Exploratory Analysis

Data Transformations

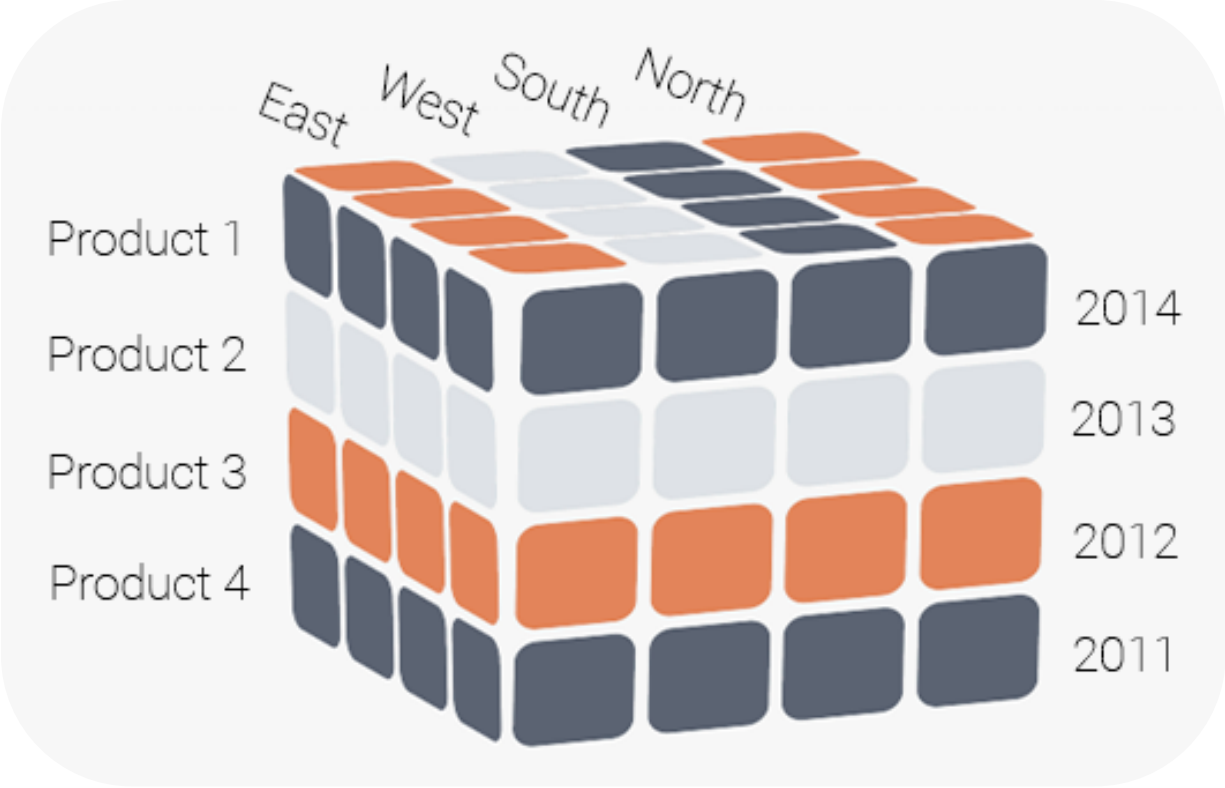
You've already done this with DataFrames in Pandas or R.

They look like this:

	account	campaign	date	successes	trials	rate
455	1	Campaign #76	2012-08-14 11:56:20 -0400	2	2	1.000000
449	1	Campaign #78	2012-08-14 12:06:20 -0400	2	2	1.000000
438	1	Campaign #87	2012-08-14 18:06:30 -0400	27	118	0.228814
431	1	Campaign #95	2012-08-15 00:07:42 -0400	22	118	0.186441
422	1	Campaign #99	2012-08-15 01:27:48 -0400	25	120	0.208333

Data Transformations

But you think of them like this:



Common Transformations

Normalize

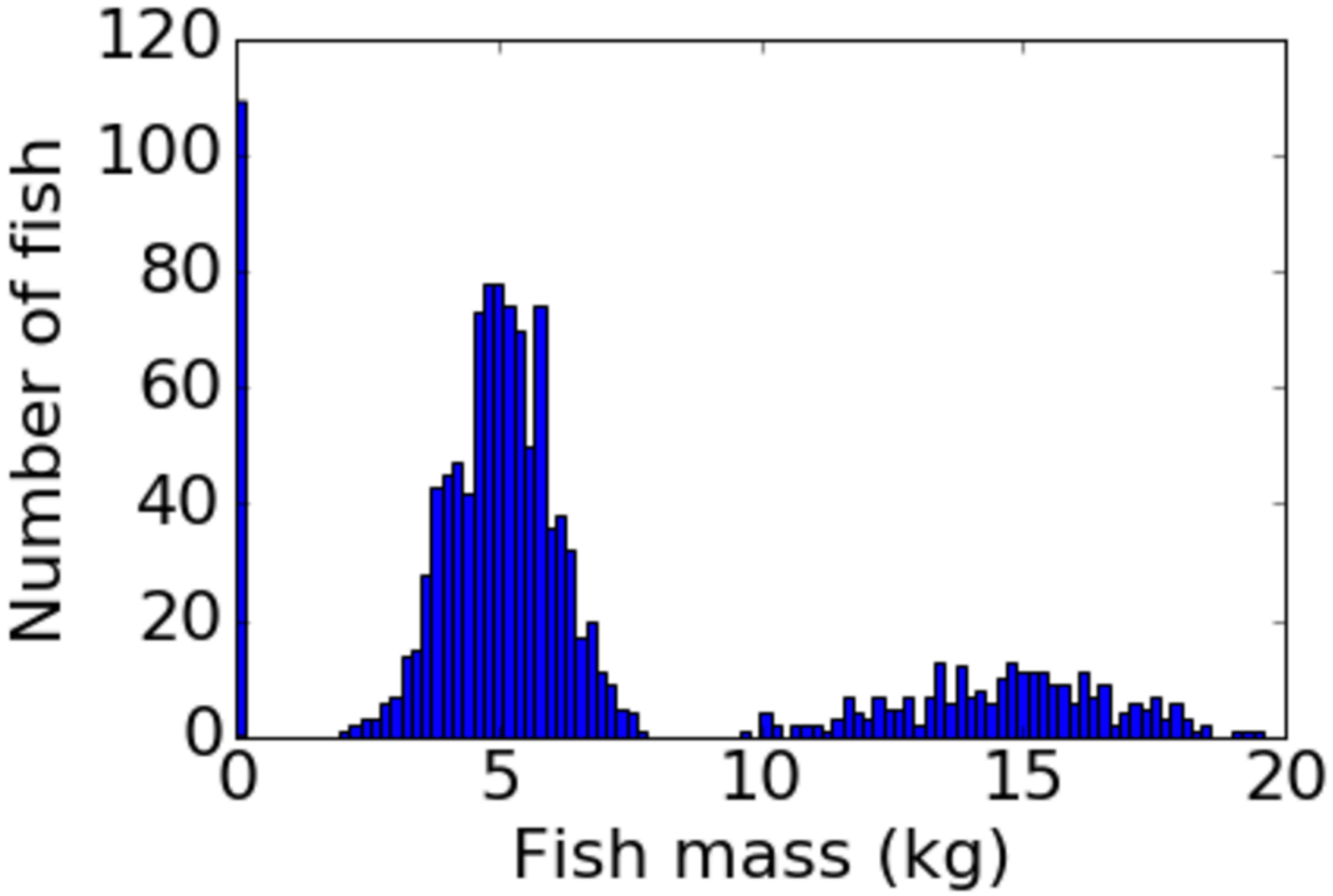
Log

Power

Binning

Grouping

What Does A Look Like?

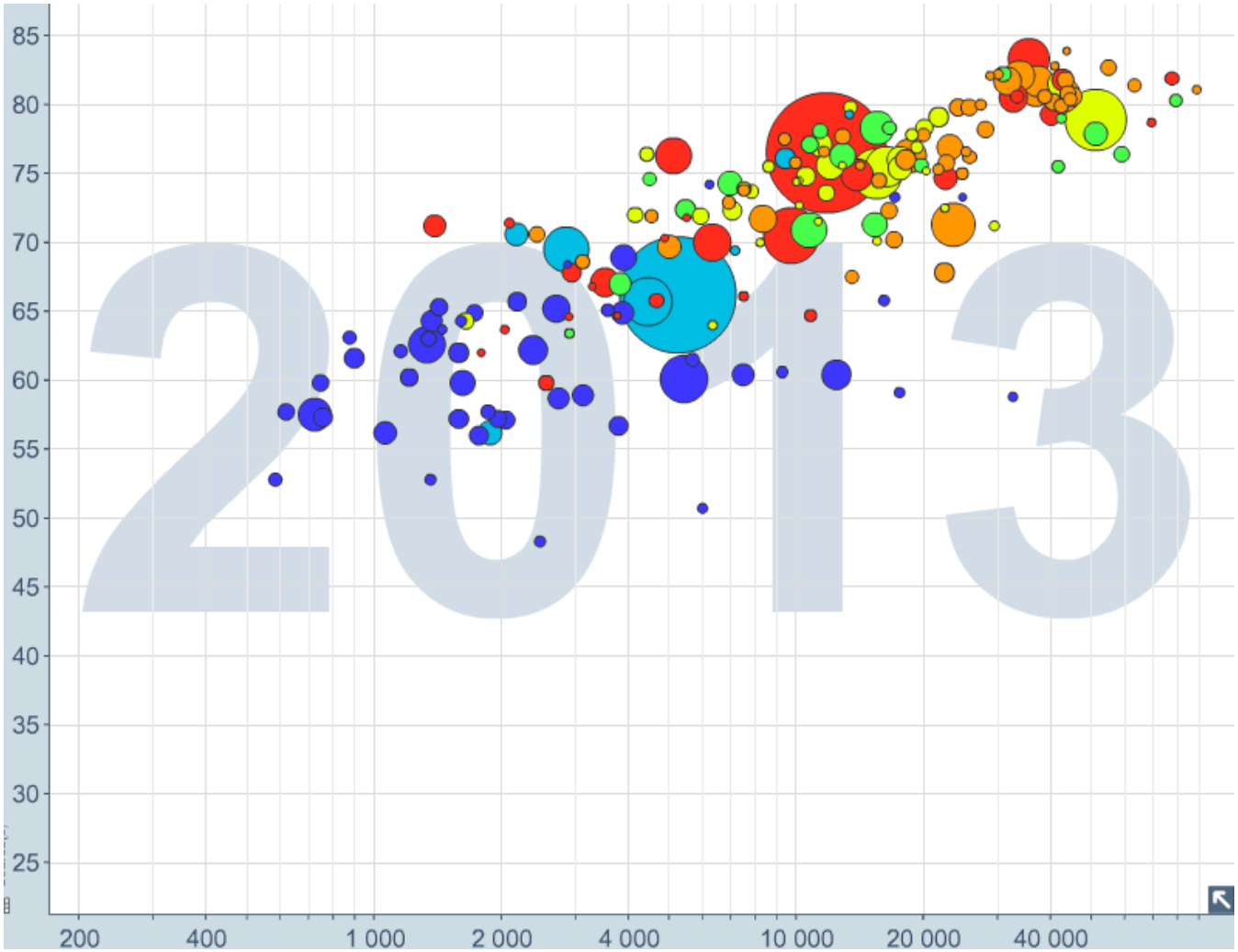


Histograms, Histograms, Histograms

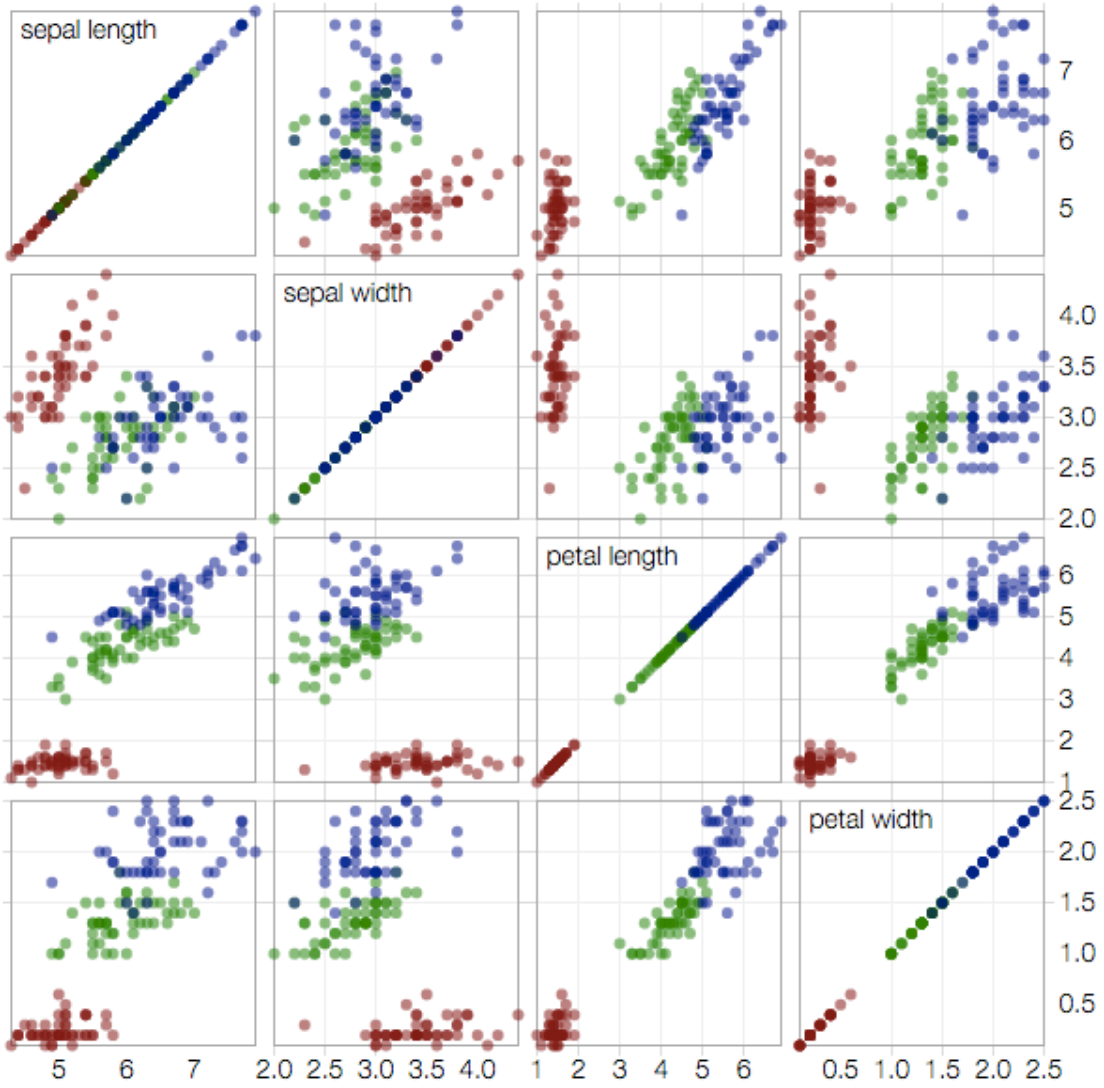
An cornerstone in the EDA toolbox!

“Above all else, show the data”

What's the Relationship Between A and B?

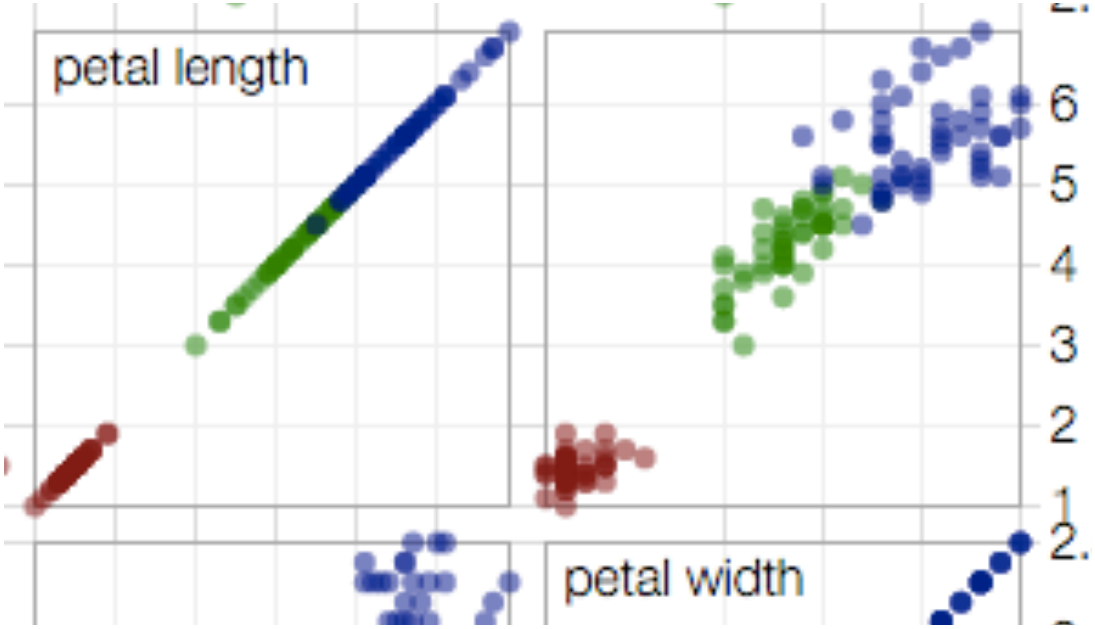


Hypothesis Generation



Hypothesis Testing

E.g., given



we find R^2=… and p=...

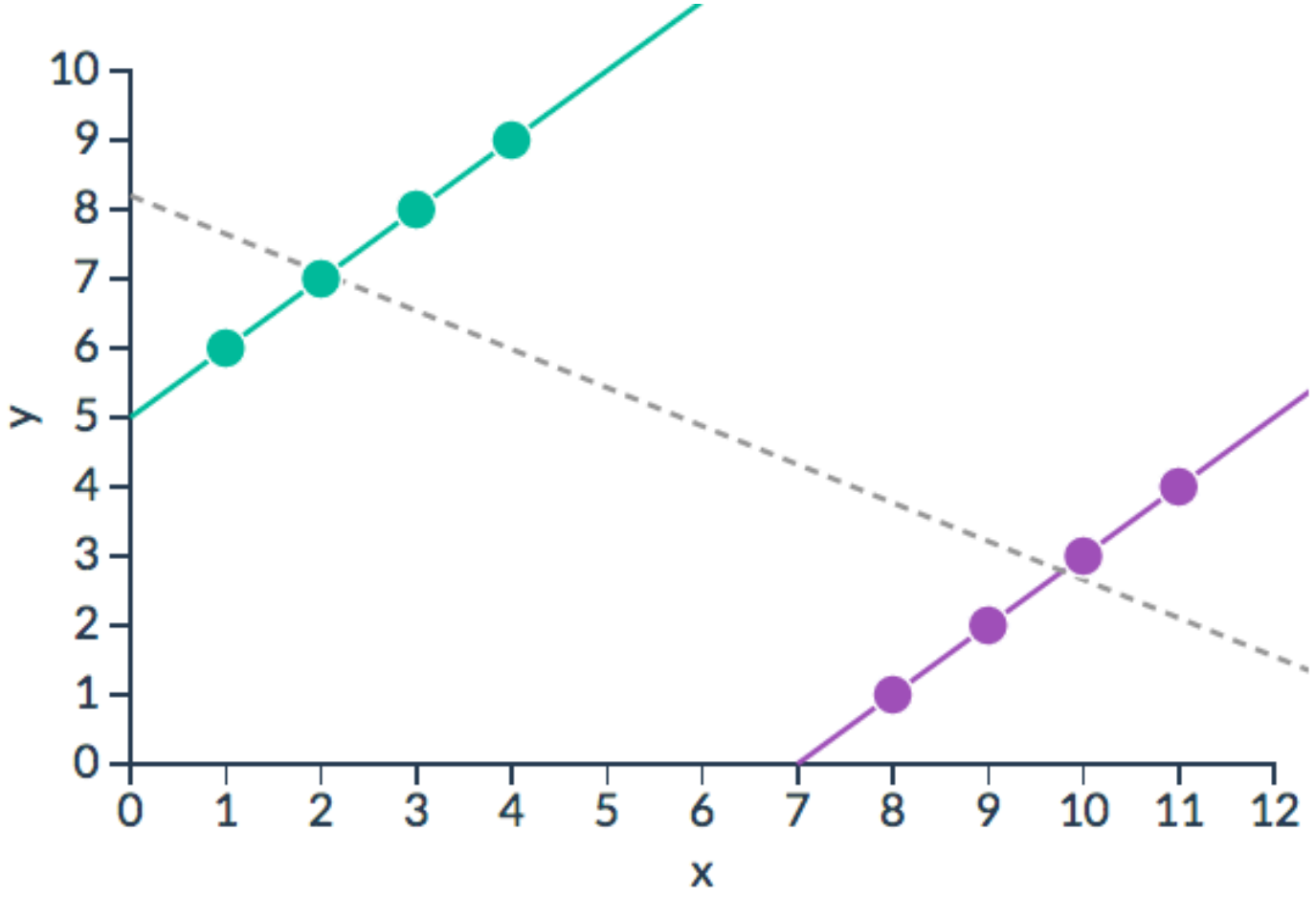
Mantras

Be skeptical: What assumptions have been made?

Explore iteratively: Start simple, keep asking questions.

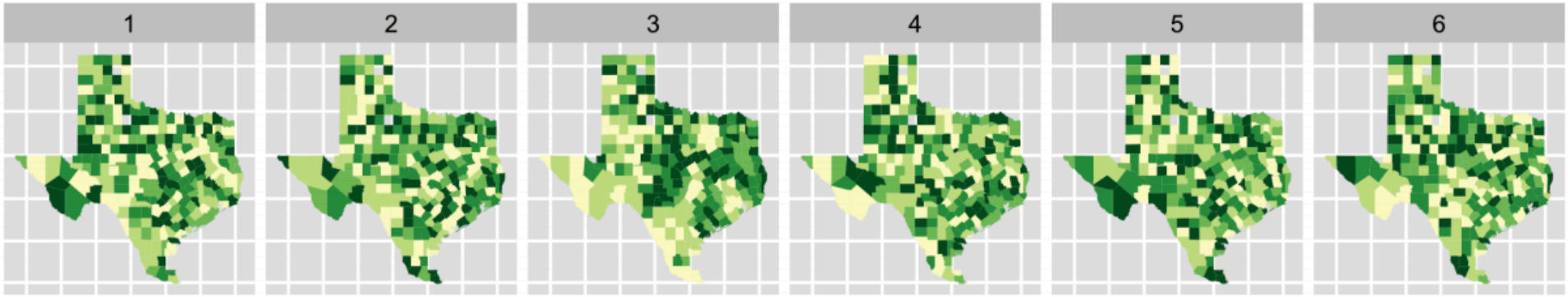
Avoid fixation: Use a variety of graphics to inspect more angles.

Simpson (Yule)



http://vudlab.com/simpsons/

Find the Real Data



Wickham, et al. 2010

Iterating

[example exploratory iteration]

Iterating

Measuring the happiness of words

Goal:

- Our overall aim is to assess how people feel about individual words.
- With this particular survey, we are focusing on the dual emotions of happiness and sadness.

Time required:

- 6 to 8 minutes.

Instructions and Example:

- You are to rate individual words on a 9 point unhappy-happy scale:

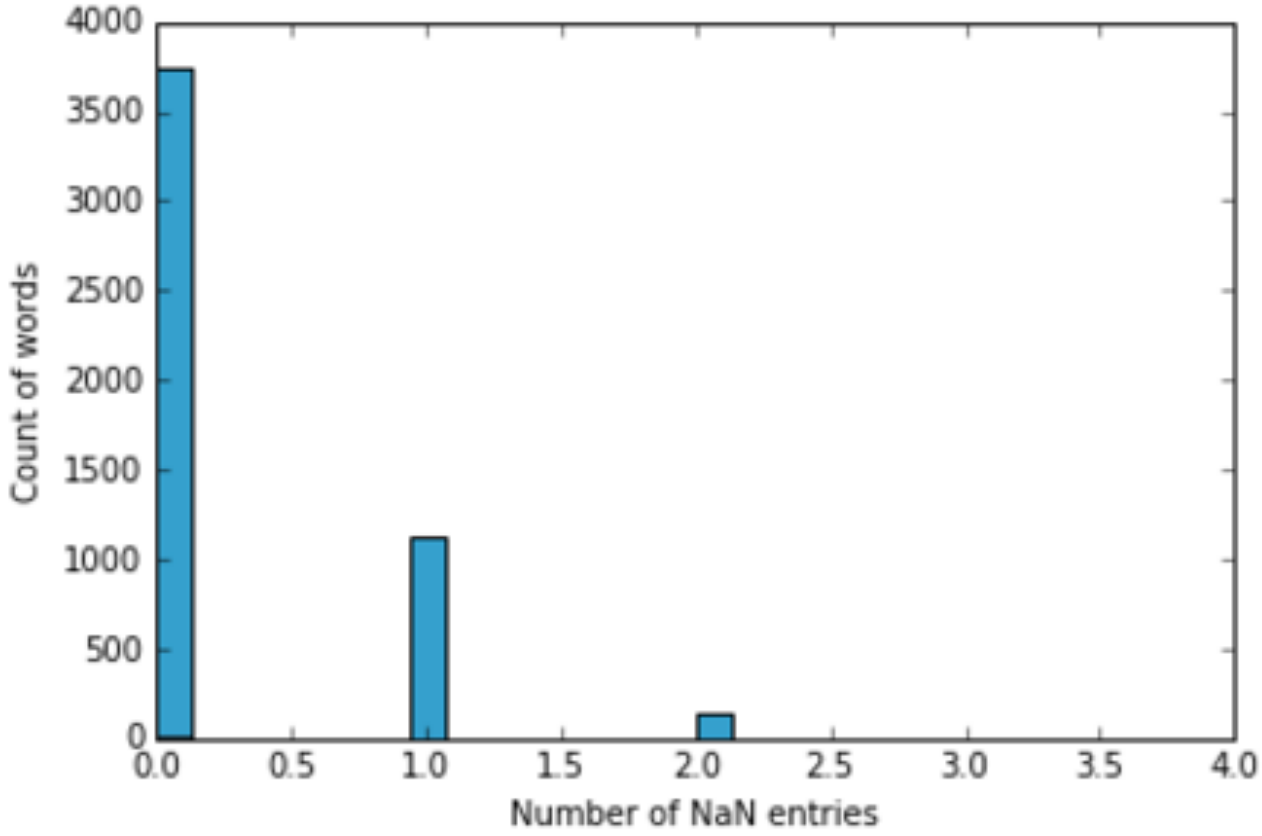
sunshine

1. Read the word, ("sunshine" in the above example) and observe your emotional response.
2. Click on the face that best corresponds to your response.

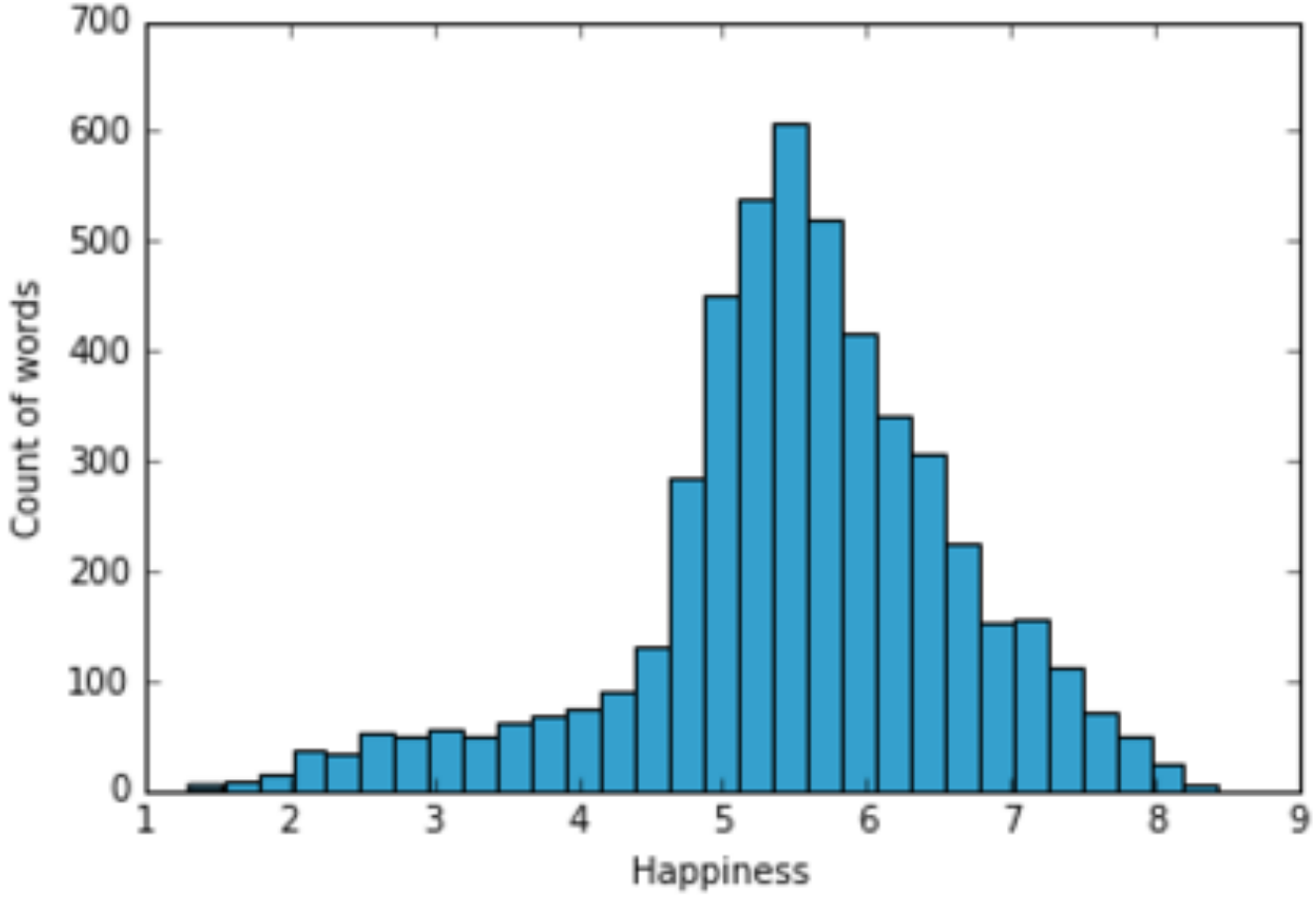
```
['HITId', 'HITTypeId', 'Title', 'Description', 'Keywords', 'Reward', 'CreationTime', 'MaxAssignments', 'RequesterAnno
tation', 'AssignmentDurationInSeconds', 'AutoApprovalDelayInSeconds', 'Expiration', 'NumberOfSimilarHITs', 'LifetimeI
nSeconds', 'AssignmentId', 'WorkerId', 'AssignmentStatus', 'AcceptTime', 'SubmitTime', 'AutoApprovalTime', 'ApprovalT
ime', 'RejectionTime', 'RequesterFeedback', 'WorkTimeInSeconds', 'Input.anew_word_1', 'Input.anew_word_2', 'Input.ane
w_word_3', 'Input.anew_word_4', 'Input.anew_word_5', 'Input.anew_word_6', 'Input.anew_word_7', 'Input.anew_word_8',
 'Input.anew_word_9', 'Input.anew_word_10', 'Input.anew_word_11', 'Input.anew_word_12', 'Input.anew_word_13', 'Input.
anew_word_14', 'Input.anew_word_15', 'Input.anew_word_16', 'Input.anew_word_17', 'Input.anew_word_18', 'Input.anew_wo
rd_19', 'Input.anew_word_20', 'Input.anew_word_21', 'Input.anew_word_22', 'Input.anew_word_23', 'Input.anew_word_24',
 'Input.anew_word_25', 'Input.anew_word_26', 'Input.anew_word_27', 'Input.anew_word_28', 'Input.anew_word_29', 'Inpu
t.anew_word_30', 'Input.anew_word_31', 'Input.anew_word_32', 'Input.anew_word_33', 'Input.anew_word_34', 'Input.anew_
word_35', 'Input.anew_word_36', 'Input.anew_word_37', 'Input.anew_word_38', 'Input.anew_word_39', 'Input.anew_word_4
0', 'Input.anew_word_41', 'Input.anew_word_42', 'Input.anew_word_43', 'Input.anew_word_44', 'Input.anew_word_45', 'In
put.anew_word_46', 'Input.anew_word_47', 'Input.anew_word_48', 'Input.anew_word_49', 'Input.anew_word_50', 'Input.ane
w_word_51', 'Input.anew_word_52', 'Input.anew_word_53', 'Input.anew_word_54', 'Input.anew_word_55', 'Input.anew_word_
56', 'Input.anew_word_57', 'Input.anew_word_58', 'Input.anew_word_59', 'Input.anew_word_60', 'Input.anew_word_61', 'I
nput.anew_word_62', 'Input.anew_word_63', 'Input.anew_word_64', 'Input.anew_word_65', 'Input.anew_word_66', 'Input.an
ew_word_67', 'Input.anew_word_68', 'Input.anew_word_69', 'Input.anew_word_70', 'Input.anew_word_71', 'Input.anew_word
_72', 'Input.anew_word_73', 'Input.anew_word_74', 'Input.anew_word_75', 'Input.anew_word_76', 'Input.anew_word_77',
 'Input.anew_word_78', 'Input.anew_word_79', 'Input.anew_word_80', 'Input.anew_word_81', 'Input.anew_word_82', 'Inpu
t.anew_word_83', 'Input.anew_word_84', 'Input.anew_word_85', 'Input.anew_word_86', 'Input.anew_word_87', 'Input.anew_
word_88', 'Input.anew_word_89', 'Input.anew_word_90', 'Input.anew_word_91', 'Input.anew_word_92', 'Input.anew_word_9
3', 'Input.anew_word_94', 'Input.anew_word_95', 'Input.anew_word_96', 'Input.anew_word_97', 'Input.anew_word_98', 'In
put.anew_word_99', 'Input.anew_word_100', 'Answer.Qpersonfeel25', 'Answer.Qpersonfeel46', 'Answer.Qpersonfeel67', 'An
swer.Q7FirstLanguage', 'Answer.Qpersonfeel88', 'Answer.Qpersonfeel26', 'Answer.Qpersonfeel47', 'Answer.Qpersonfeel6
8', 'Answer.Qpersonfeel90', 'Answer.Qpersonfeel89', 'Answer.Qpersonfeel27', 'Answer.Qpersonfeel48', 'Answer.Qpersonfe
el70', 'Answer.Qpersonfeel69', 'Answer.Qpersonfeel91', 'Answer.Q1Gender', 'Answer.Qpersonfeel50', 'Answer.Qpersonfeel
28', 'Answer.Qpersonfeel49', 'Answer.Qpersonfeel71', 'Answer.Qpersonfeel100', 'Answer.Qpersonfeel92', 'Answer.Qperson
feel51', 'Answer.Qpersonfeel29', 'Answer.Qpersonfeel72', 'Answer.Qpersonfeel93', 'Answer.Qpersonfeel30', 'Answer.Qper
sonfeel73', 'Answer.Qpersonfeel94', 'Answer.Qpersonfeel10', 'Answer.Qpersonfeel31', 'Answer.Qpersonfeel52', 'Answer.Q
personfeel95', 'Answer.Qpersonfeel11', 'Answer.Qpersonfeel32', 'Answer.Qpersonfeel53', 'Answer.comment', 'Answer.Qper
sonfeel74', 'Answer.Qpersonfeel12', 'Answer.Q4income', 'Answer.Qpersonfeel33', 'Answer.Qpersonfeel54', 'Answer.Q2ag
e', 'Answer.Qpersonfeel75', 'Answer.Qpersonfeel96', 'Answer.Qpersonfeel1', 'Answer.Qpersonfeel13', 'Answer.Qpersonfee
l34', 'Answer.Qpersonfeel55', 'Answer.Qpersonfeel76', 'Answer.Qpersonfeel97', 'Answer.Qpersonfeel2', 'Answer.Qpersonf
eel14', 'Answer.Qpersonfeel35', 'Answer.Qpersonfeel56', 'Answer.Qpersonfeel77', 'Answer.Qpersonfeel98', 'Answer.Qpers
onfeel3', 'Answer.Qpersonfeel15', 'Answer.Qpersonfeel36', 'Answer.Qpersonfeel57', 'Answer.Qpersonfeel78', 'Answer.Qpe
rsonfeel99', 'Answer.Qpersonfeel58', 'Answer.Qpersonfeel16', 'Answer.Qpersonfeel4', 'Answer.Qpersonfeel37', 'Answer.Q
personfeel80', 'Answer.Qpersonfeel79', 'Answer.Qpersonfeel5', 'Answer.Qpersonfeel59', 'Answer.Qpersonfeel60', 'Answe
r.country-from', 'Answer.Qpersonfeel17', 'Answer.Qpersonfeel38', 'Answer.Qpersonfeel81', 'Answer.Qpersonfeel6', 'Answ
er.Qpersonfeel61', 'Answer.Qpersonfeel40', 'Answer.Qpersonfeel18', 'Answer.Qpersonfeel39', 'Answer.Qpersonfeel82', 'A
nswer.Qpersonfeel7', 'Answer.Qpersonfeel62', 'Answer.Qpersonfeel83', 'Answer.Qpersonfeel20', 'Answer.Qpersonfeel41',
 'Answer.Qpersonfeel19', 'Answer.Qpersonfeel42', 'Answer.country-live', 'Answer.Qpersonfeel8', 'Answer.Qpersonfeel6
3', 'Answer.Qpersonfeel84', 'Answer.Qpersonfeel21', 'Answer.Qpersonfeel43', 'Answer.Qpersonfeel9', 'Answer.Qpersonfee
l64', 'Answer.Q3eductation', 'Answer.Qpersonfeel22', 'Answer.Qpersonfeel85', 'Answer.Qpersonfeel44', 'Answer.Qpersonf
eel65', 'Answer.Qpersonfeel23', 'Answer.Qpersonfeel86', 'Answer.Qpersonfeel45', 'Answer.Qpersonfeel66', 'Answer.Qpers
onfeel24', 'Answer.Qpersonfeel87', 'Approve', 'Reject']
```

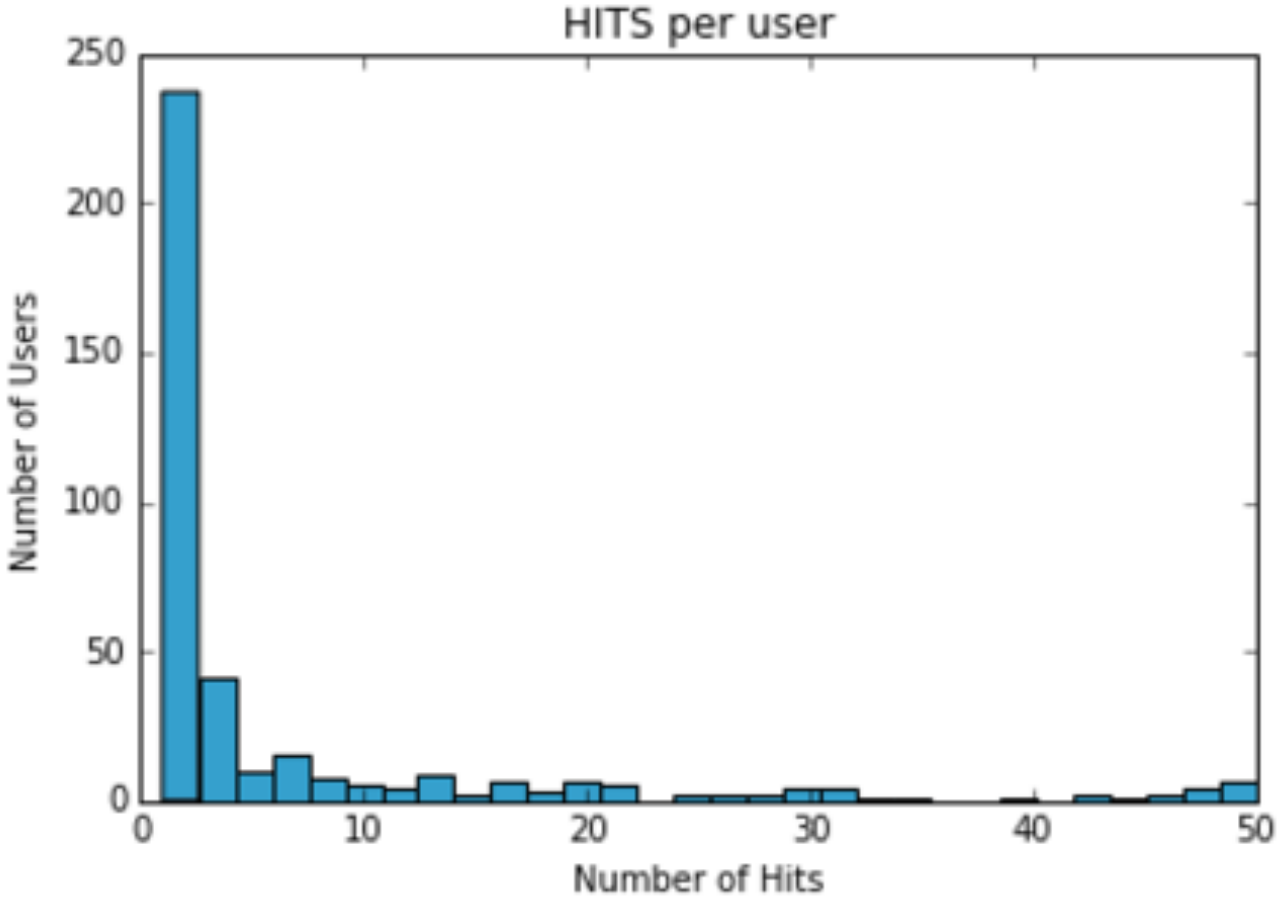
```
['"1DJIXOHIZKIZTZFYMD6MLFOY47FL8K"', '"1KJ0TI5VZG1SMAGRPH0SYSHX13HHL2"', '"Survey - word evaluations"', '"Rate your e
motional association of 100 words"', '"survey', ' words"', '"$0.60"', '"Wed Sep 29 16:29:10 GMT 2010"', '"50"', '""',
 '"7200"', '"1209600"', '"Fri Oct 29 16:29:10 GMT 2010"', '"40"', '""', '"11JLPS0QOH977DP44B5GXB767NJKT9"', '"A2BFQKE
Y8QKGQW"', '"Submitted"', '"Wed Sep 29 16:48:17 GMT 2010"', '"Wed Sep 29 16:53:38 GMT 2010"', '"Wed Oct 13 16:53:38 G
MT 2010"', '""', '""', '""', '"321"', '"mario"', '"intelligence"', '"agreed"', '"tengo"', '"slow"', '"pancakes"', '"b
ody"', '"joy"', '"charge"', '"countries"', '"ps3"', '"scores"', '"les"', '"joshua"', '"advanced"', '"betty"', '"hope
d"', '"fred"', '"previous"', '"appeal"', '"managed"', '"dan"', '"emily"', '"goes"', '"ko"', '"tryin"', '"academic"',
 '"shore"', '"talkin"', '"vital"', '"burger"', '"ltd"', '"library"', '"emeritus"', '"@ddlovato"', '"responsible"',
 '"fellow"', '"deh"', '"walmart"', '"taught"', '"tips"', '"oo"', '"rent"', '"lights"', '"clients"', '"mitchell"', '"b
et"', '"rude"', '"flying"', '"gripe"', '"game"', '"celebration"', '"super"', '"due"', '"conducted"', '"houston"', '"s
unday"', '"loud"', '"americans"', '"boards"', '"suite"', '"send"', '"tools"', '"mistake"', '"care"', '"projects"',
 '"safety"', '"services"', '"glee"', '"prosecutors"', '"birth"', '"diamond"', '"balance"', '"nd"', '"je"', '"branc
h"', '"reply"', '"weekly"', '"tough"', '"howard"', '"allowed"', '"partnership"', '"criticism"', '"judy"', '"medicar
e"', '"how"', '"posts"', '"models"', '"course"', '"bread"', '"historic"', '"truck"', '"enough"', '"toy"', '"lyrics"',
 '"sunset"', '"maintenance"', '"cried"', '"date"', '"rid"', '"6"', '"5"', '"7"', '"Yes|"', '"4"', '"5"', '"5"',
 '"5"', '"5"', '"5"', '"6"', '"3"', '"5"', '"6"', '"5"', '"Female"', '"3"', '"6"', '"5"', '"6"', '"4"', '"5"', '"5"',
 '"5"', '"7"', '"6"', '"5"', '"5"', '"5"', '"5"', '"6"', '"6"', '"6"', '"5"', '"5"', '"7"', '""', '"5"', '"5"', '"$7
5', '000 - $87', '499"', '"8"', '"4"', '"32"', '"5"', '"7"', '"4"', '"6"', '"7"', '"5"', '"5"', '"5"', '"9"', '"5"',
 '"5"', '"5"', '"5"', '"5"', '"5"', '"7"', '"6"', '"5"', '"5"', '"5"', '"5"', '"4"', '"4"', '"5"', '"4"', '"5"',
 '"1"', '"6"', '"5"', '"USA_MI"', '"7"', '"5"', '"5"', '"7"', '"6"', '"5"', '"6"', '"3"', '"7"', '"7"', '"5"', '"3"',
 '"7"', '"5"', '"5"', '"6"', '"USA_MI"', '"6"', '"4"', '"4"', '"6"', '"5"', '"5"', '"4"', '"Bachelors degree"',
 '"4"', '"5"', '"5"', '"5"', '"5"', '"5"', '"5"', '"5"', '"5"', '"5"']
```

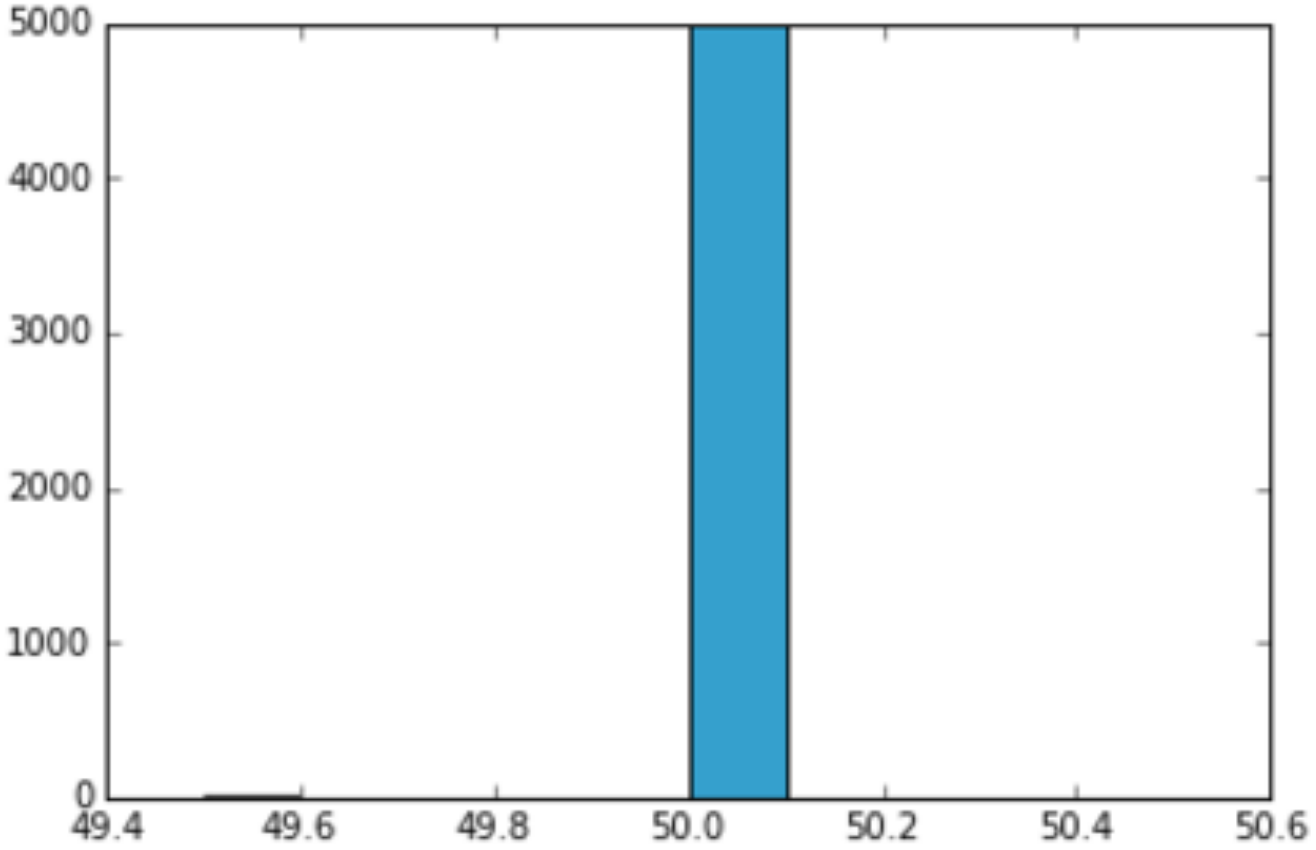
Check on the NaNs…

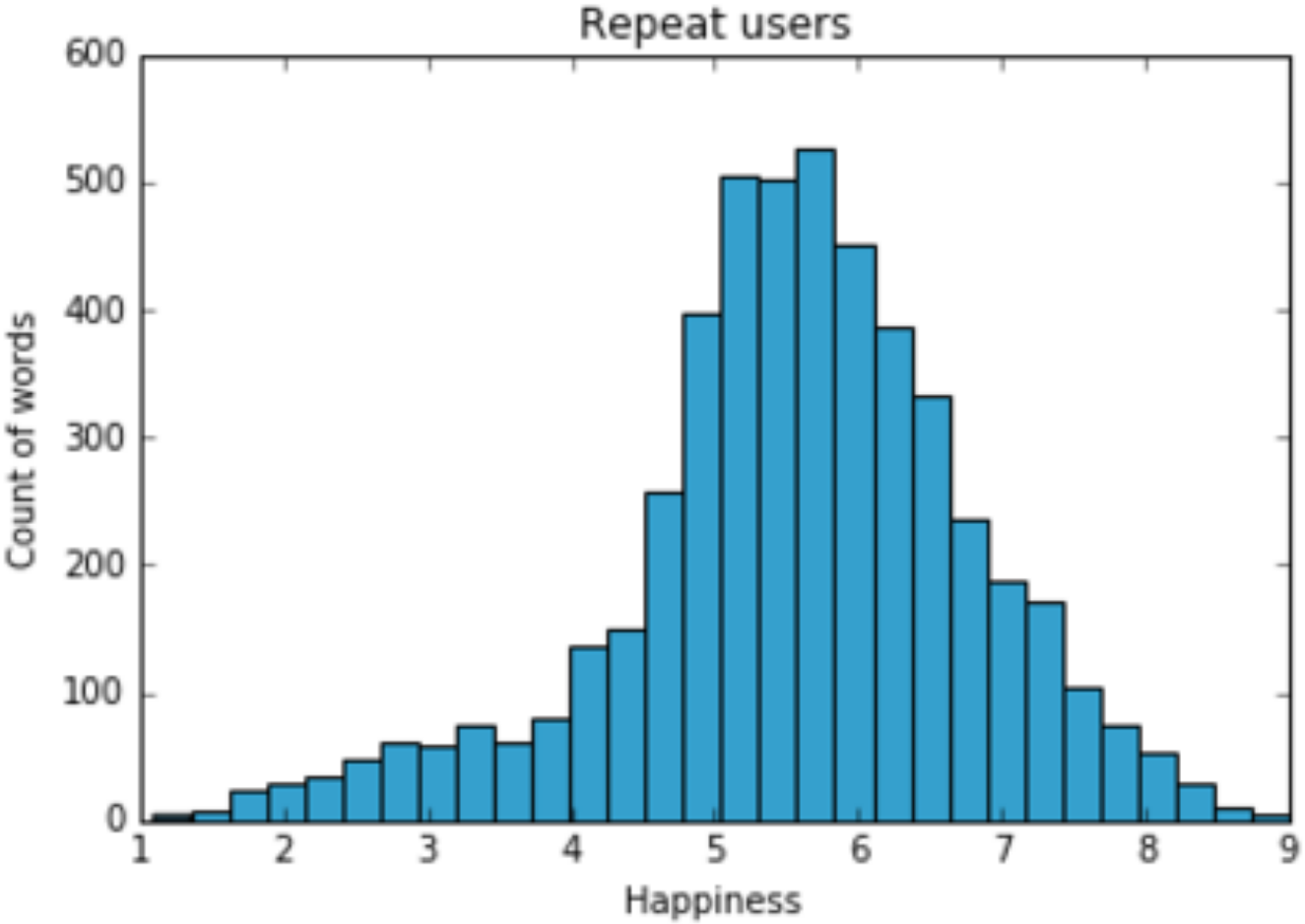


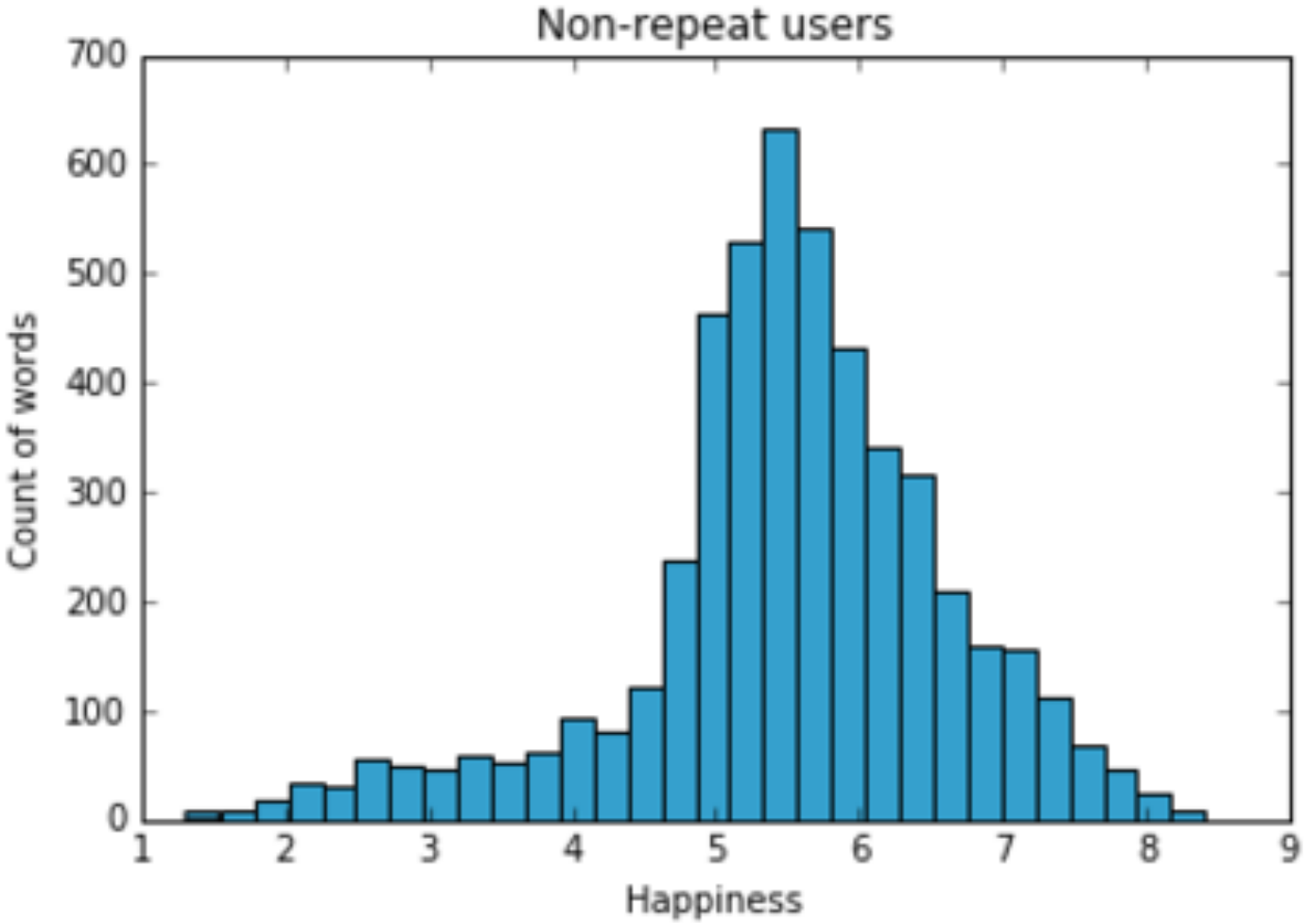
The Pollyanna Hypothesis Holds!

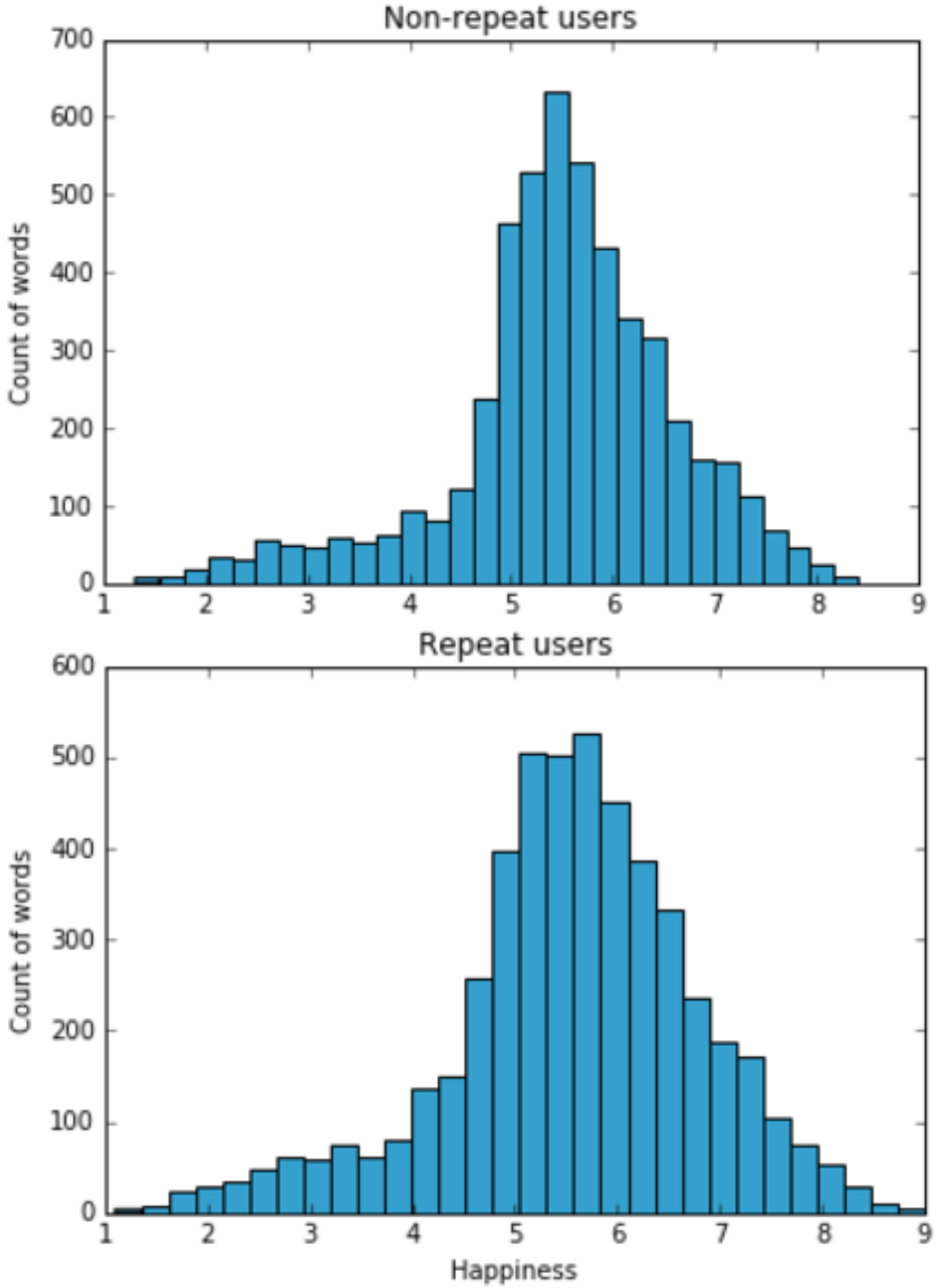



HITS per user
Number of Users
Number of Hits
250
200
150
100
50
0
0
10
20
30
40
50


5000
4000
3000
2000
1000
0
49.4
49.6
49.8
50.0
50.2
50.4
50.6


Repeat users
Count of words
Happiness
0
100
200
300
400
500
600
1
2
3
4
5
6
7
8
9


Non-repeat users
Count of words
Happiness
0
100
200
300
400
500
600
700
1
2
3
4
5
6
7
8
9


Non-repeat users
Count of words
0
100
200
300
400
500
600
700
1
2
3
4
5
6
7
8
9
Repeat users
Count of words
0
100
200
300
400
500
600
1
2
3
4
5
6
7
8
9
Happiness

Berkeley SCHOOL OF INFORMATION