Allisonia CDP	117	107

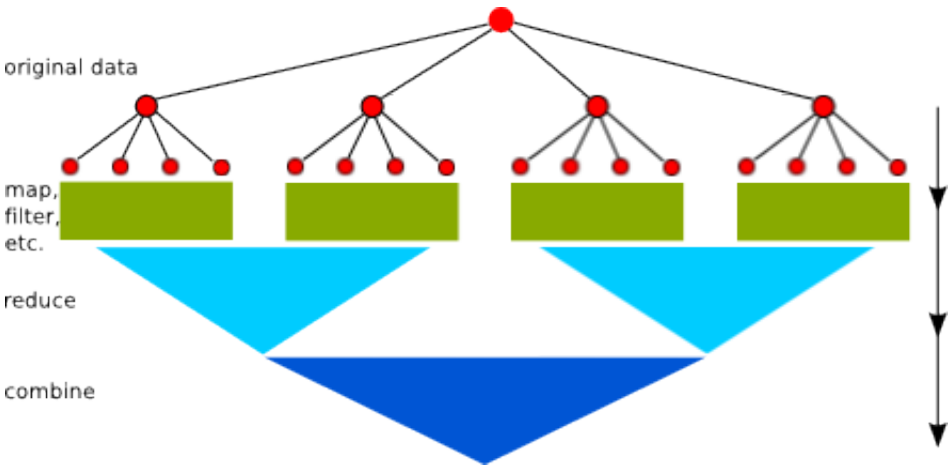

(def x (ref 0))
(println @x) => 0
future-call
dosync
(ref-set x 42)
waiting...
(println @x) => 0
(println @x) => 42

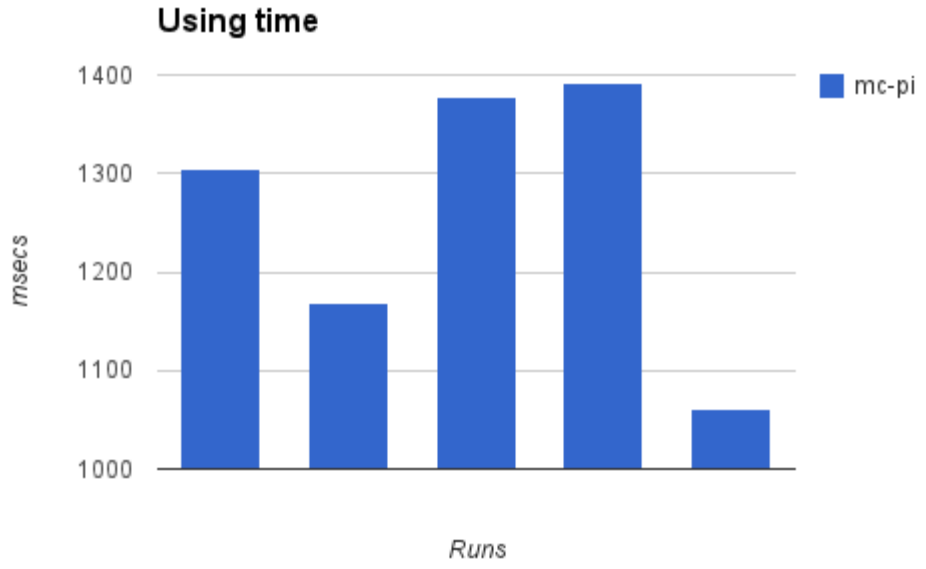

reader
(read-row)
caster
(coerce-row)
validate
sink (conj)

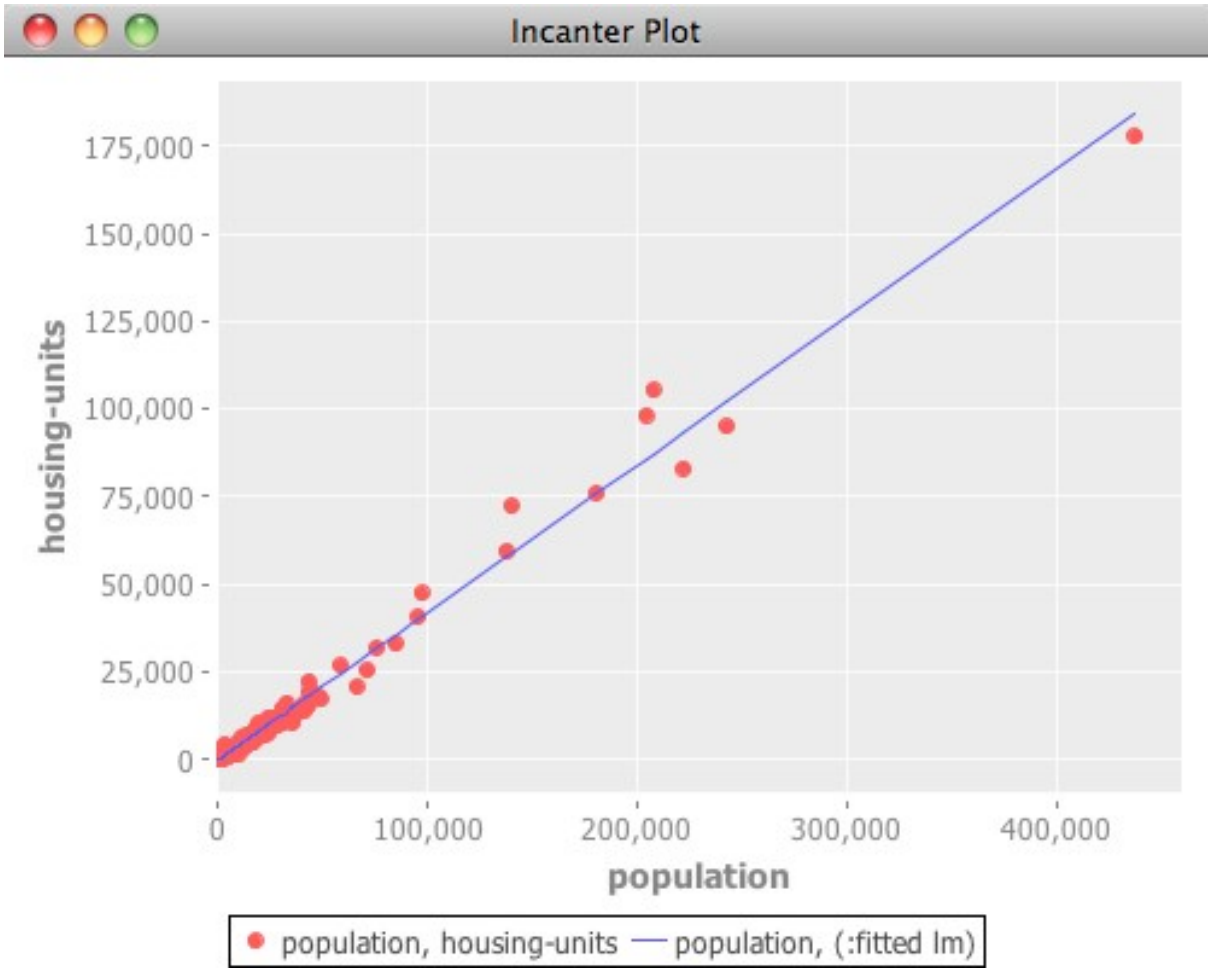

Timings for map and pmap
40
30
20
10
0
seconds
500
1000
Iterations
map
pmap


repeatedly n rand-point vec
Partition 1
Partition 2
...
...
(r/map center-dist)
(r/folder #(<=901.0))
Count-items
Count-items
Count-items
Count-items
+
+
+
r/fold
result

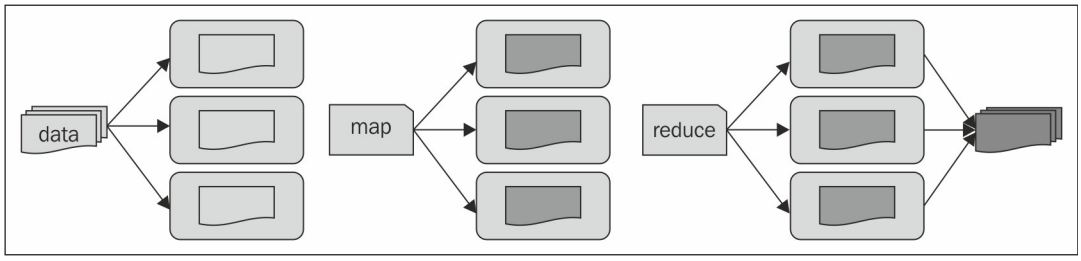

Monte Carlo timings
1000
750
500
250
0
msecs
mean
stand...
standard
pmap
chunked
pmap


original data
map,
filter,
etc.
reduce
combine


Using time
1400
1300
1200
1100
1000
msecs
Runs
mc-pi


Incanter Plot
175,000
150,000
125,000
100,000
75,000
50,000
25,000
0
housing-units
0
100,000
200,000
300,000
400,000
population
population, housing-units
population, (:fitted lm)

Chapter 5, *Distributed Data Processing with Cascalog*

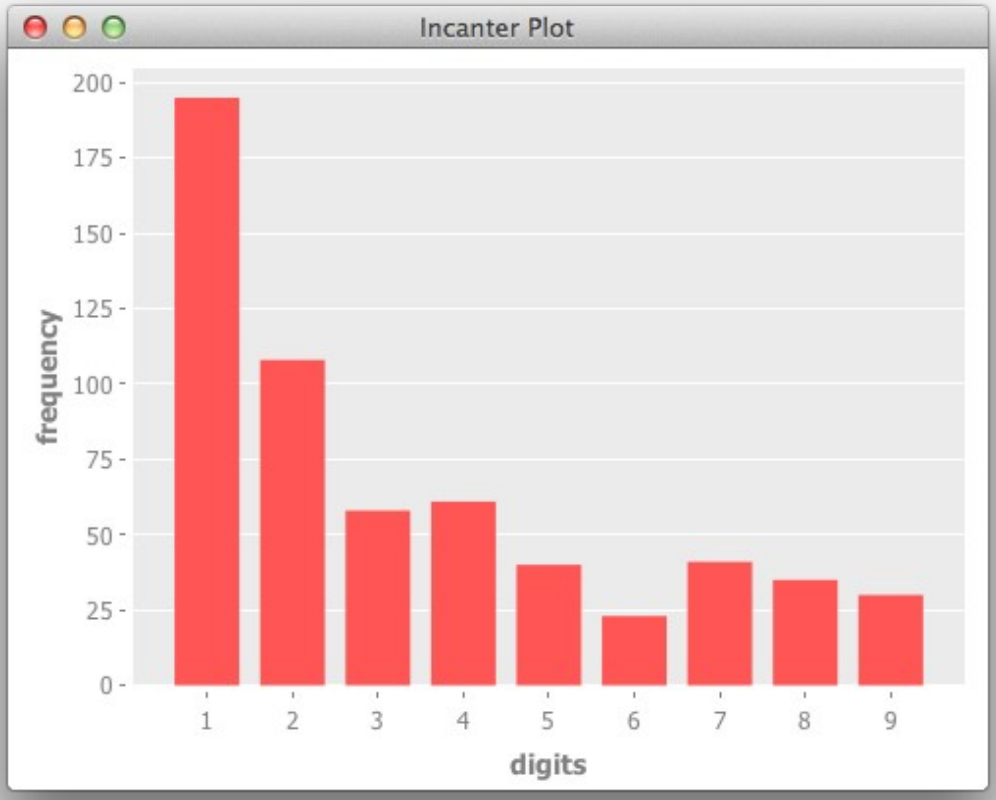


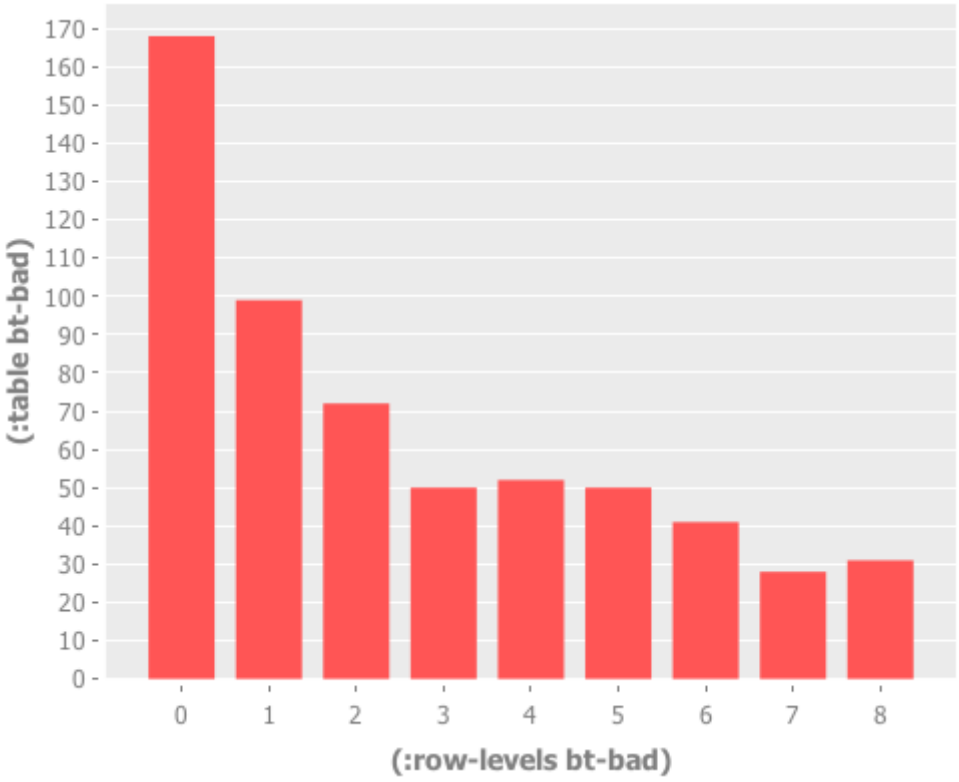
Chapter 6, *Working with Incanter Datasets*

Incanter Dataset

:Sepal.Leng...	:Sepal.Width	:Petal.Length	:Petal.Width	:Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa

Chapter 7, *Statistical Data Analysis with Incanter*


Incanter Plot
200
175
150
125
100
75
50
25
0
frequency
1
2
3
4
5
6
7
8
9
digits




```
(i/$rollup :mean :POP100 :STATE census)
```

Input dataset `census`

:STATE	:NAME	:POP100
6	Ashland CDP	21925
6	San Bernardino city	209924
6	Trinidad city	367
36	Smallwood CDP	580
36	North Babylon CDP	17509
36	Pleasant Valley CDP	1145
36	Upper Nyack village	2063
51	Crosspointe CDP	5802
51	Flint Hill CDP	209

Group by `:STATE`

6

:STATE	:NAME	:POP100
6	Ashland CDP	21925
6	San Bernardino city	209924
6	Trinidad city	367

36

:STATE	:NAME	:POP100
36	Smallwood CDP	580
36	North Babylon CDP	17509
36	Pleasant Valley CDP	1145
36	Upper Nyack village	2063

51

:STATE	:NAME	:POP100
51	Crosspointe CDP	5802
51	Flint Hill CDP	209

Extract `:POP100`

6

:POP100
21925
209924
367

36

:POP100
580
17509
1145
2063

51

:POP100
5802
209

Rollup with `:mean`

6

:mean
232216/3

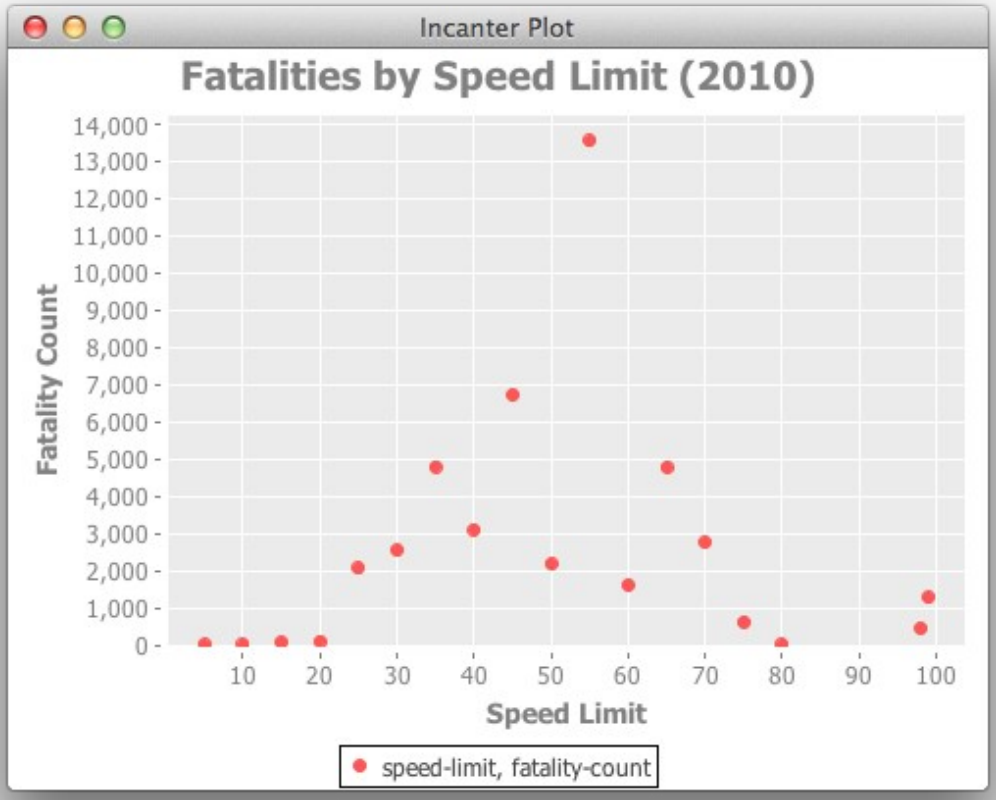
36

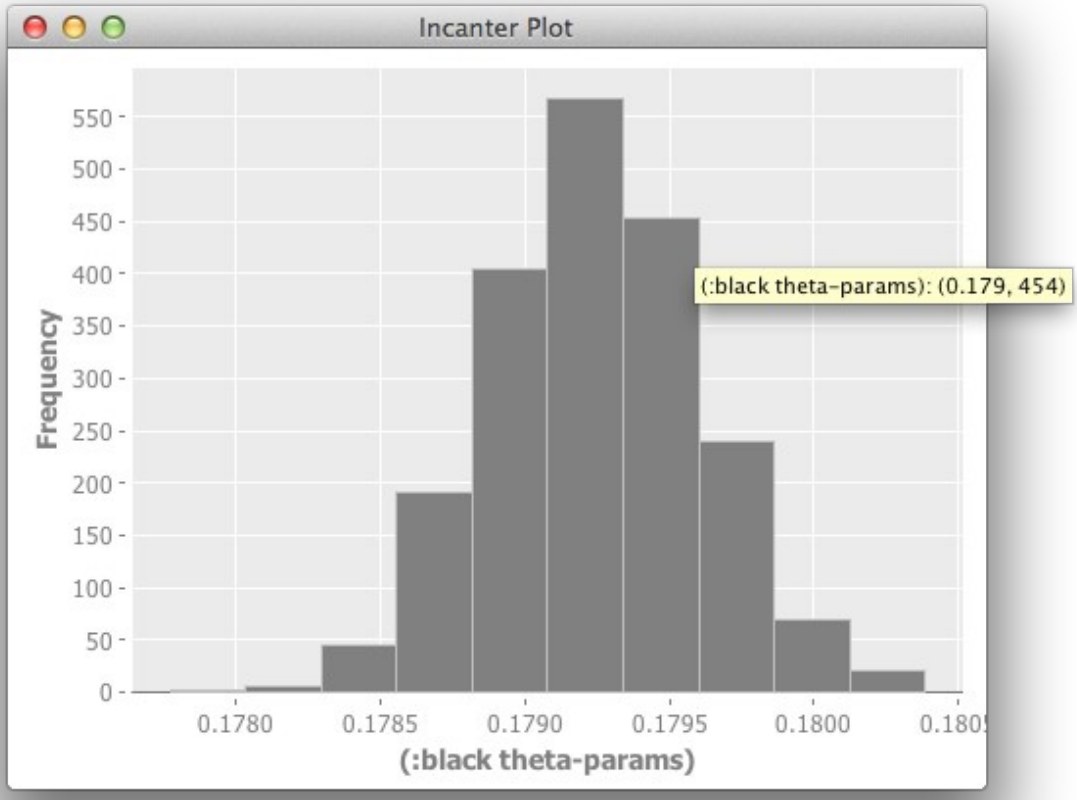
:mean
21297/4

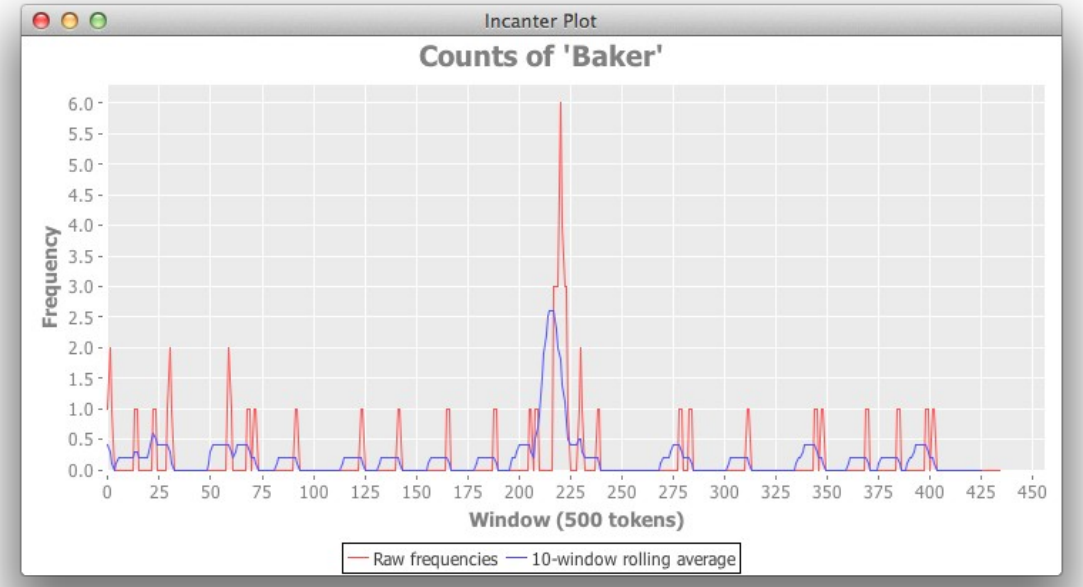
51

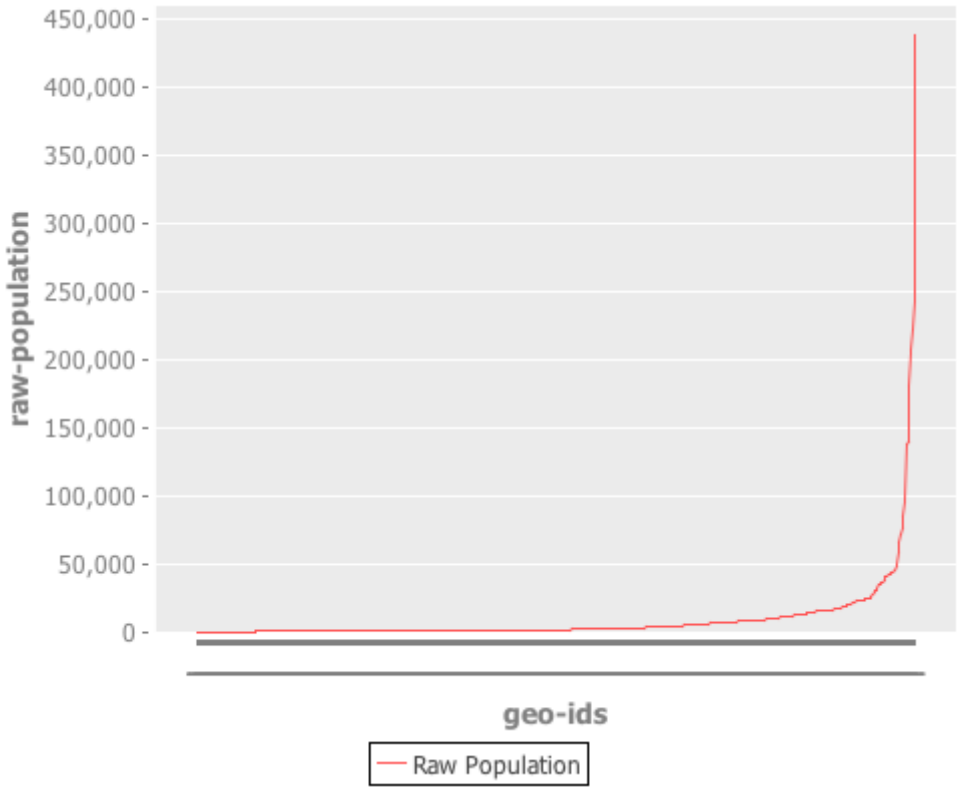
:mean
6011/2

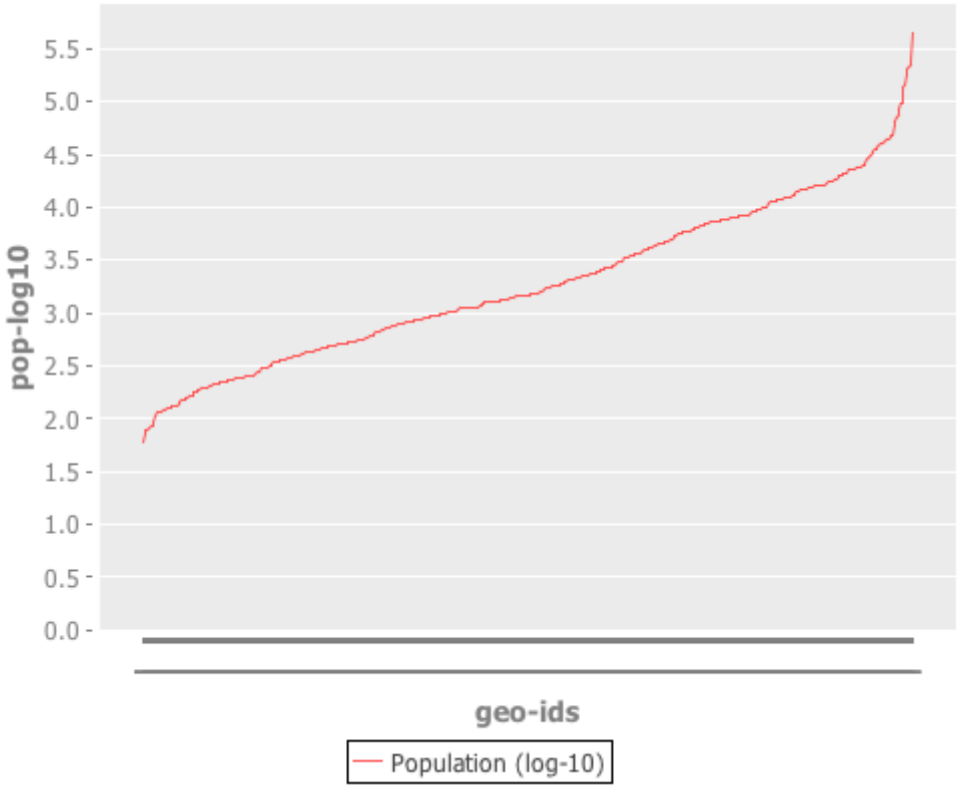
Incanter Plot
NYSE:IBM (Zoo)
220
215
210
205
200
195
190
185
180
175
170
Closing
Date
date, closing
5-day Running Mean
30-day Running Mean

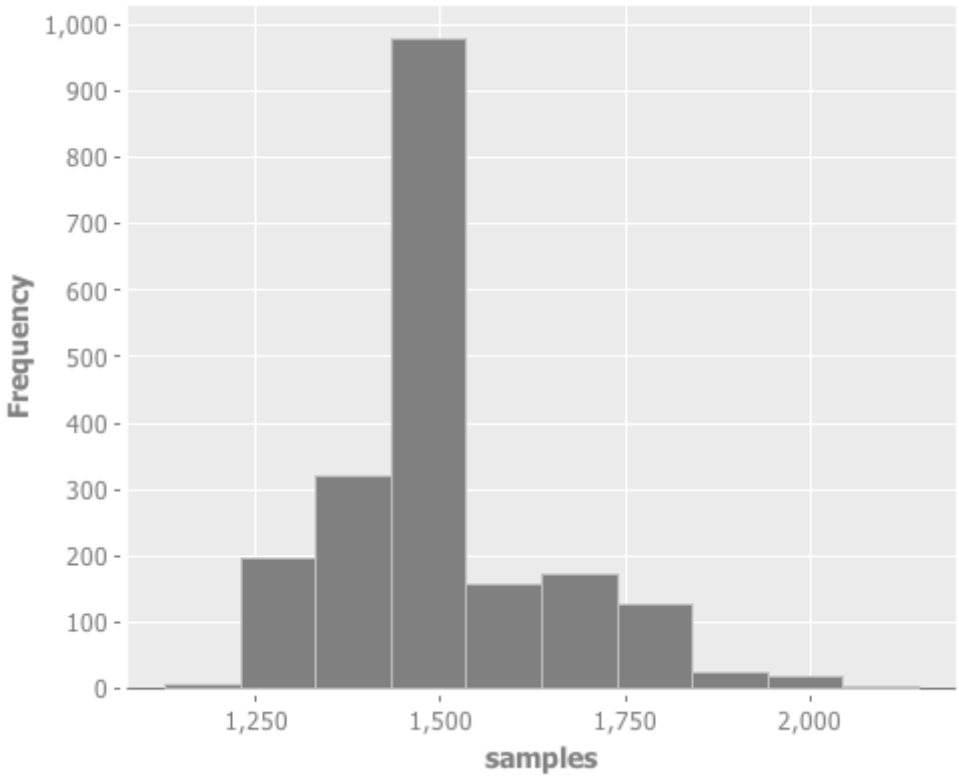

Incanter Plot
Fatalities by Speed Limit (2010)
Fatality Count
14,000
13,000
12,000
11,000
10,000
9,000
8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
10
20
30
40
50
60
70
80
90
100
Speed Limit
speed-limit, fatality-count

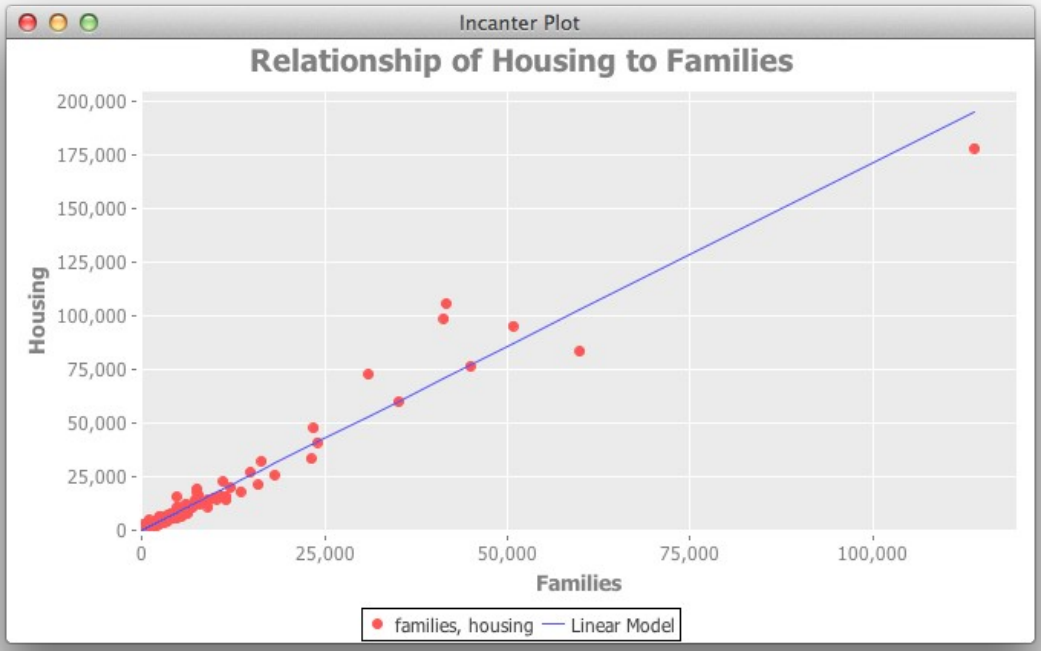

Incanter Plot
Frequency
550
500
450
400
350
300
250
200
150
100
50
0
0.1780
0.1785
0.1790
0.1795
0.1800
(:black theta-params)
(:black theta-params): (0.179, 454)

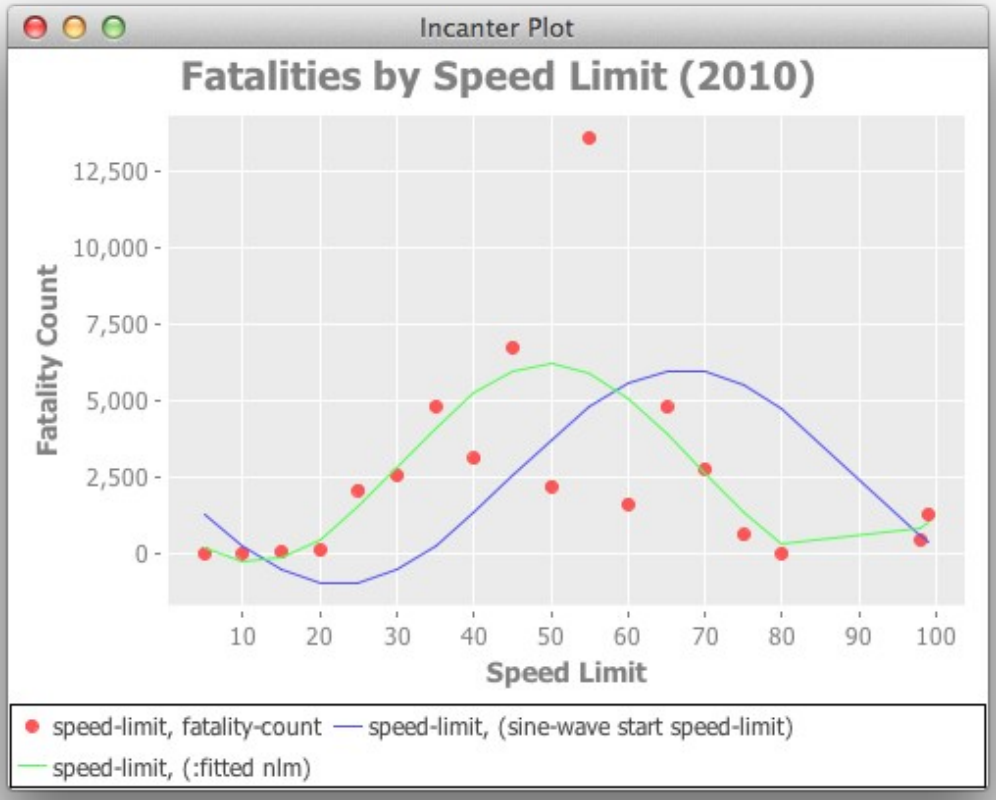

Incanter Plot
Counts of 'Baker'
Frequency
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
25
50
75
100
125
150
175
200
225
250
275
300
325
350
375
400
425
450
Window (500 tokens)
Raw frequencies
10-window rolling average


450,000
400,000
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
raw-population
geo-ids
Raw Population

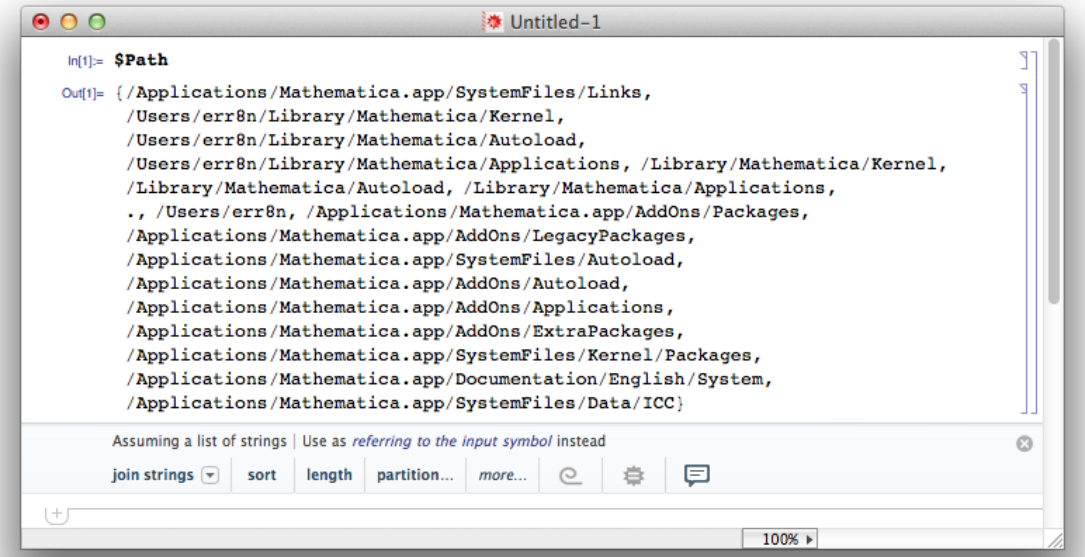

5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
pop-log10
geo-ids
Population (log-10)

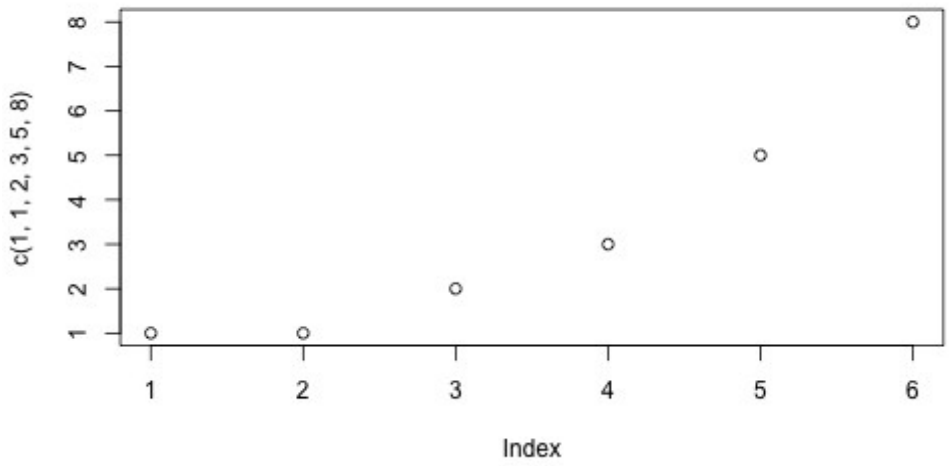

1,000
900
800
700
600
500
400
300
200
100
0
Frequency
1,250
1,500
1,750
2,000
samples

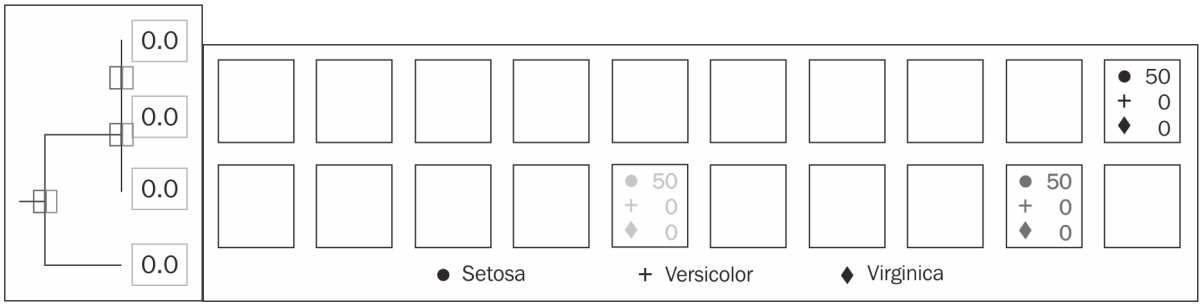

Incanter Plot
Relationship of Housing to Families
200,000
175,000
150,000
125,000
100,000
75,000
50,000
25,000
0
Housing
0
25,000
50,000
75,000
100,000
Families
families, housing
Linear Model

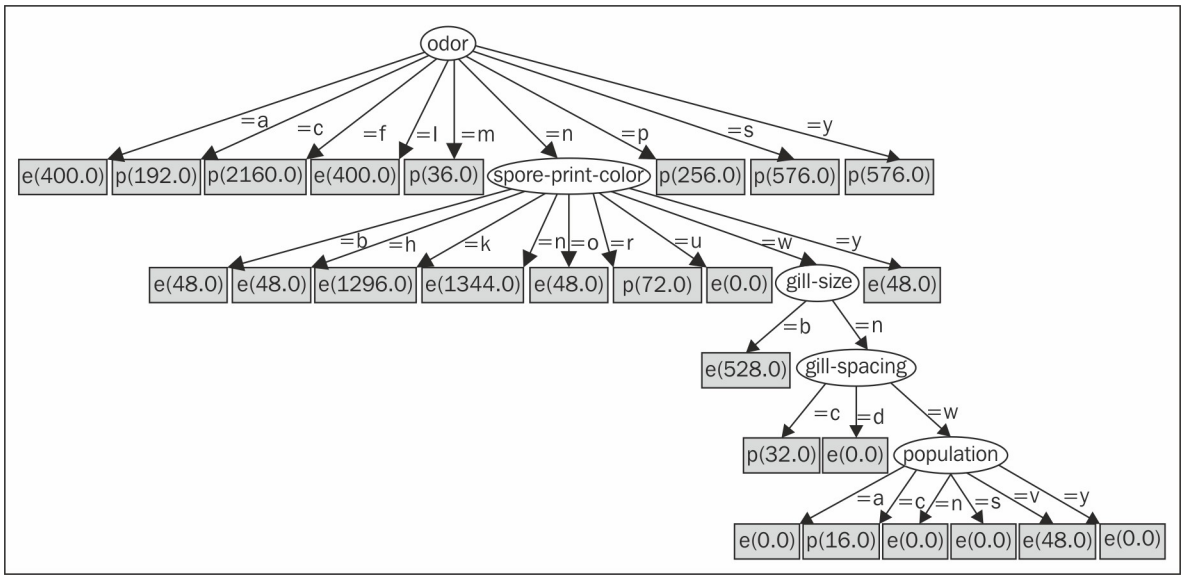


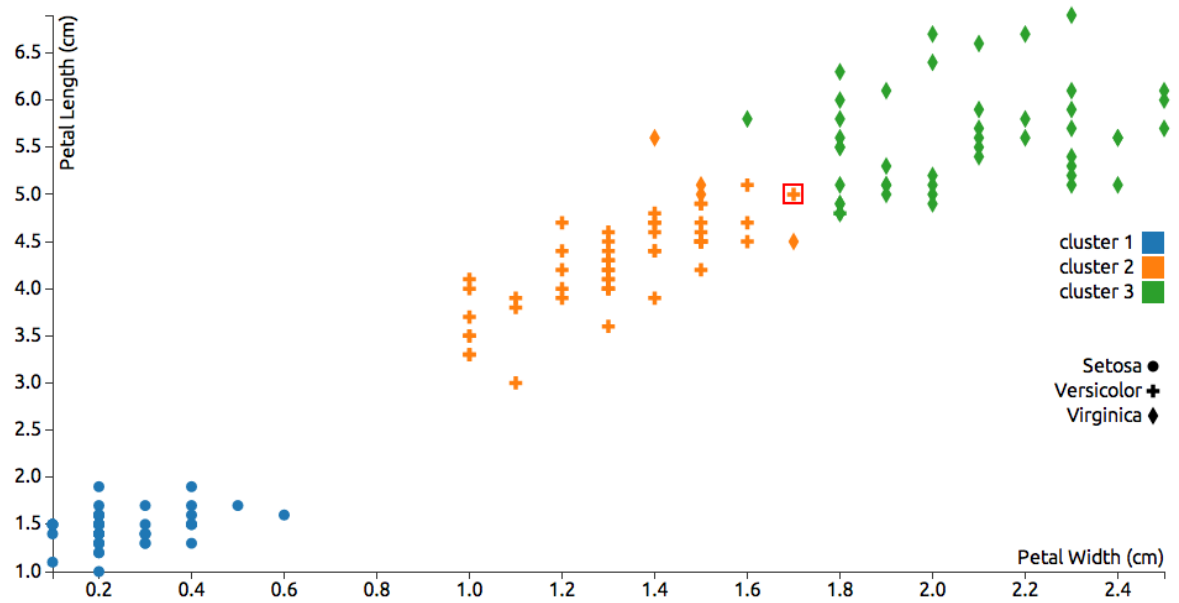
Chapter 8, *Working with Mathematica and R*,

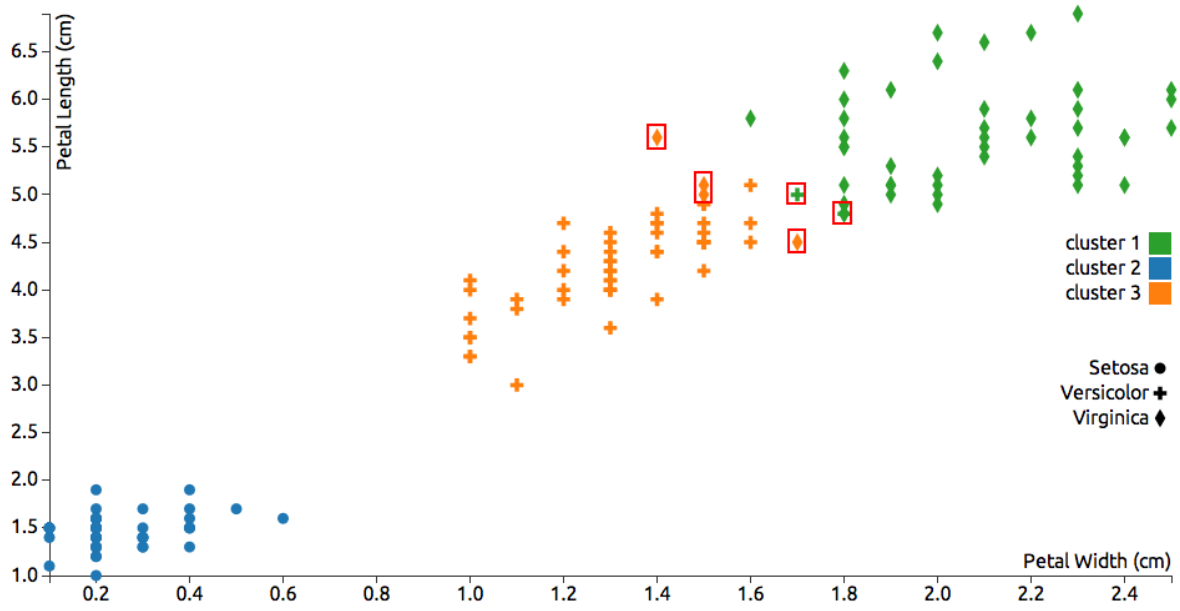

Untitled-1
In[1]:= $Path
Out[1]= {/Applications/Mathematica.app/SystemFiles/Links,
/Users/err8n/Library/Mathematica/Kernel,
/Users/err8n/Library/Mathematica/Autoload,
/Users/err8n/Library/Mathematica/Applications, /Library/Mathematica/Kernel,
/Library/Mathematica/Autoload, /Library/Mathematica/Applications,
., /Users/err8n, /Applications/Mathematica.app/AddOns/Packages,
/Applications/Mathematica.app/AddOns/LegacyPackages,
/Applications/Mathematica.app/SystemFiles/Autoload,
/Applications/Mathematica.app/AddOns/Autoload,
/Applications/Mathematica.app/AddOns/Applications,
/Applications/Mathematica.app/AddOns/ExtraPackages,
/Applications/Mathematica.app/SystemFiles/Kernel/Packages,
/Applications/Mathematica.app/Documentation/English/System,
/Applications/Mathematica.app/SystemFiles/Data/ICC}
Assuming a list of strings | Use as referring to the input symbol instead
join strings
sort
length
partition...
more...
100%


c(1, 1, 2, 3, 5, 8)
1
2
3
4
5
6
7
8
1
2
3
4
5
6
Index


0.0
0.0
0.0
0.0
50
0
0
50
0
0
50
0
0
Setosa
Versicolor
Virginica


odor
=a
=c
=f
=l
=m
=n
=p
=s
=y
e(400.0)
p(192.0)
p(2160.0)
e(400.0)
p(36.0)
spore-print-color
p(256.0)
p(576.0)
p(576.0)
=b
=h
=k
=n
=o
=r
=u
=w
=y
e(48.0)
e(48.0)
e(1296.0)
e(1344.0)
e(48.0)
p(72.0)
e(0.0)
gill-size
e(48.0)
=b
=n
e(528.0)
gill-spacing
=c
=d
=w
p(32.0)
e(0.0)
population
=a
=c
=n
=s
=v
=y
e(0.0)
p(16.0)
e(0.0)
e(0.0)
e(48.0)
e(0.0)

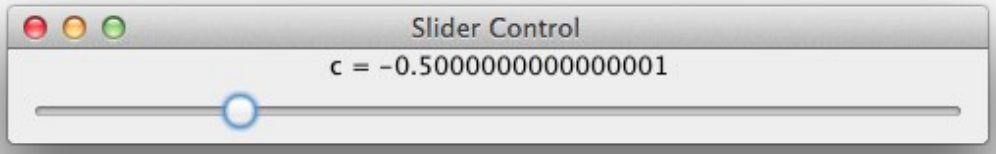


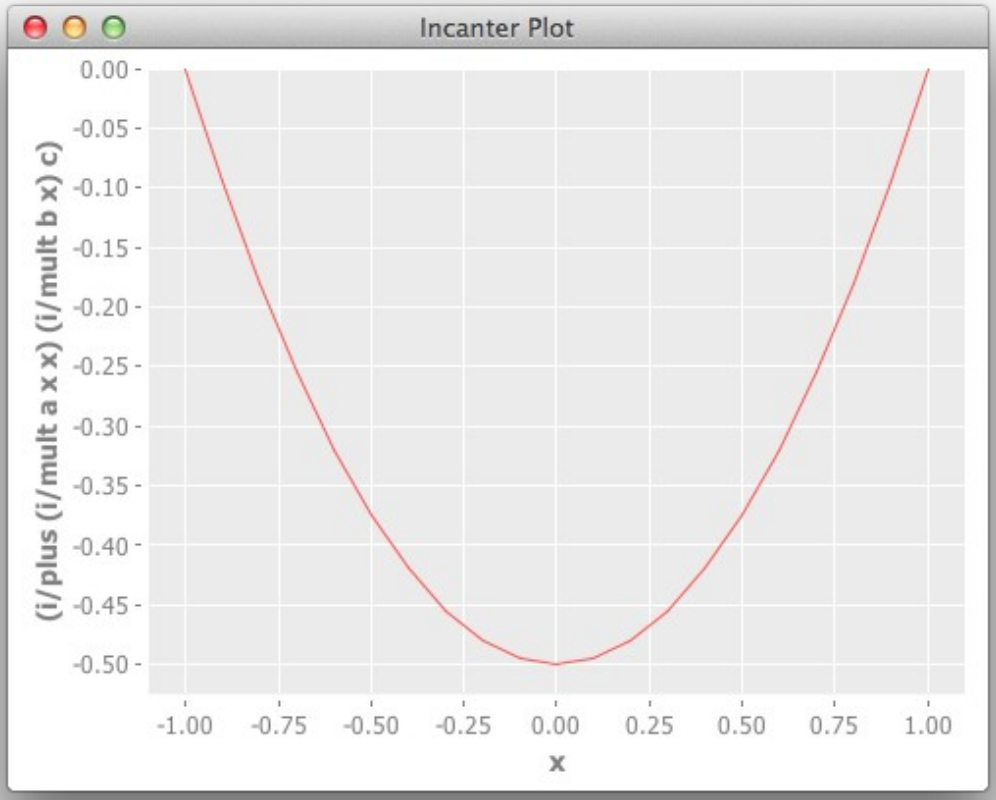


$$tf(t,d) = 0.5 + \frac{0.5 \times f(t,d)}{\max\{f(w,d) : \mathrm{w} \in \mathrm{d}\}}$$

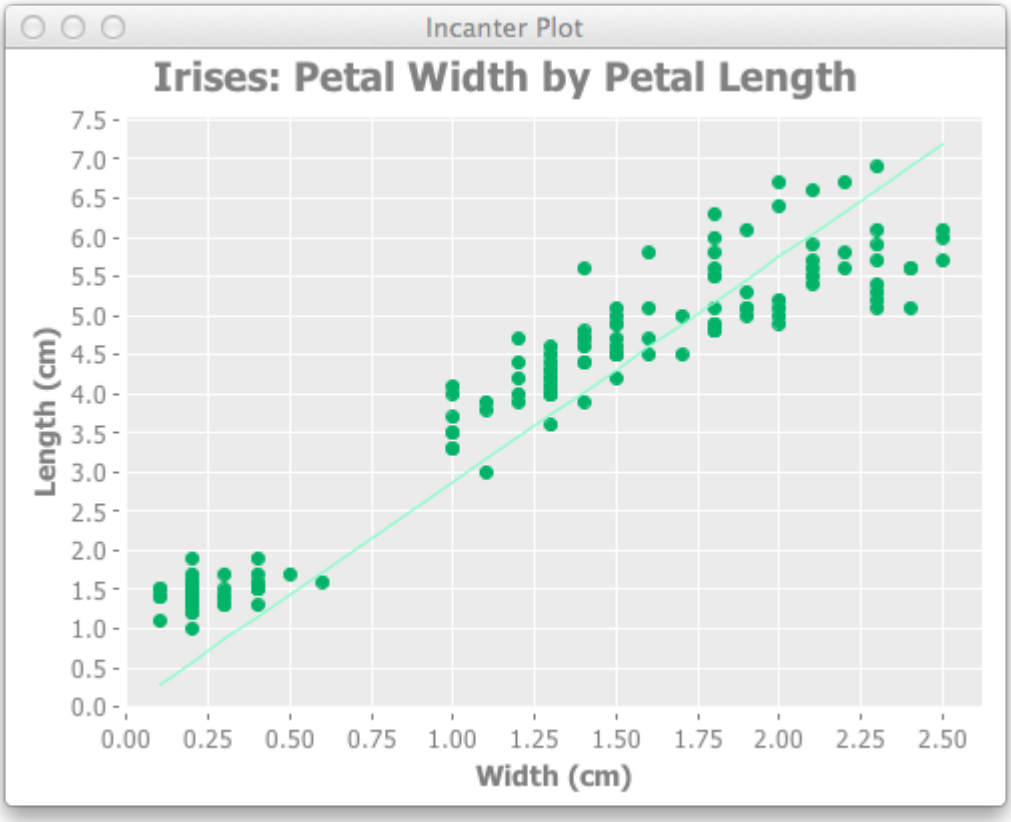
$$idf(t,D) = \log \frac{N}{\left|\{d \in D : t \in d\}\right|}$$

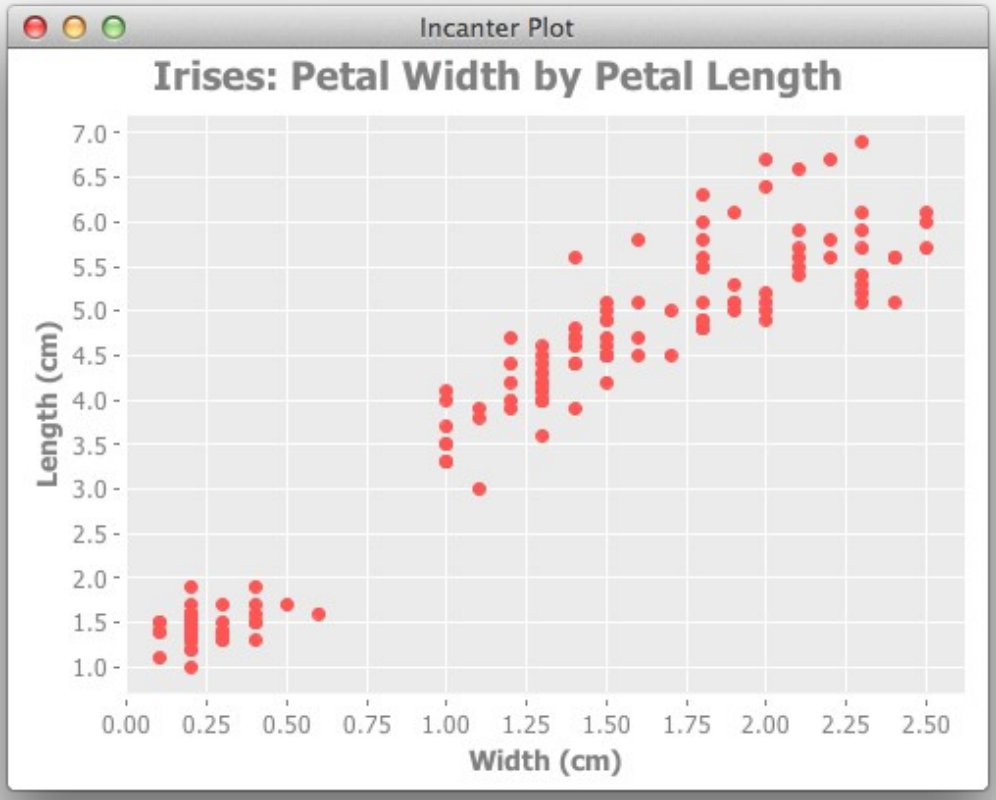
$$tfidf(t,d,D) = tf(t,d) \times idf(t,D)$$

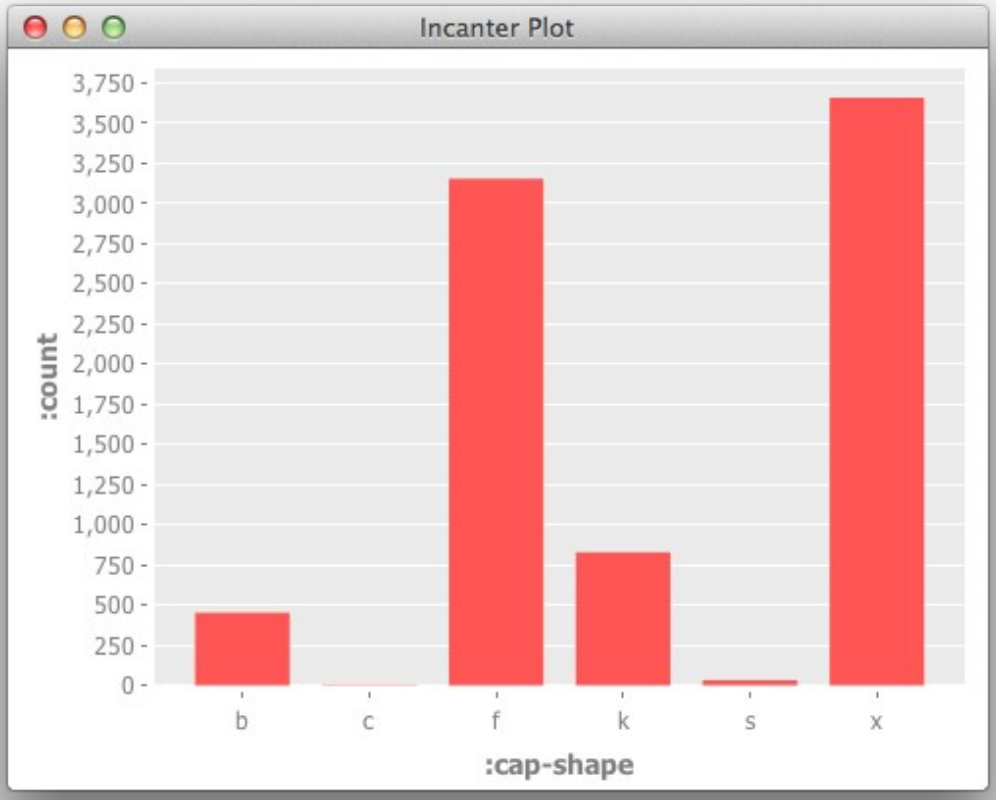

Slider Control
c = -0.5000000000000001

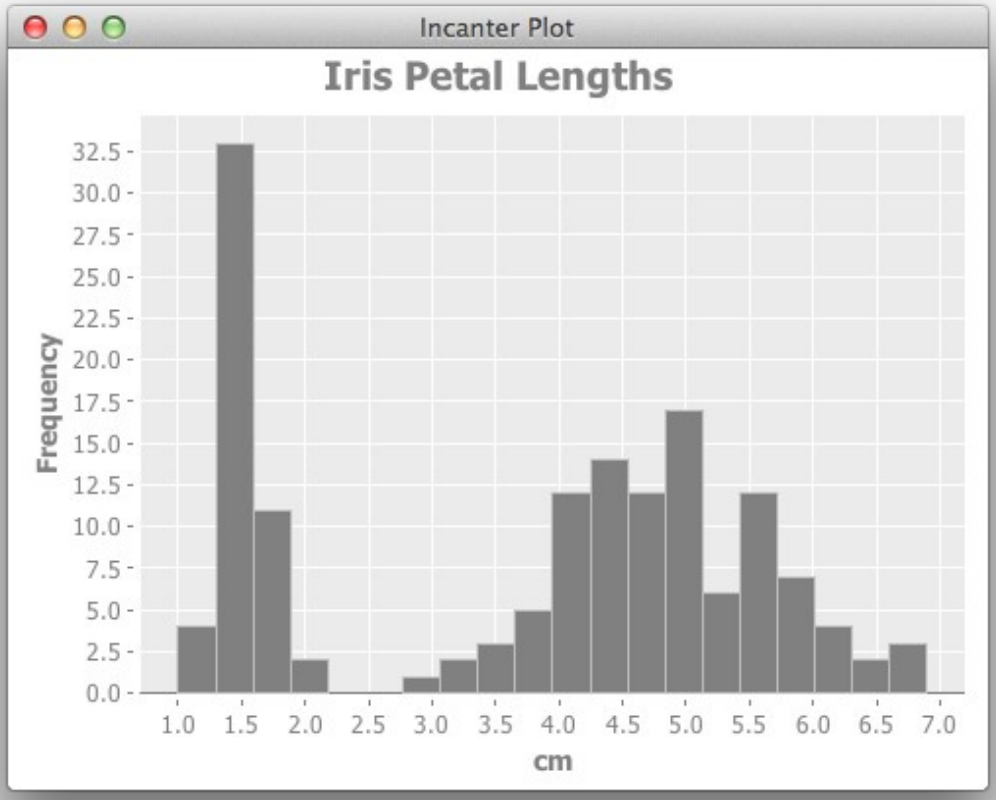

Incanter Plot
(i/plus (i/mult a x x) (i/mult b x) c)
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
-0.30
-0.35
-0.40
-0.45
-0.50
-1.00
-0.75
-0.50
-0.25
0.00
0.25
0.50
0.75
1.00
x

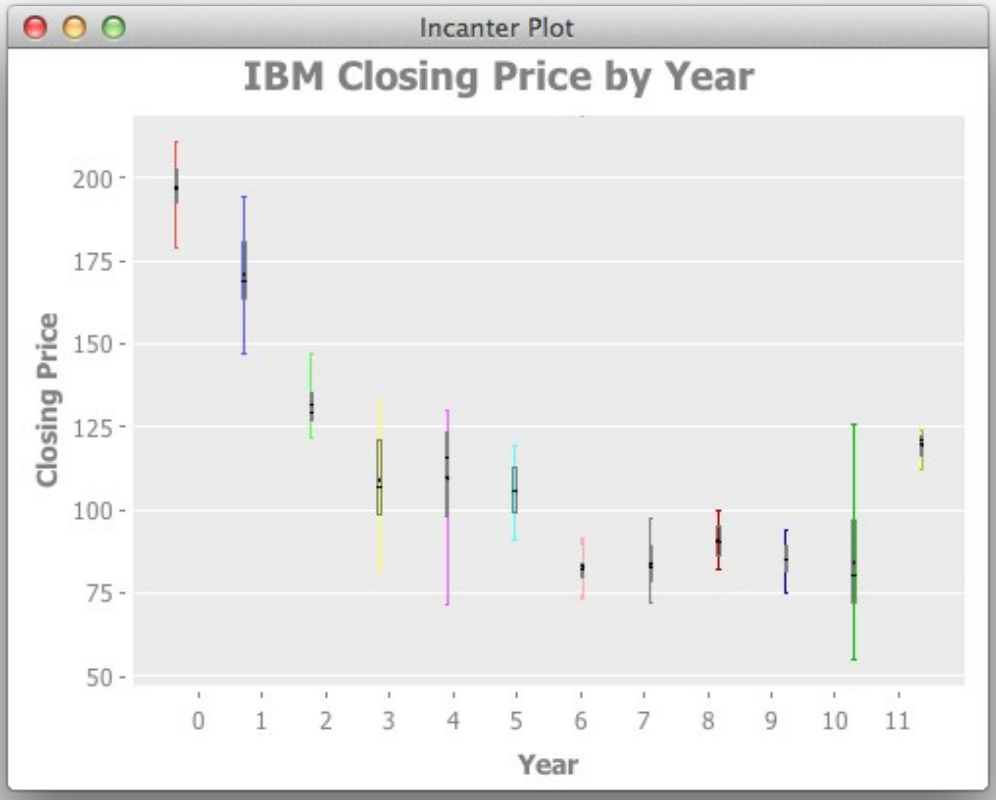
Incanter Plot
Iris' Dimensions
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
cm
setosa
versicolor
virginica
Species
Petal Width
Sepal Width
Petal Length
Sepal Length

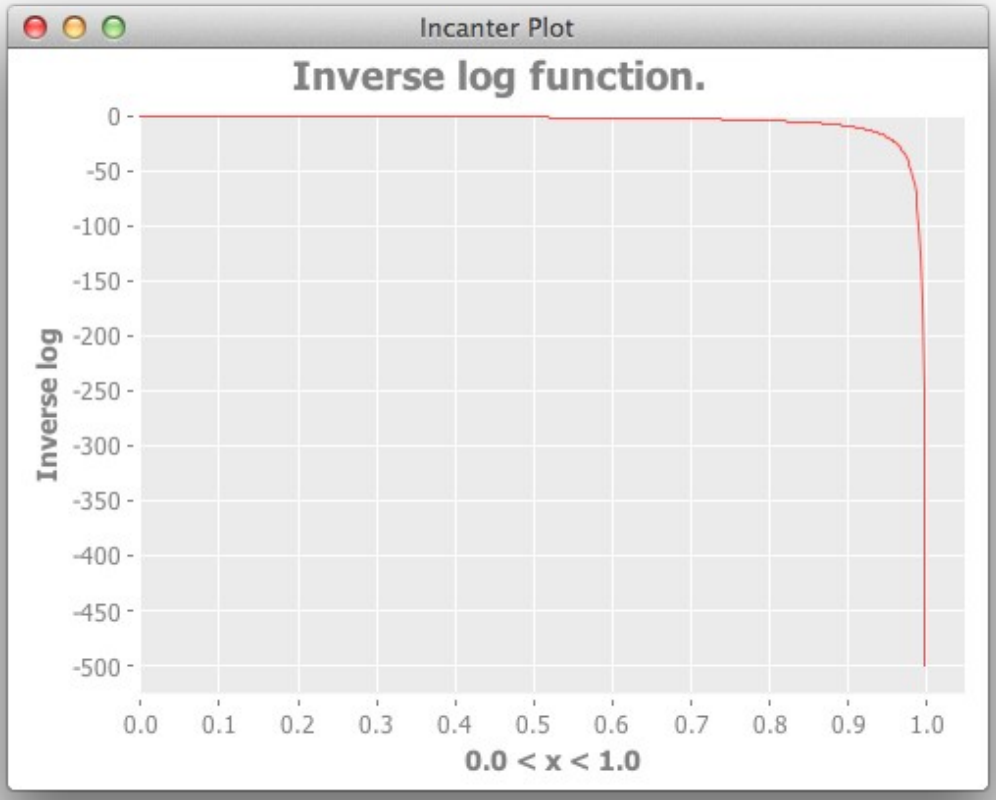

Incanter Plot
Irises: Petal Width by Petal Length
Length (cm)
Width (cm)
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
2.25
2.50

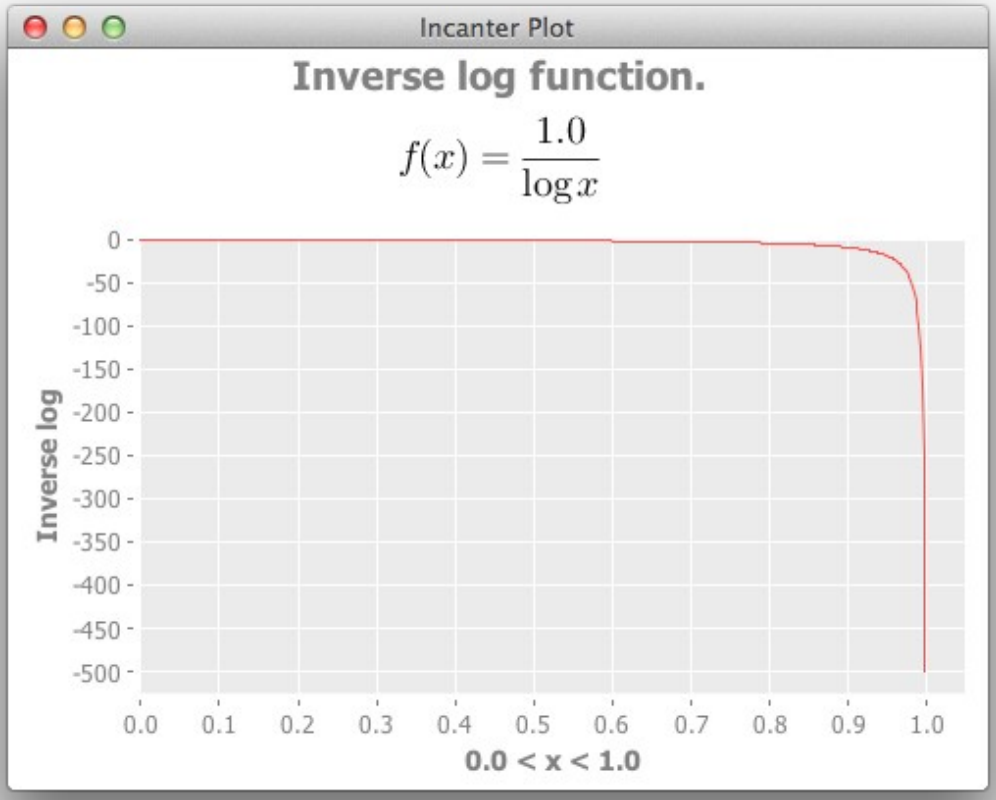

Incanter Plot
Irises: Petal Width by Petal Length
Length (cm)
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
2.25
2.50
Width (cm)

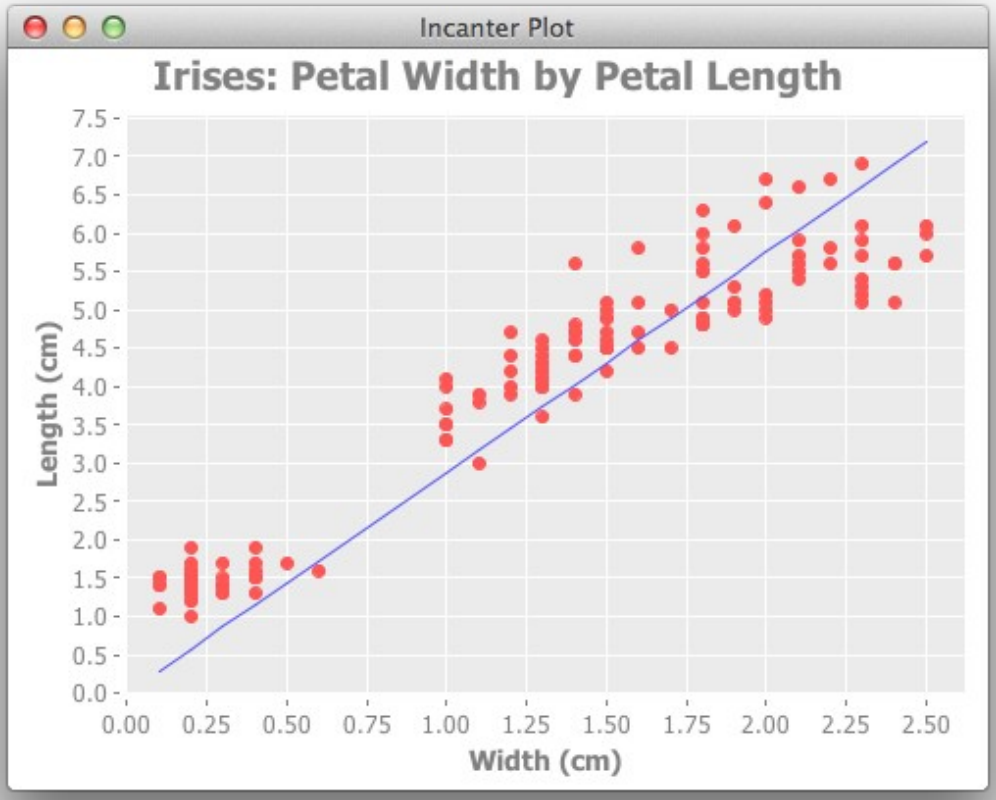

Incanter Plot
3,750
3,500
3,250
3,000
2,750
2,500
2,250
2,000
1,750
1,500
1,250
1,000
750
500
250
0
:count
b
c
f
k
s
x
:cap-shape

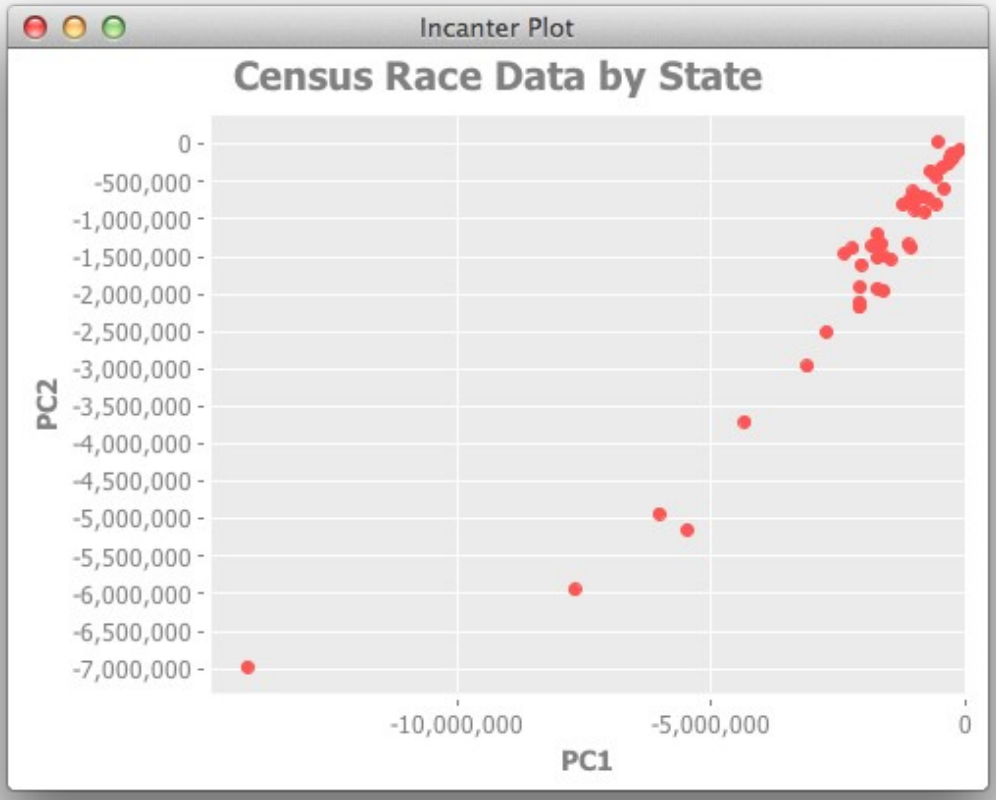

Incanter Plot
Iris Petal Lengths
Frequency
32.5
30.0
27.5
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
cm

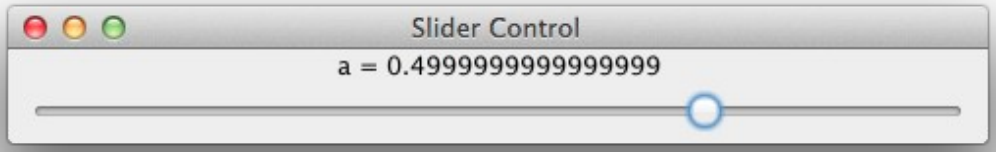

Incanter Plot
IBM Closing Price by Year
Closing Price
200
175
150
125
100
75
50
0
1
2
3
4
5
6
7
8
9
10
11
Year

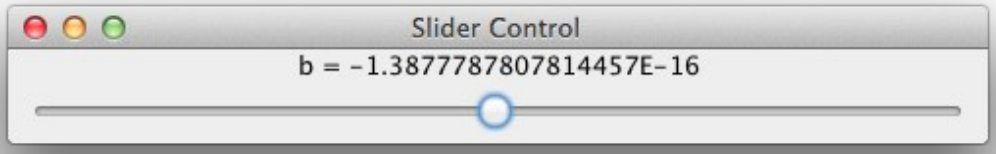

Incanter Plot
Inverse log function.
Inverse log
0
-50
-100
-150
-200
-250
-300
-350
-400
-450
-500
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0.0 < x < 1.0

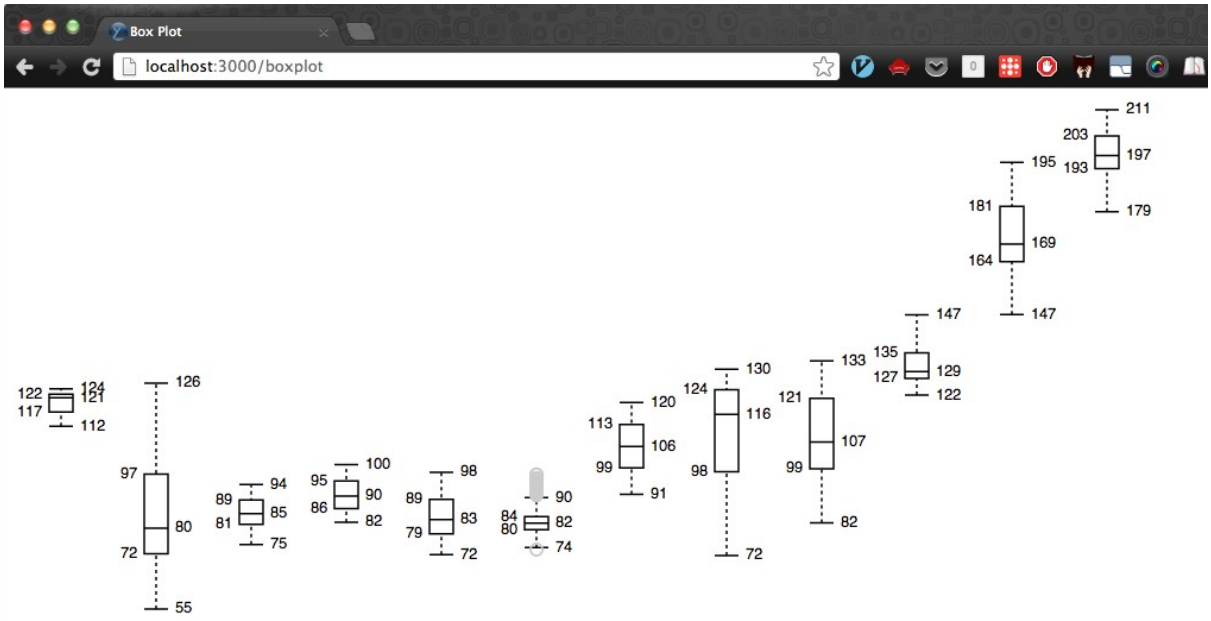

Incanter Plot
Inverse log function.
$f(x) = \frac{1.0}{\log x}$
Inverse log
0
-50
-100
-150
-200
-250
-300
-350
-400
-450
-500
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0.0 < x < 1.0

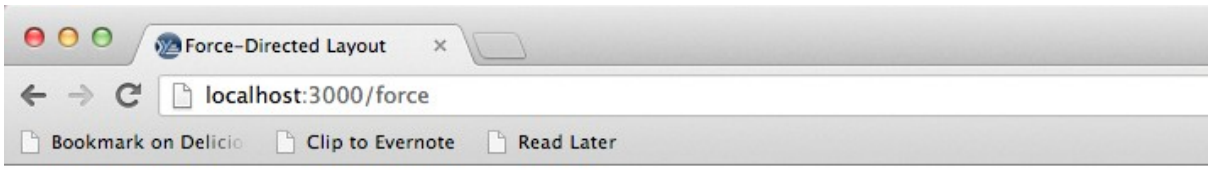

Incanter Plot
Irises: Petal Width by Petal Length
Length (cm)
Width (cm)
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
2.25
2.50

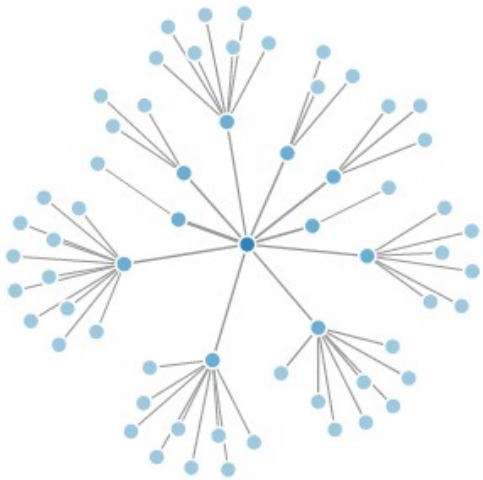

Incanter Plot
Census Race Data by State
0
-500,000
-1,000,000
-1,500,000
-2,000,000
-2,500,000
-3,000,000
-3,500,000
-4,000,000
-4,500,000
-5,000,000
-5,500,000
-6,000,000
-6,500,000
-7,000,000
PC2
-10,000,000
-5,000,000
0
PC1

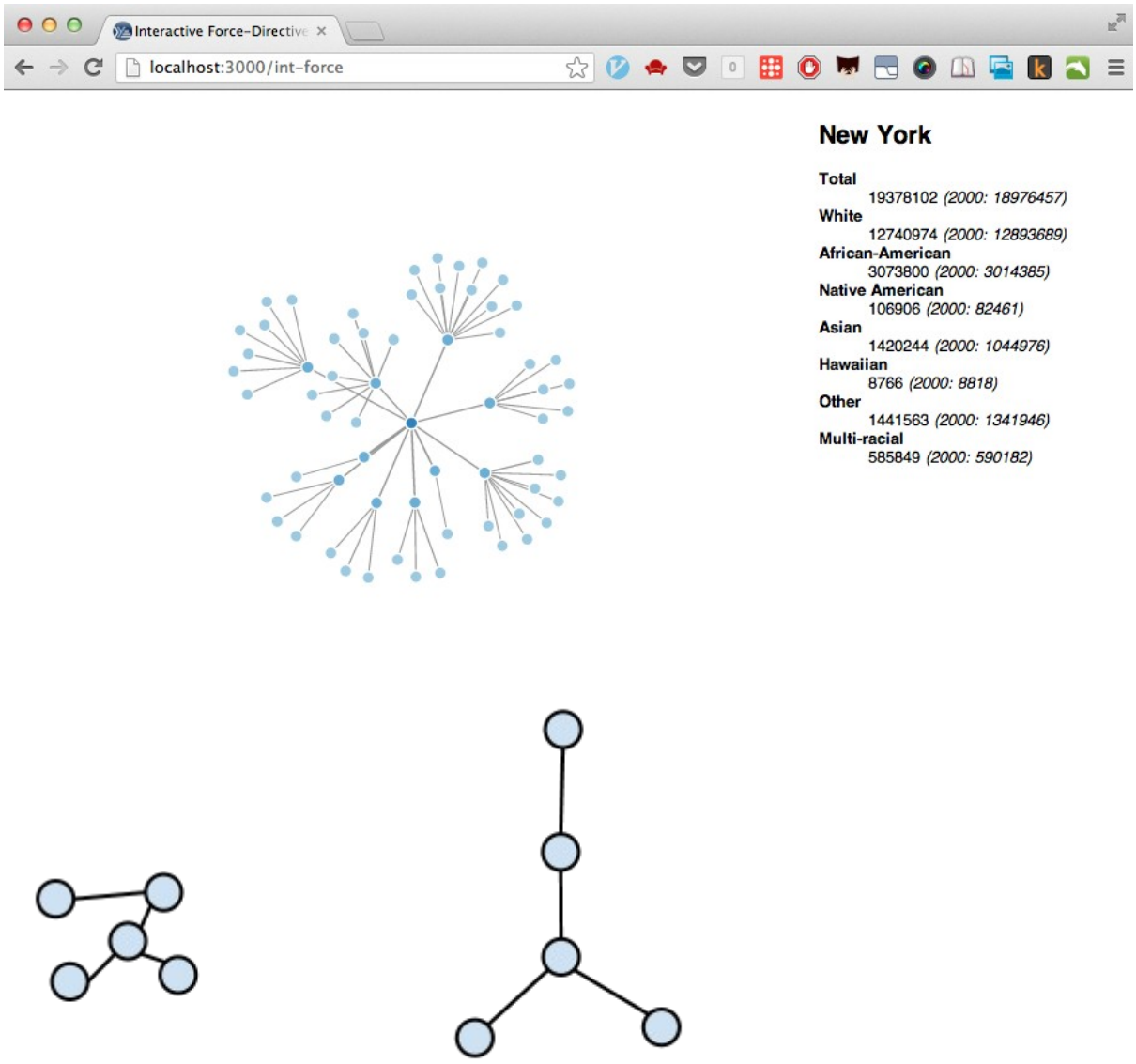

Slider Control
a = 0.4999999999999999

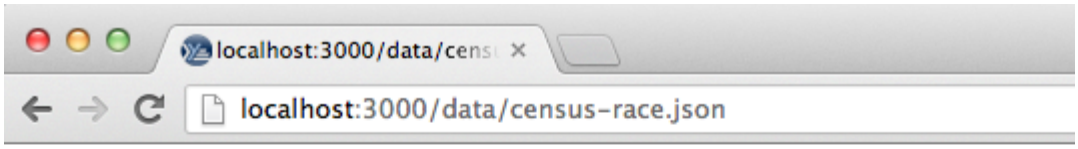

Slider Control
b = -1.3877787807814457E-16


Box Plot
localhost:3000/boxplot
211
203
197
193
195
179
181
169
164
147
147
135
129
127
122
147
133
121
107
99
82
130
124
116
98
72
120
113
106
99
91
90
84
82
80
74
98
89
83
79
72
100
95
90
86
82
94
89
85
81
75
126
97
80
72
55
124
122
121
117
112

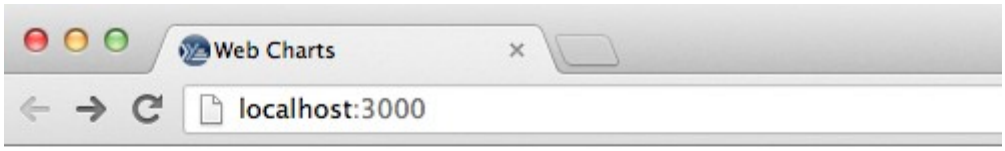

Force-Directed Layout
localhost:3000/force
Bookmark on Delicio
Clip to Evernote
Read Later




Interactive Force-Directive
localhost:3000/int-force
New York
Total
19378102 (2000: 18976457)
White
12740974 (2000: 12893689)
African-American
3073800 (2000: 3014385)
Native American
106906 (2000: 82461)
Asian
1420244 (2000: 1044976)
Hawaiian
8766 (2000: 8818)
Other
1441563 (2000: 1341946)
Multi-racial
585849 (2000: 590182)

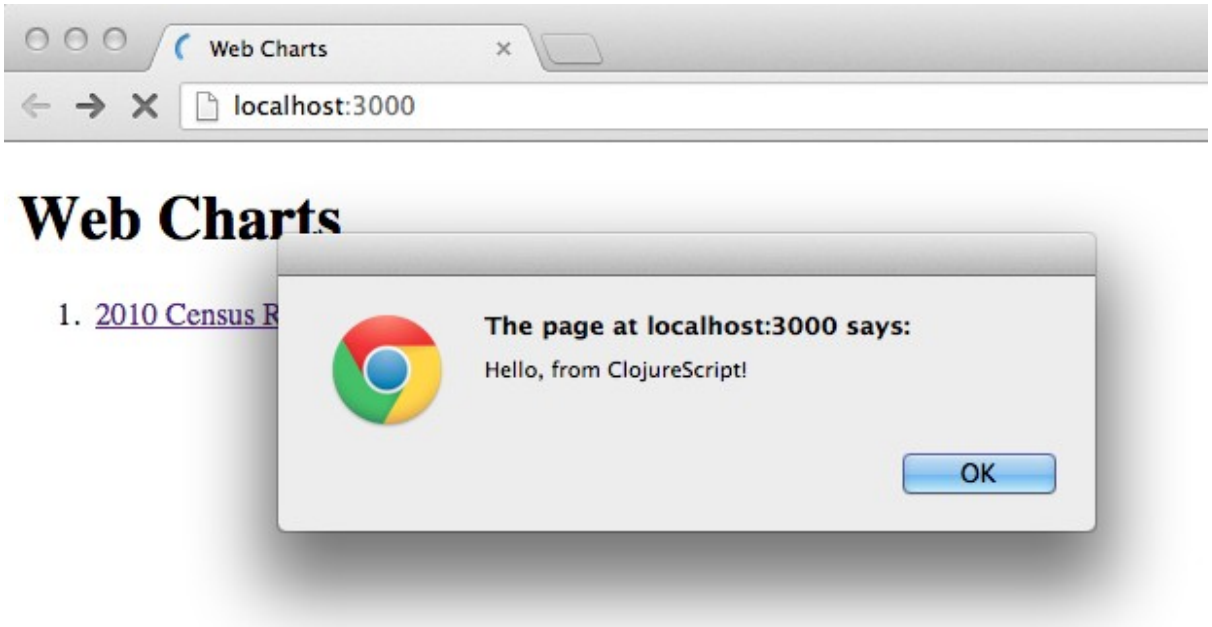


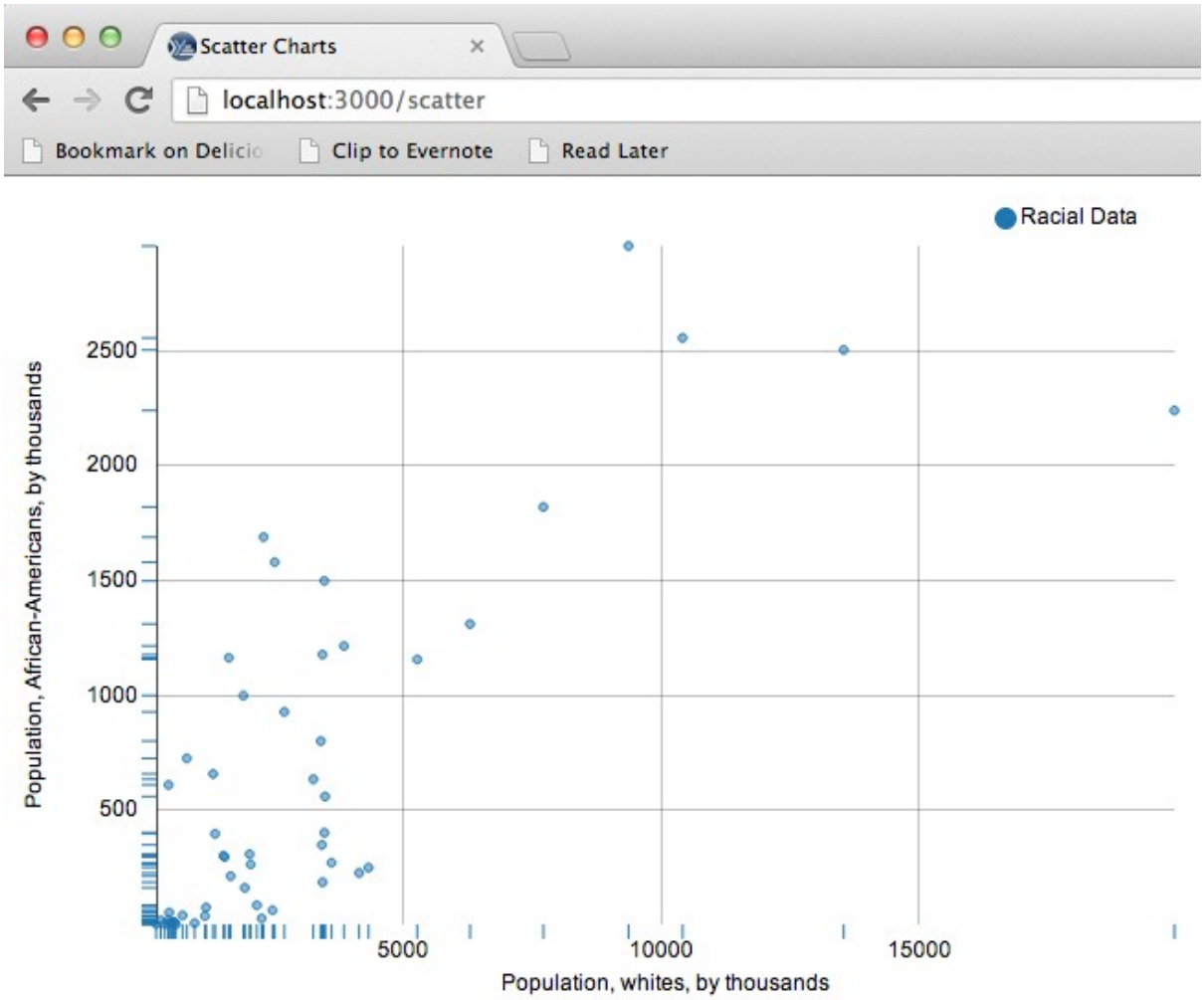

```
[{"black":58,"geoid":100100,"white":129,"white_2000":null,"state":
banda
CDP","other_2000":null,"asian":0,"total_2000":null,"hu100_2000":nu
aiian":0,"asian_2000":null,"multiracial":3,"multiracial_2000":null
{"black":1113,"geoid":100124,"white":1463,"white_2000":1692,"state
me":"Abbeville
city","other_2000":85,"asian":26,"total_2000":2987,"hu100_2000":13
ian":0,"asian_2000":2,"multiracial":31,"multiracial_2000":15,"hawa
{"black":2030,"geoid":100460,"white":2366,"white_2000":3763,"state
me":"Adamsville
city","other_2000":8,"asian":14,"total_2000":4965,"hu100_2000":204
ian":1,"asian_2000":7,"multiracial":38,"multiracial_2000":33,"hawa
{"black":1,"geoid":100484,"white":751,"white_2000":719,"state":1,"
ison
town","other_2000":0,"asian":1,"total_2000":723,"hu100_2000":339,"
,"asian_2000":1,"multiracial":5,"multiracial_2000":1,"hawaiian_200
```

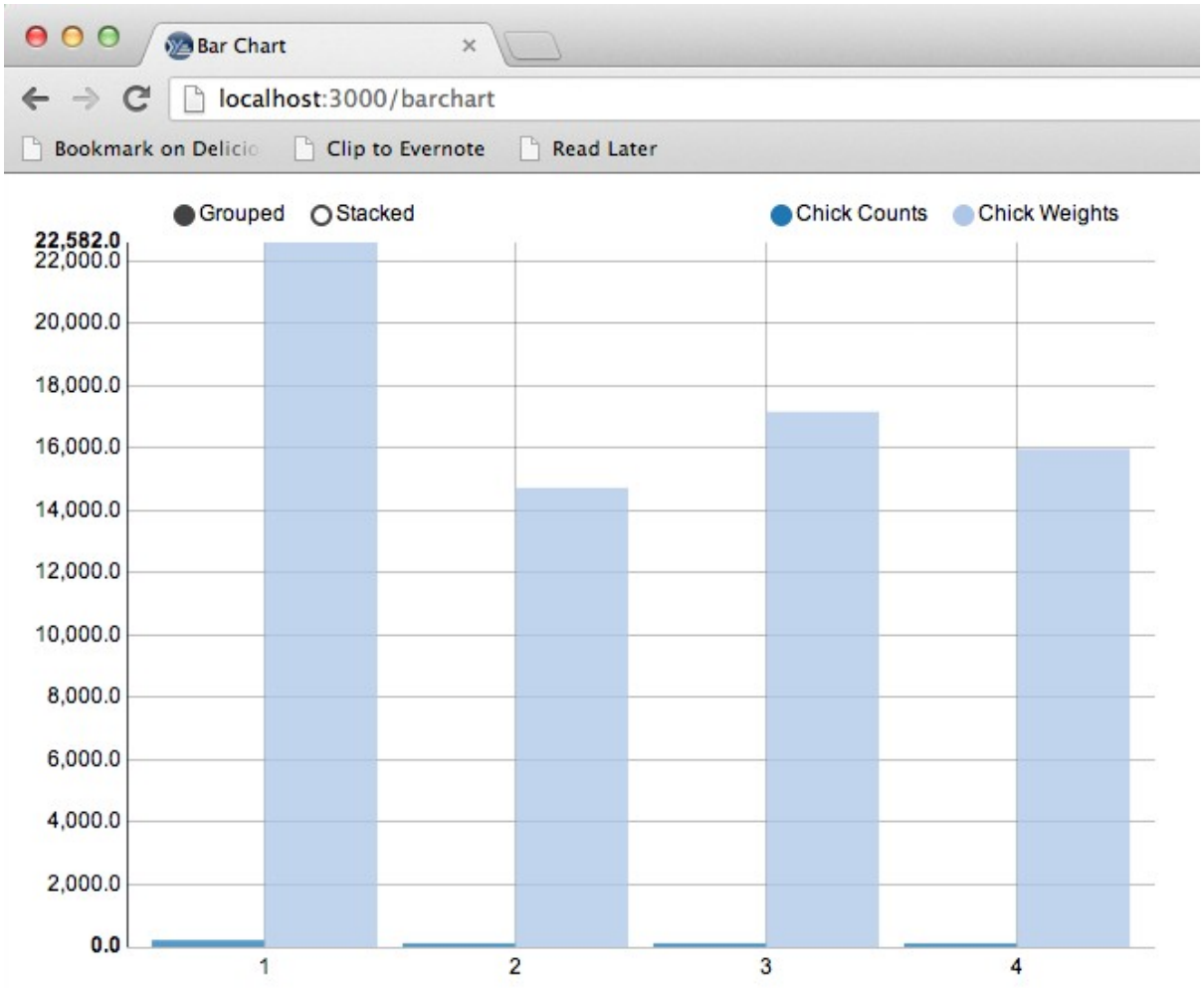


Web Charts

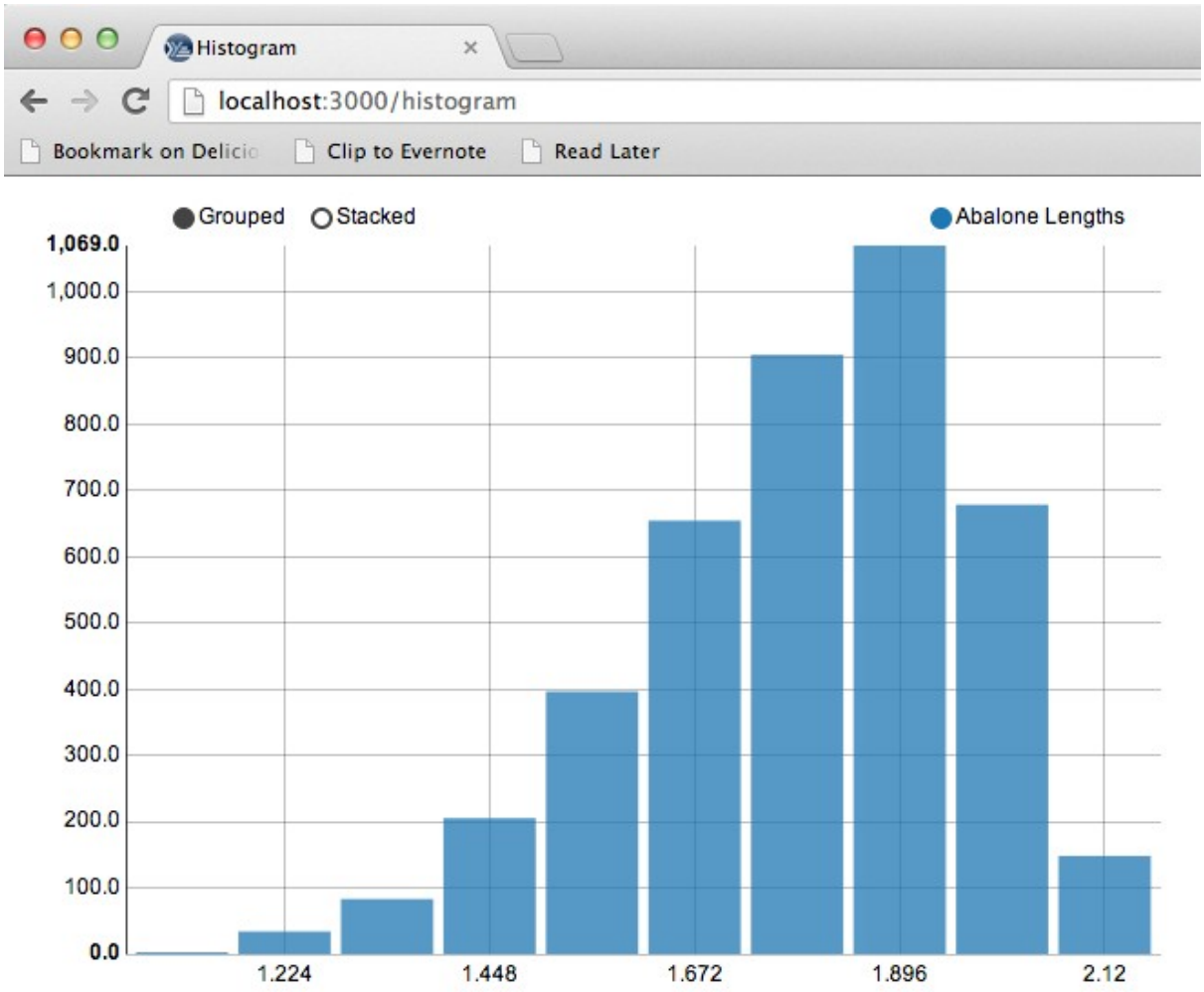
1. 2010 Census Race Data




Scatter Charts
localhost:3000/scatter
Bookmark on Delicio
Clip to Evernote
Read Later
Racial Data
Population, African-Americans, by thousands
2500
2000
1500
1000
500
5000
10000
15000
Population, whites, by thousands


Bar Chart
localhost:3000/barchart
Bookmark on Delicio
Clip to Evernote
Read Later
Grouped
Stacked
Chick Counts
Chick Weights
22,582.0
22,000.0
20,000.0
18,000.0
16,000.0
14,000.0
12,000.0
10,000.0
8,000.0
6,000.0
4,000.0
2,000.0
0.0
1
2
3
4

Line Chart
localhost:3000/lineplot
IBM
211.00
200.00
180.00
160.00
140.00
120.00
100.00
80.00
60.00
55.07
2001
2004
2006
2008
2010
2012
2012
Date


Histogram
localhost:3000/histogram
Bookmark on Delicio
Clip to Evernote
Read Later
Grouped
Stacked
Abalone Lengths
1,069.0
1,000.0
900.0
800.0
700.0
600.0
500.0
400.0
300.0
200.0
100.0
0.0
1.224
1.448
1.672
1.896
2.12