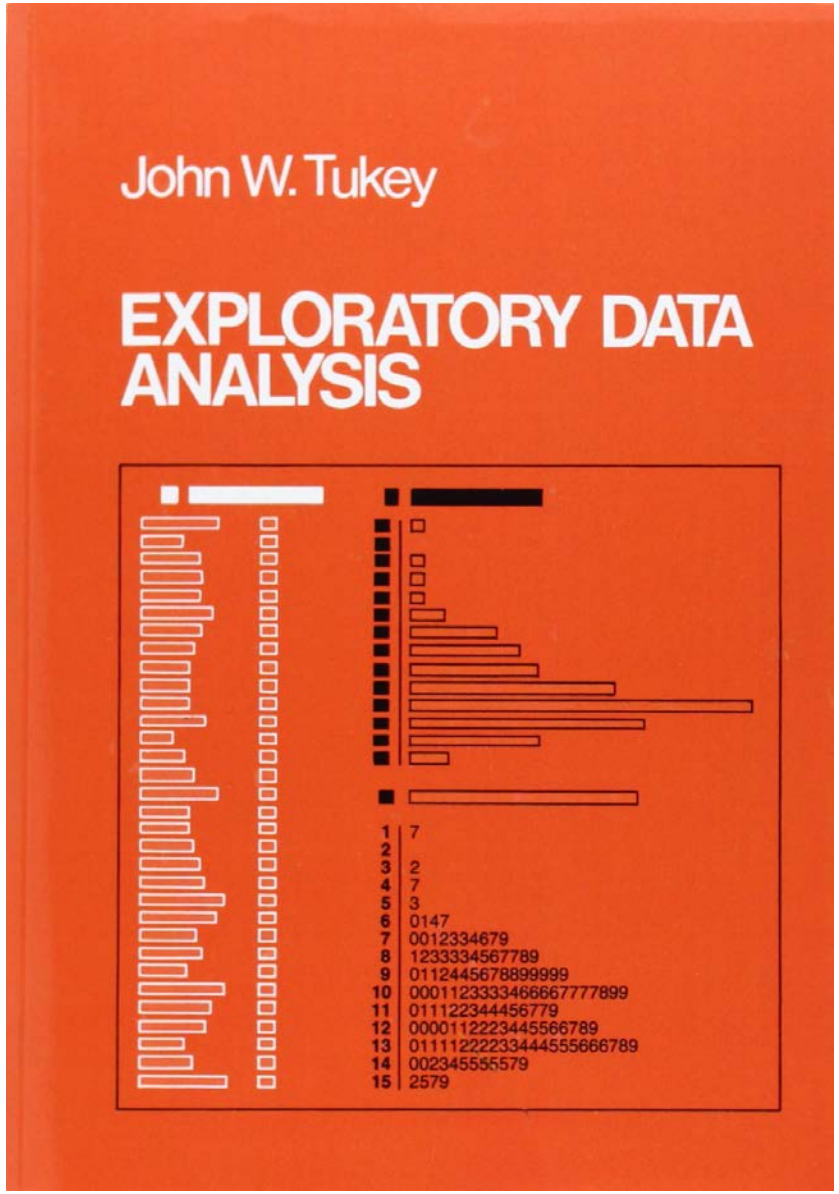
VISUAL DATA ANALYSIS

UNIVERSITY OF
CALGARY

“EXPLORATORY DATA ANALYSIS”

JOHN TUKEY

(IN CONTRAST TO “**CONFIRMATORY**” DATA ANALYSIS)



Based on insights developed at **Bell Labs** in the 60's

Introduced a number of novel techniques for visualizing and summarizing data:

- 5-number summary
- Box plots
- Stem and leaf diagrams

EXPLORATORY ANALYSIS IS ABOUT UNDERSTANDING DATA AND CHECKING ASSUMPTIONS

- IS THE DATA CORRECT?
- DOES IT MATCH OUR PREVIOUS EXPECTATIONS?
- IS THERE A RELATIONSHIP?
  A CORRELATION?
  A TREND?
  ETC.?

E.D.A. CIRCA ~1970

- Mostly done by hand
  (computation is expensive and inaccessible)

- Simple statistical summaries and charts

TUKEY'S 5-NUMBER SUMMARY

The sample minimum (smallest observation)

The lower quartile

The median (middle value)

The upper quartile

The sample maximum (largest observation)

STEM-AND-LEAF PLOTS

Volcano heights:

900 feet
1957 feet
823 feet
2620 feet
19300 feet
730 feet
1753 feet
603 feet
2930 feet
12400 feet
650 feet
3663 feet

0 | 9 = 900 feet

Stem-and-leaf displays:
heights of 218 volcanoes, unit 100 feet.

19 | 3 = 19,300 feet

0	98766562
1	97719630
2	699877665444222211009850
3	876655412099551426
4	99988443319294333611 07
5	97666666554422210097731
6	898665441077761065
7	98855431100652108073
8	653322122937
9	377655421000493
10	0984433165212
11	4963201631
12	45421164
13	47830
14	00
15	676
16	52
17	92
18	5
19	39730

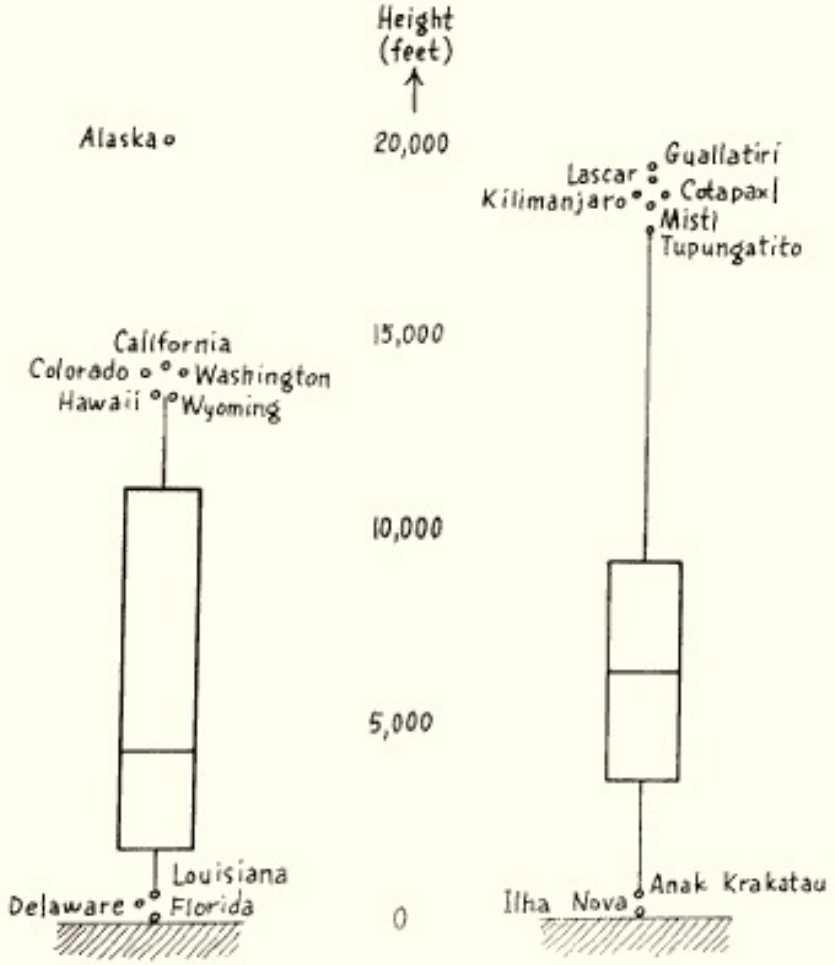
BOX PLOTS

exhibit 6 of chapter 2: various heights

Box-and-whisker plots with end values identified

A) HEIGHTS of 50 STATES

B) HEIGHTS of 219 VOLCANOS



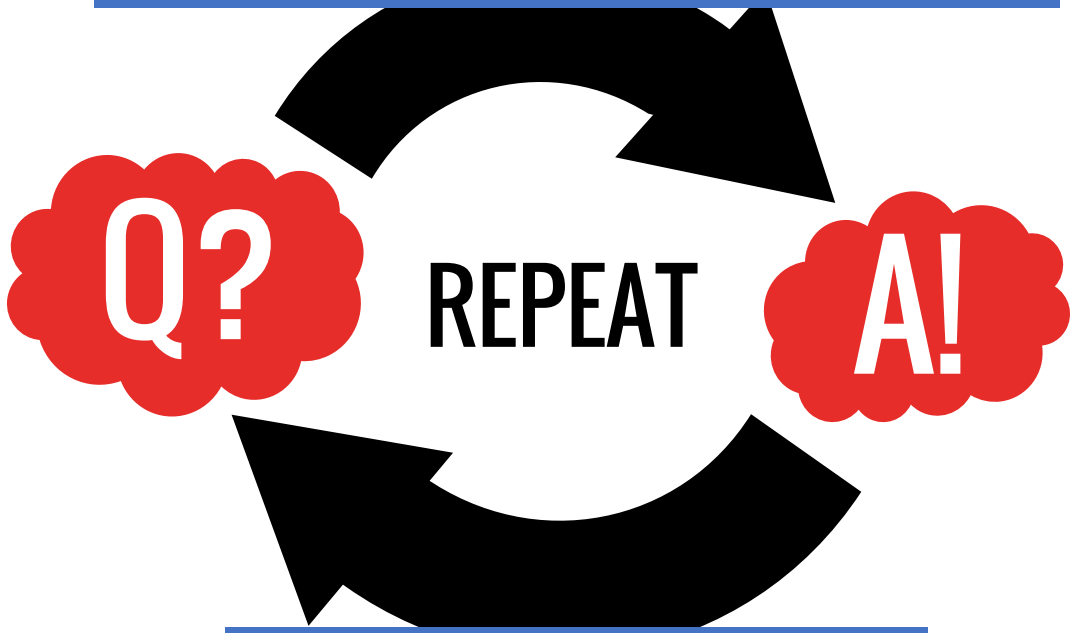
EXPLORATORY ANALYSIS IS ABOUT UNDERSTANDING DATA AND CHECKING ASSUMPTIONS

- IS THE DATA CORRECT?
- DOES IT MATCH OUR PREVIOUS EXPECTATIONS?
- IS THERE A RELATIONSHIP?
  A CORRELATION?
  A TREND?
  ETC.?

BUT, HOW SHOULD WE GO ABOUT DOING THIS?

ANALYSIS IS A CYCLE

GATHERING DATA,
APPLYING STATISTICAL TOOLS, AND
CONSTRUCTING GRAPHICS TO
ADDRESS QUESTIONS



INSPECT "ANSWERS" AND
ASSESS NEW QUESTIONS

START SIMPLE

IT'S EASY TO GET SIDETRACKED TRYING TO DO
COMPLICATED ANALYSES AND MISS THE BASIC STUFF

SOME FIRST STEPS TO START WITH

1. Plot the raw data
2. Plot simple statistics
3. Examine distributions
4. Look at attributes together

DON'T TRY TO CREATE A WHOLE NEW CHART ALL AT ONCE!
CHECK YOUR LOGIC AT EVERY STEP.

LOOKING AT DATA WITH "THE PAINTER'S EYE"

J. BERTIN

EMBRACING "SLOW DATA"

STEPHEN FEW

PLOT THE RAW DATA

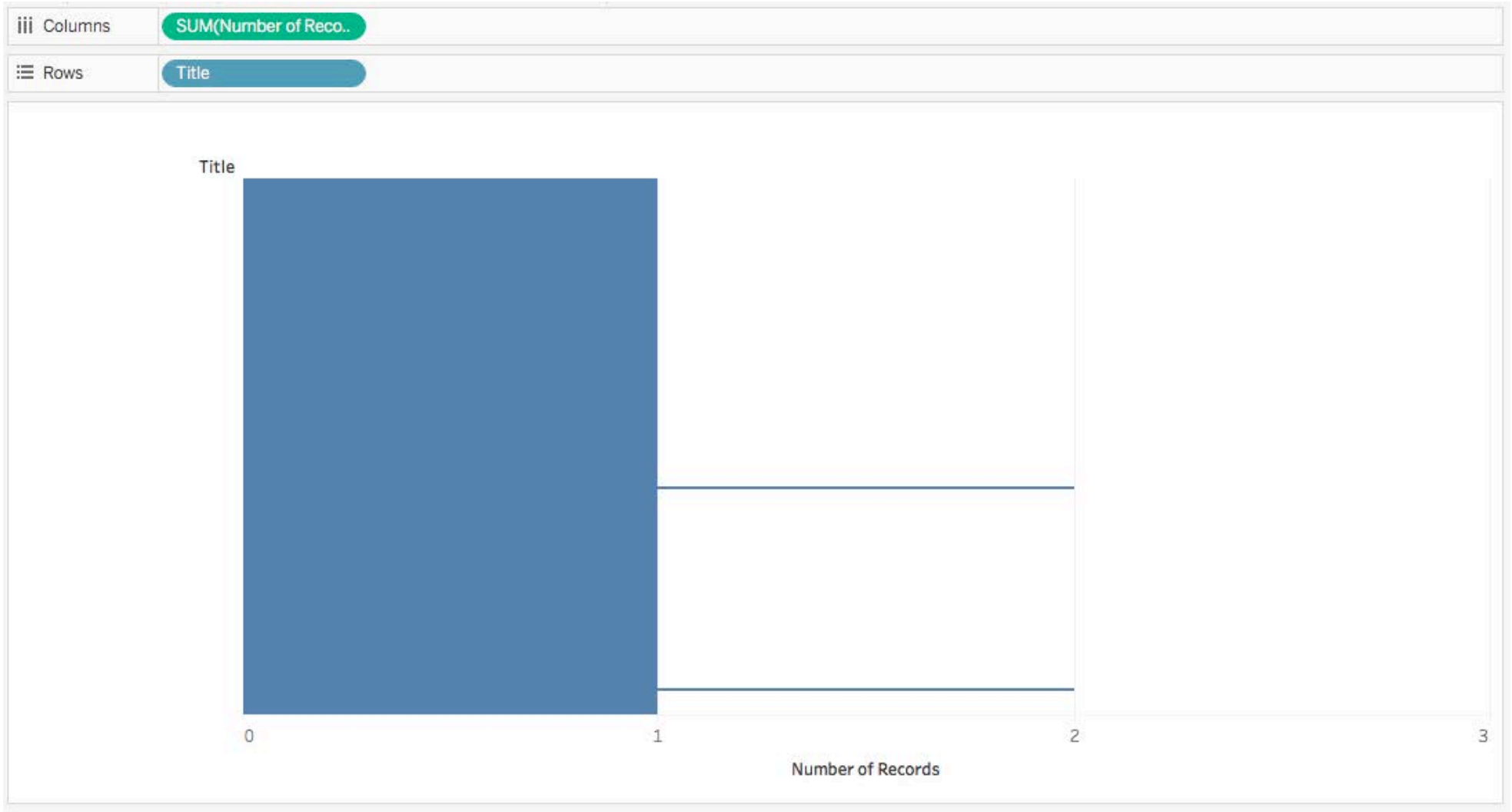
ARE THE FIELDS CORRECT?

WHAT ABOUT THE DATA TYPES?

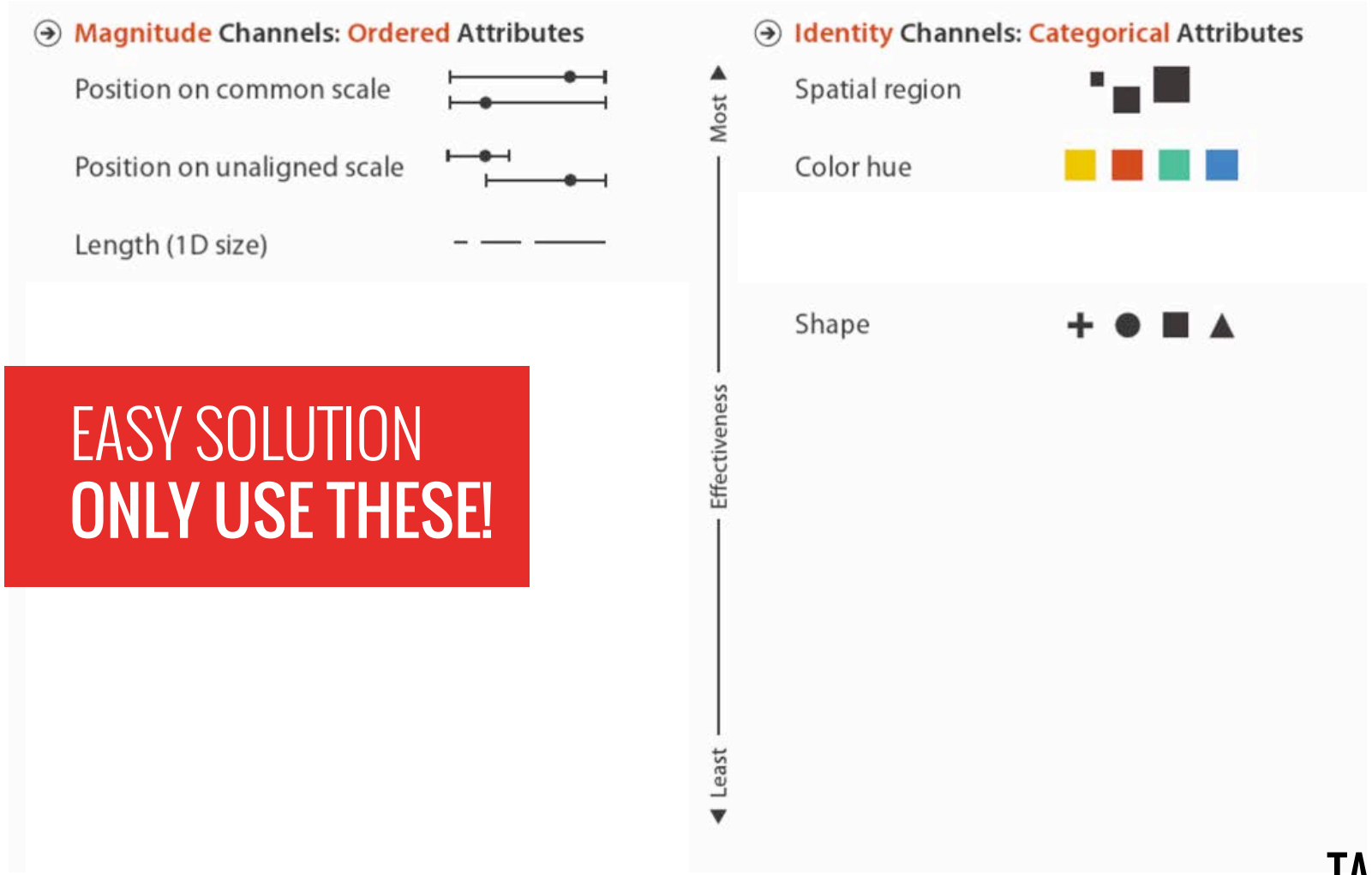
WHAT ABOUT THE VALUES?

# movies.csv Movie Id	Abc movies.csv Title	Abc movies.csv Genres	# ratings.csv User Id	# ratings.csv movieId (ratings.c...	# ratings.csv Rating	# ratings.csv Timestamp	=# Calculation Year
			2	1	5.00000	859,046,895	1995.00
			16	2	3.00000	849,188,326	1995.00
3	Grumpier Old Men (1...	Comedy|Romance	2	3	2.00000	859,046,959	1995.00
4	Waiting to Exhale (1...	Comedy|Drama|Rom...	80	4	3.50000	1,253,152,402	1995.00
5	Father of the Bride P...	Comedy	2	5	3.00000	859,046,959	1995.00
6	Heat (1995)	Action|Crime|Thriller	9	6	4.00000	842,686,600	1995.00
7	Sabrina (1995)	Comedy|Romance	3	7	3.00000	841,484,087	1995.00
8	Tom and Huck (1995)	Adventure|Children	156				00
9	Sudden Death (1995)	Action	16				00
10	GoldenEye (1995)	Action|Adventure|Th...	7	10	4.00000	1,322,062,970	1995.00
11	American President, ...	Comedy|Drama|Rom...	3	11	4.00000	841,483,689	1995.00
12	Dracula: Dead and Lo...	Comedy|Horror	29	12	3.00000	840,548,213	1995.00

USE THE SIMPLEST REPRESENTATION YOU CAN TO EVALUATE ALL OF THE DATA



CHOOSE REPRESENTATIONS THAT MAKE IT EASY TO COMPARE DIFFERENCES AND SEE PATTERNS



EASY SOLUTION
ONLY USE THESE!

TAMARA MUNZNER

DEFAULT TO SIMPLE AND EFFECTIVE CHART TYPES

the BAR

the LINE

the SCATTER

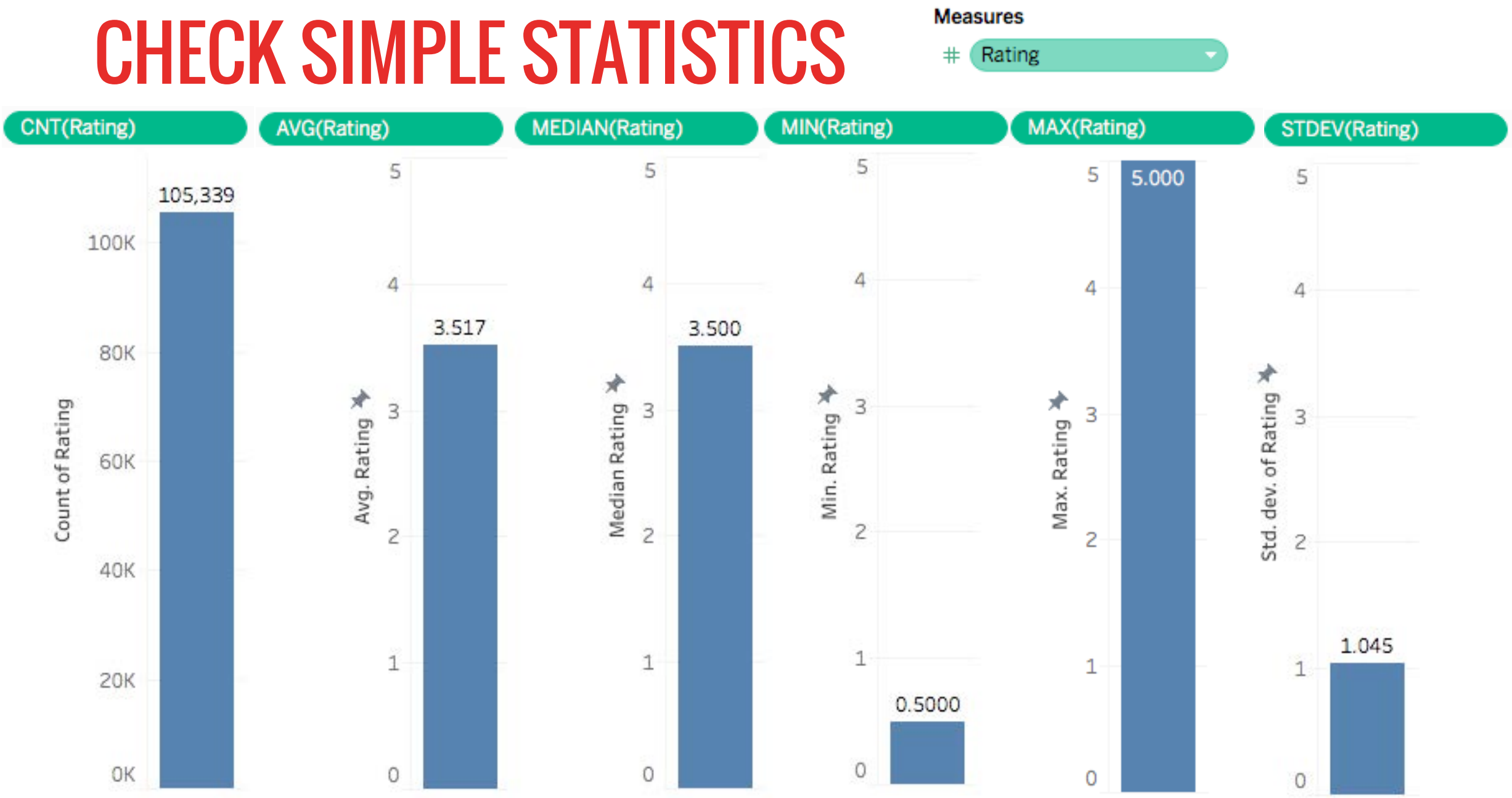
[Images from Nathan Yau]

+ COLOUR & SHAPE
TO SHOW CATEGORIES

SOME FIRST STEPS TO START WITH

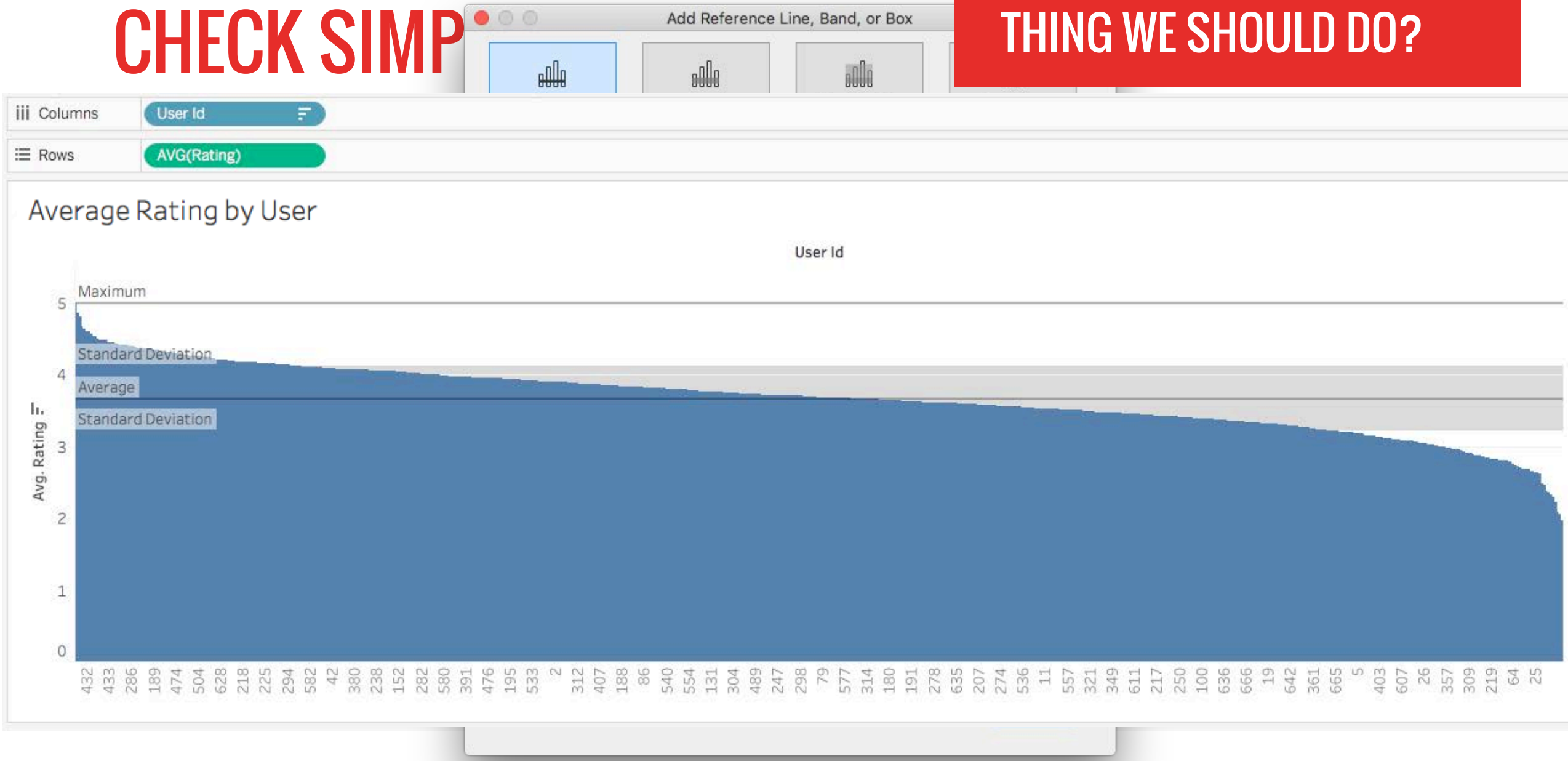
1. Plot the raw data
2. Plot simple statistics
3. Examine distributions
4. Look at attributes together

CHECK SIMPLE STATISTICS



CHECK SIMP

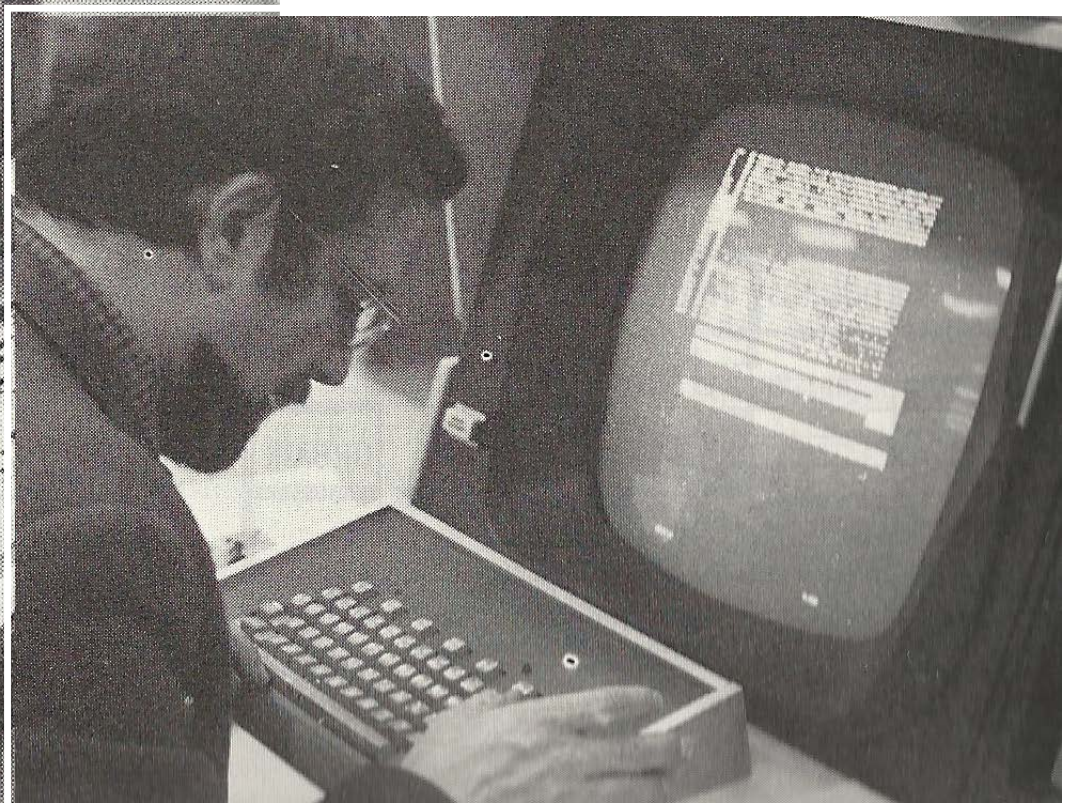
WHAT'S ONE MORE EASY THING WE SHOULD DO?



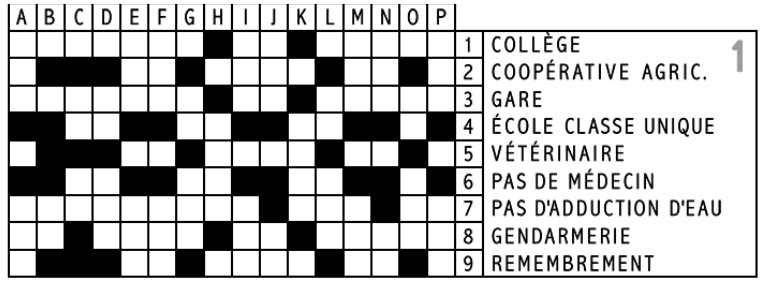
JACQUES BERTIN

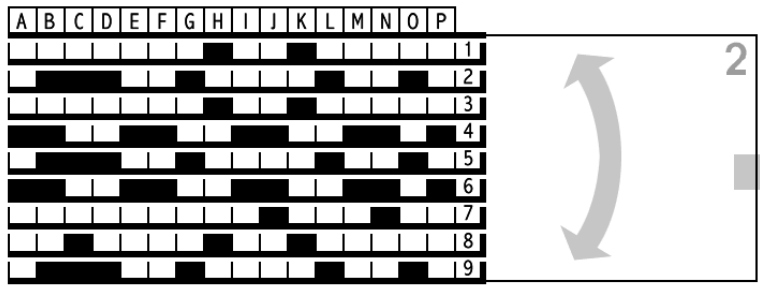
LA GRAPHIQUE ET LE TRAITEMENT GRAPHIQUE DE L'INFORMATION (1977)

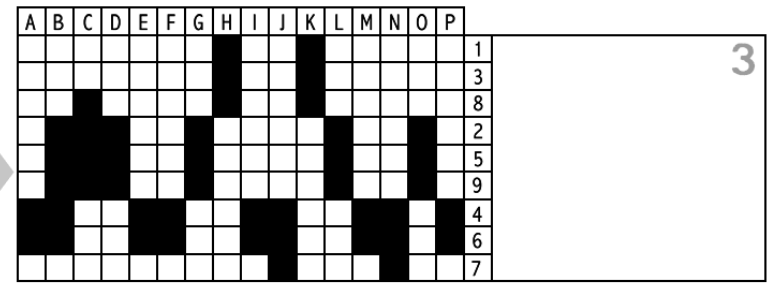
MATRIX REORGANIZATION [BERTIN 1977]

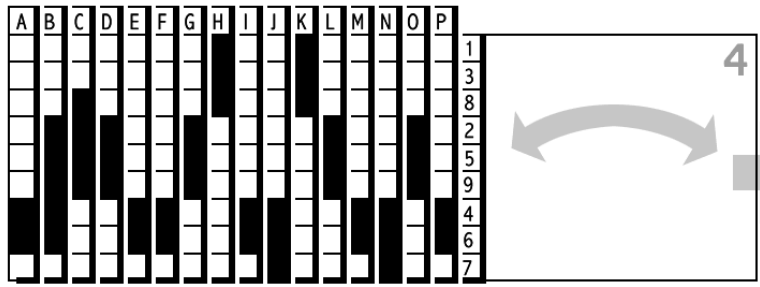


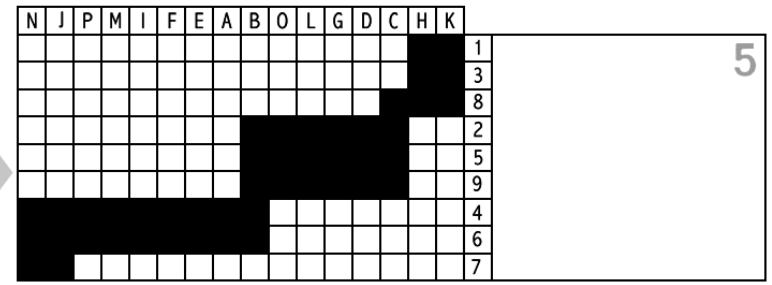
LA MATRICE ORDONNABLE











SOME FIRST STEPS TO START WITH

1. Plot the raw data
2. Plot simple statistics
3. Examine distributions
4. Look at attributes together

UNDERSTANDING DISTRIBUTIONS

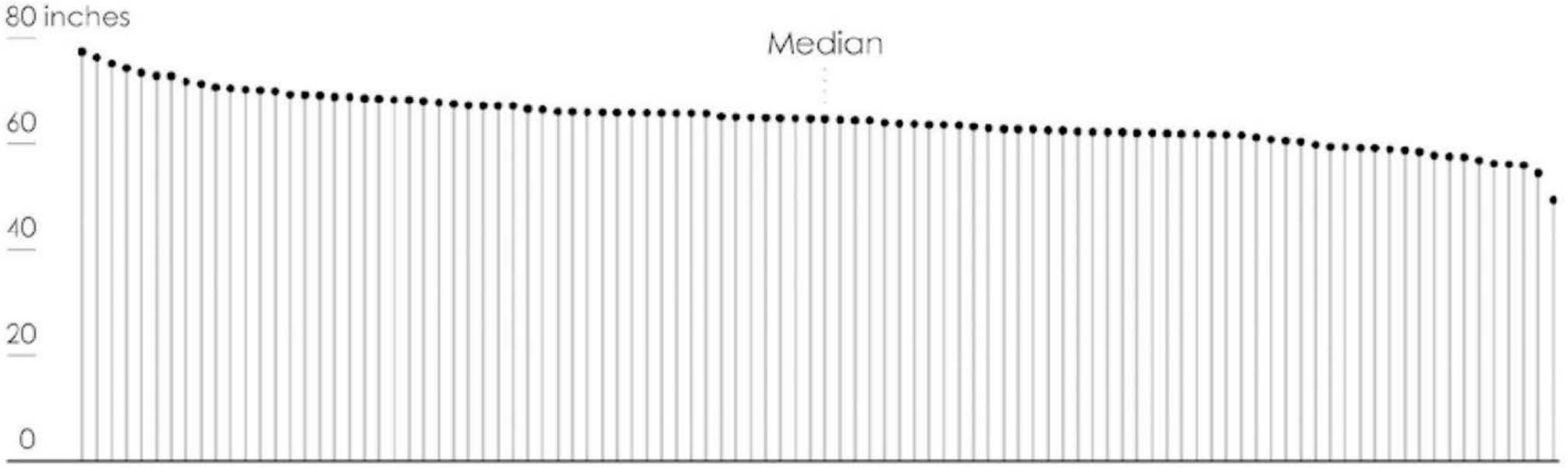
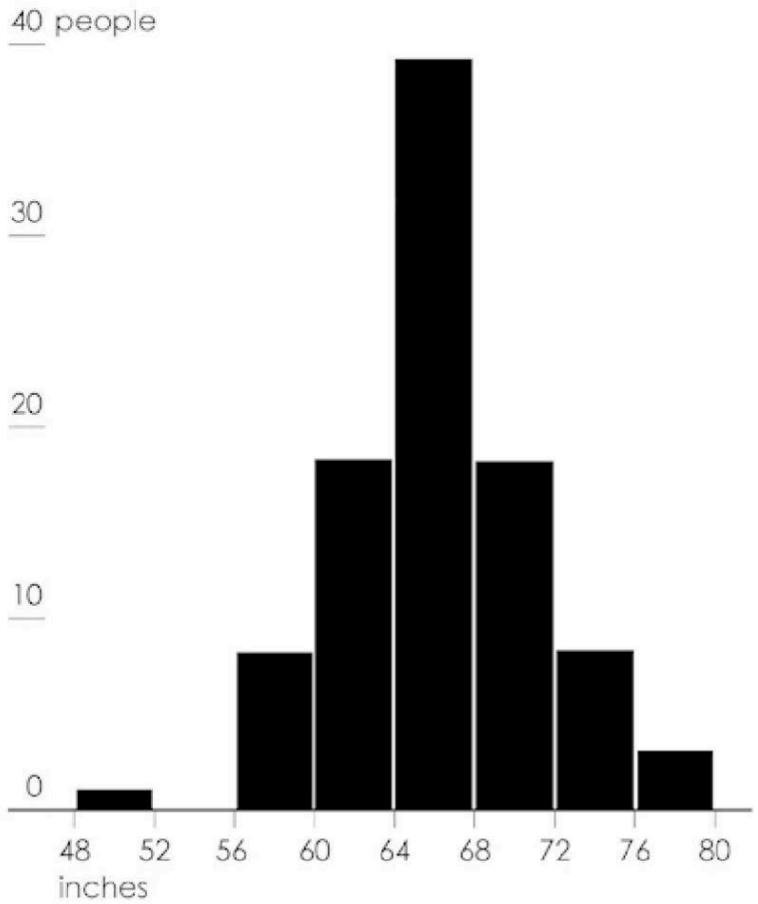
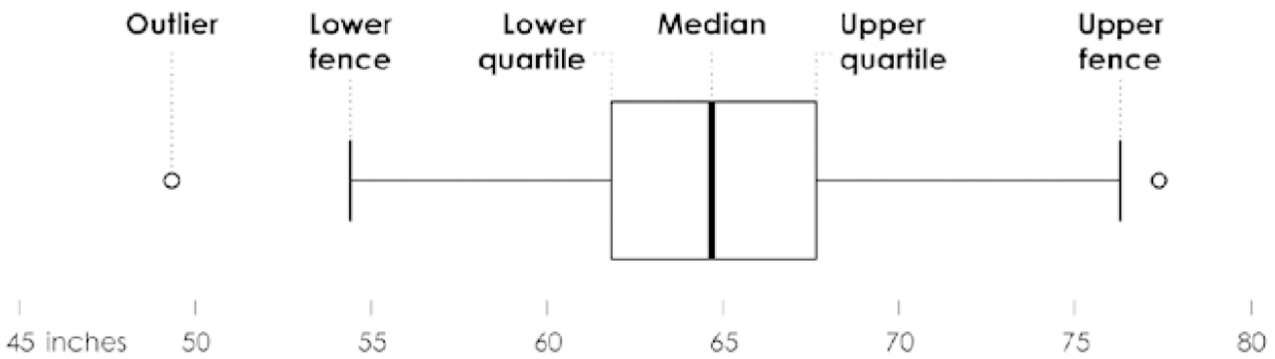


FIGURE 4-52 *Heights of imaginary people, sorted from shortest to tallest*



HISTOGRAMS

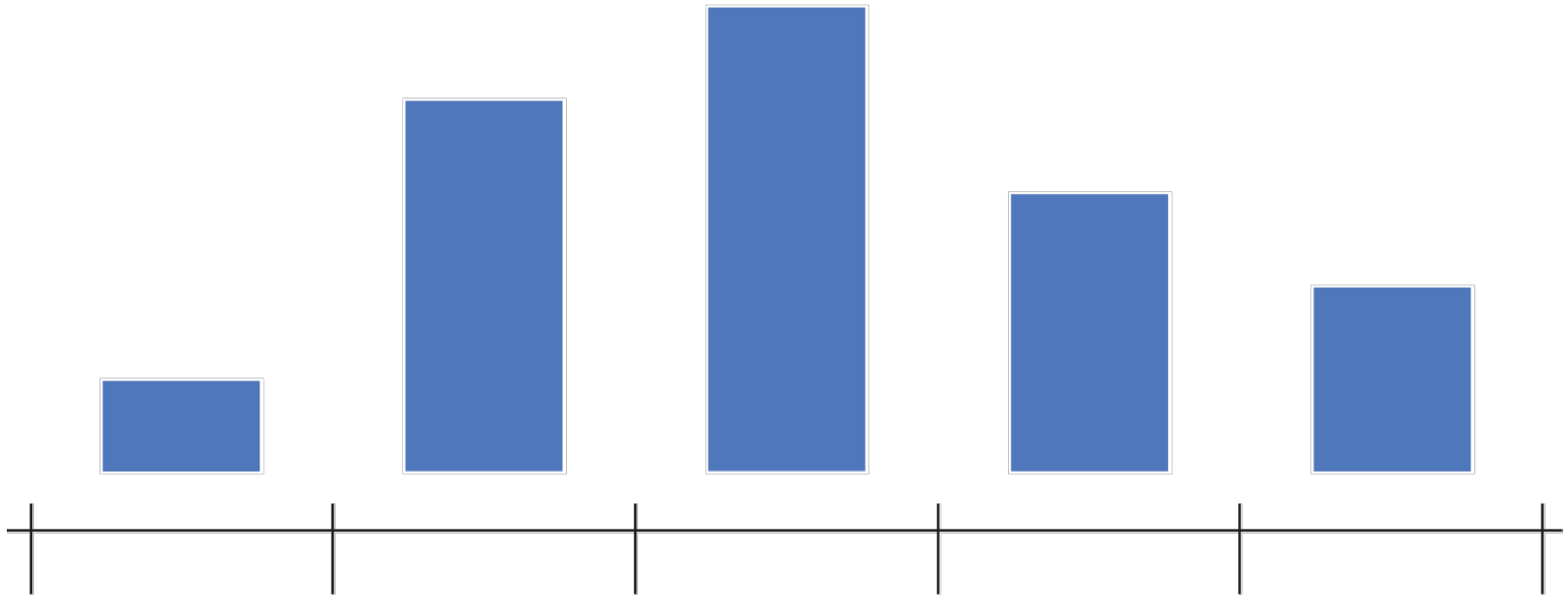


BOX PLOTS

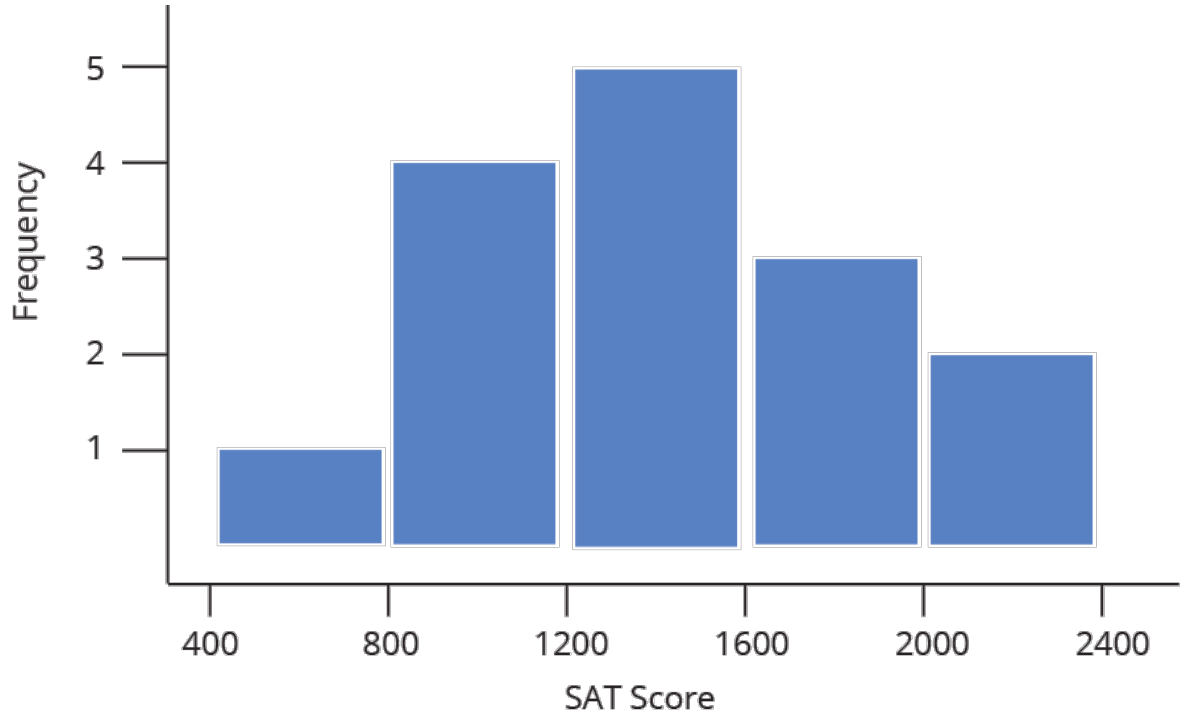
HISTOGRAMS

HISTOGRAMS

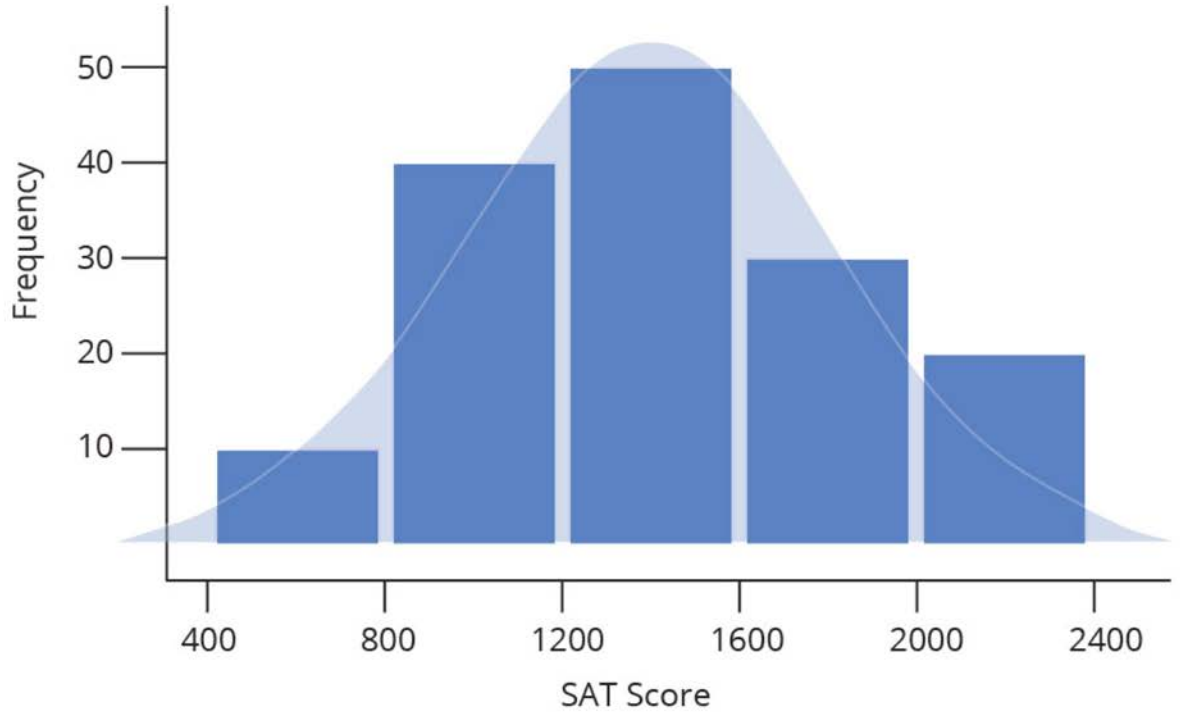
HISTOGRAMS



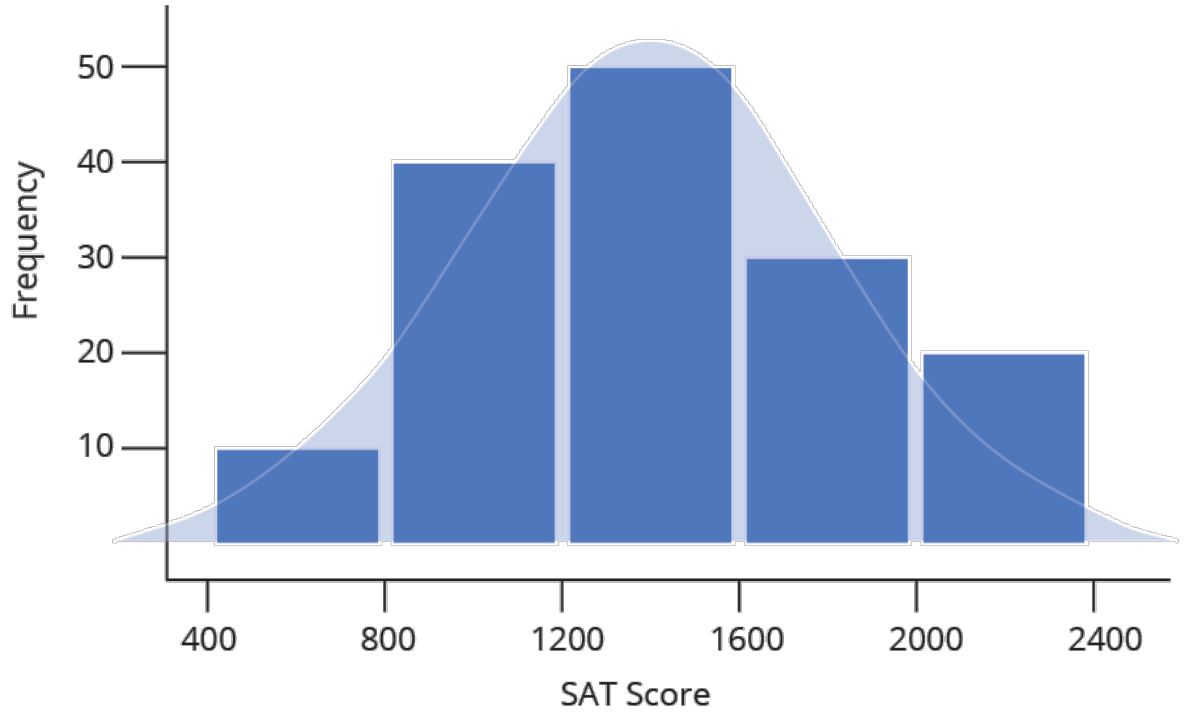
HISTOGRAMS



HISTOGRAMS

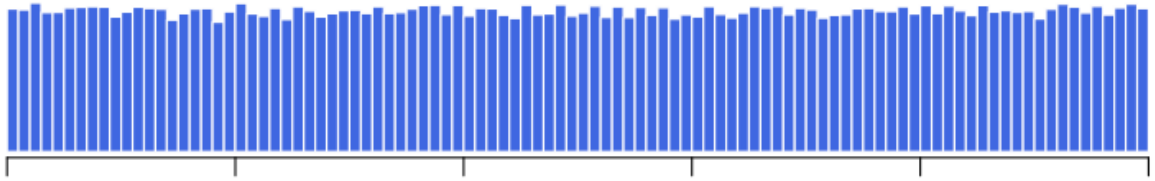


HISTOGRAMS



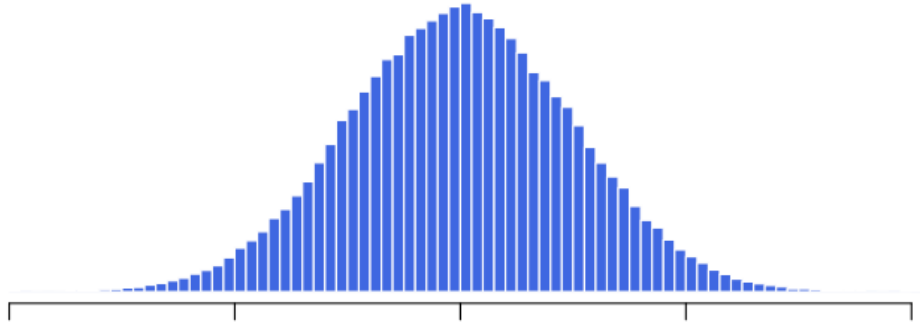
SOME SIMPLE DISTRIBUTIONS

UNIFORM



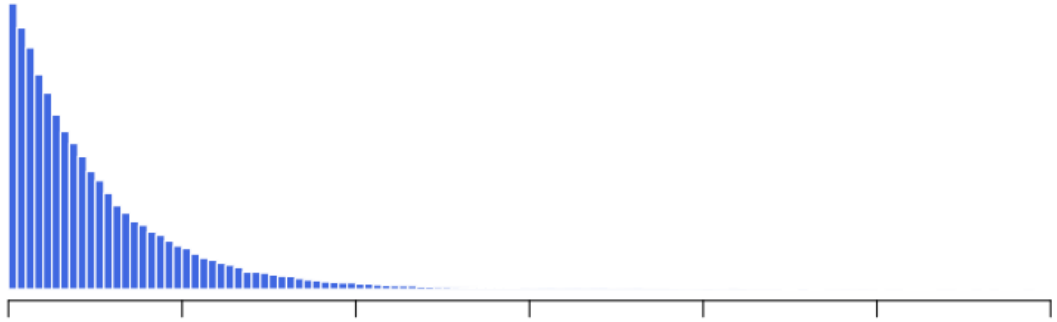
NORMAL

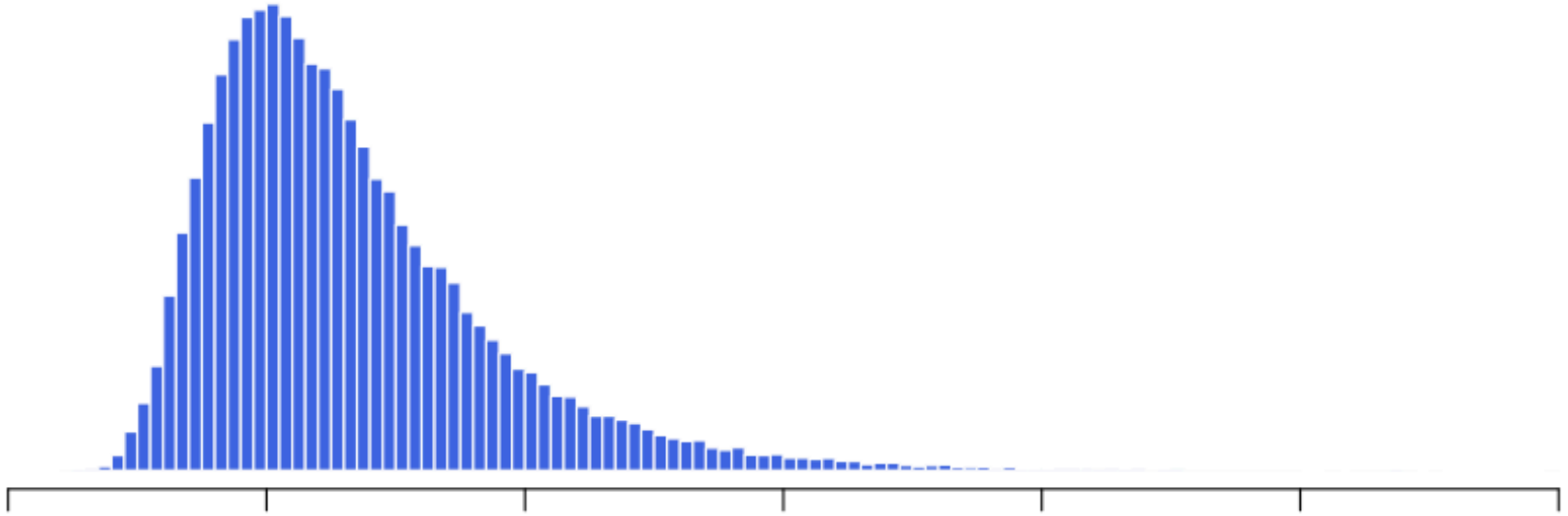
REPEATED MEASURES, VARIATION
IN POPULATIONS, ETC.



EXPONENTIAL

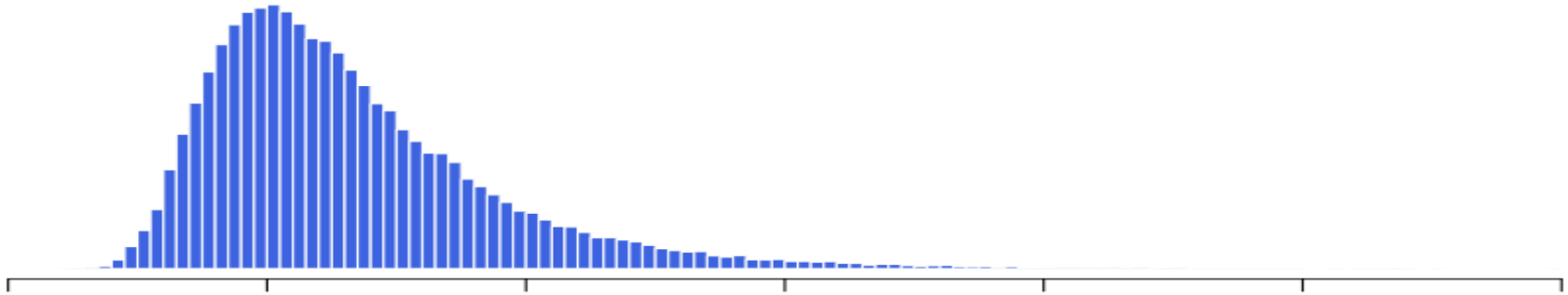
DURATIONS BETWEEN EVENTS, ETC.



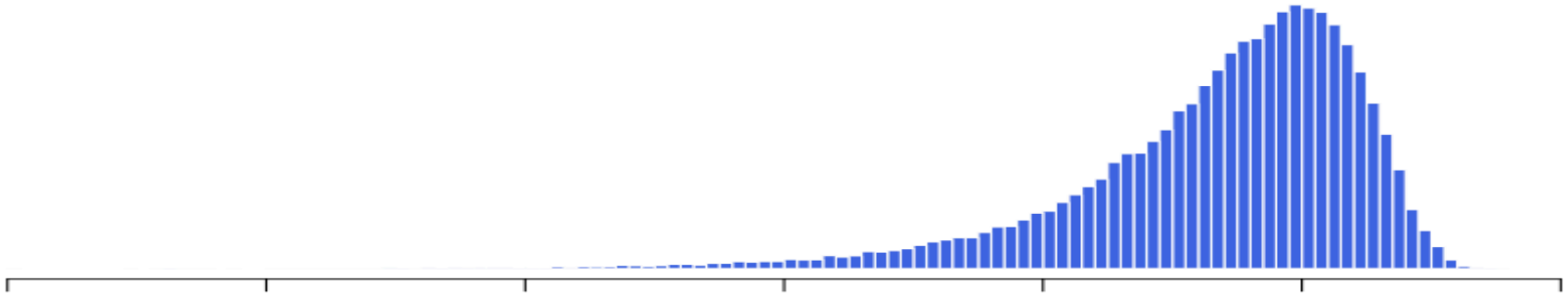


LOG-NORMAL

MANY BIOLOGICAL AND SOCIAL PROCESSES
(TISSUE GROWTH, CITY SIZE, # OF FOLLOWERS)

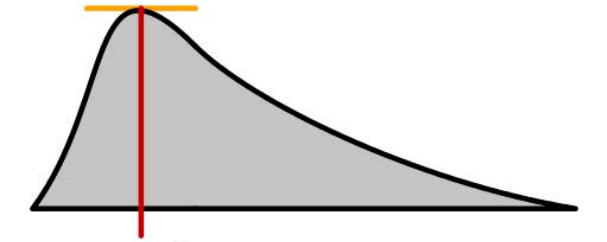


SKEWED LEFT

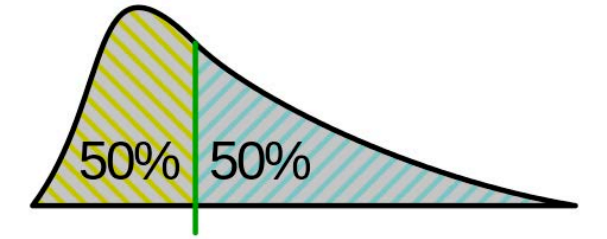


SKEWED RIGHT

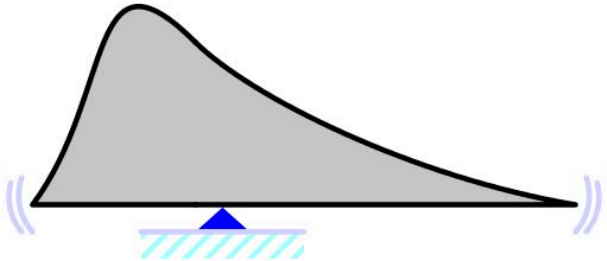
DISTRIBUTIONS AND SOME COMMON MEASURES OF CENTRAL TENDENCY



mode

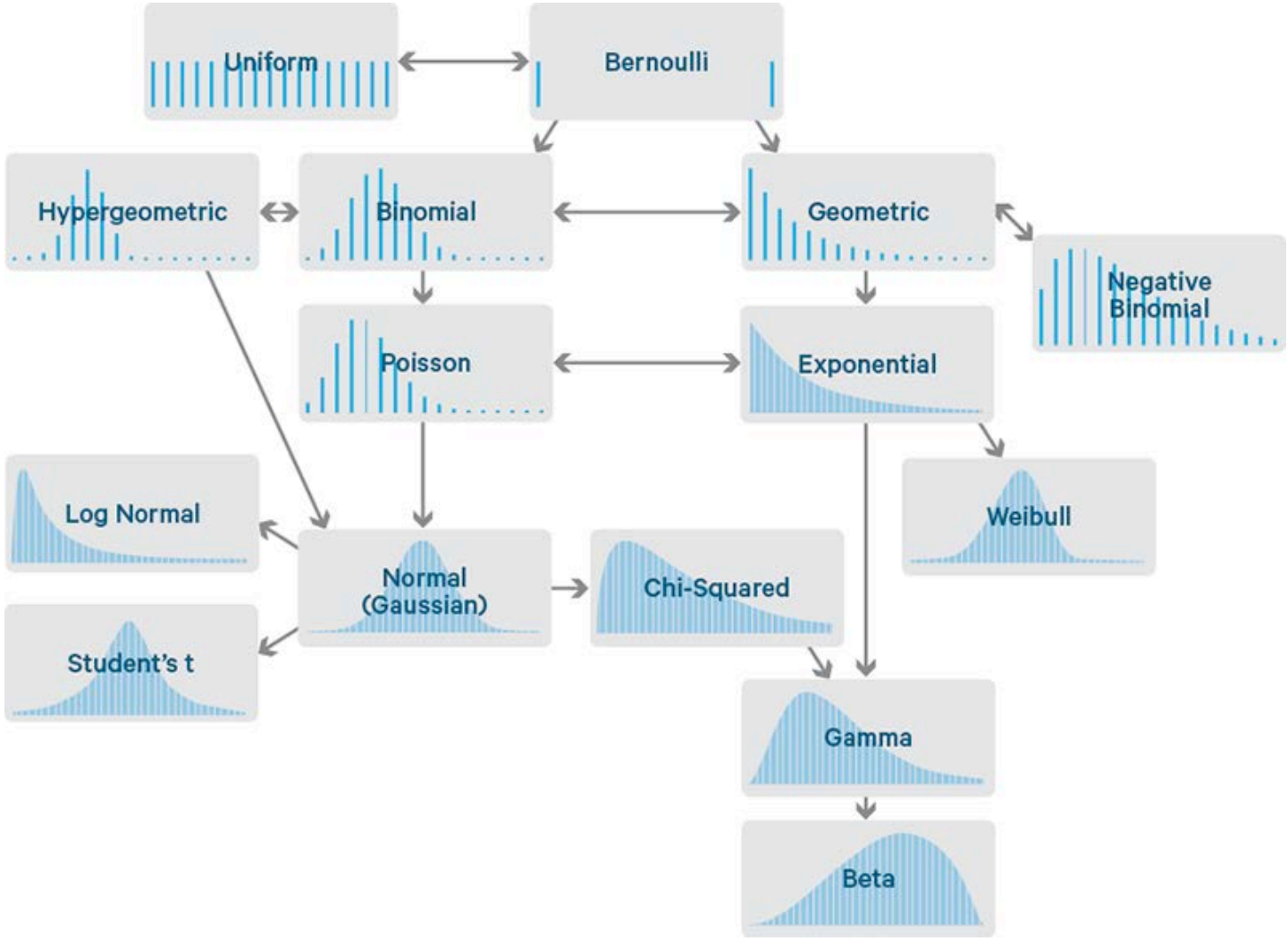


median



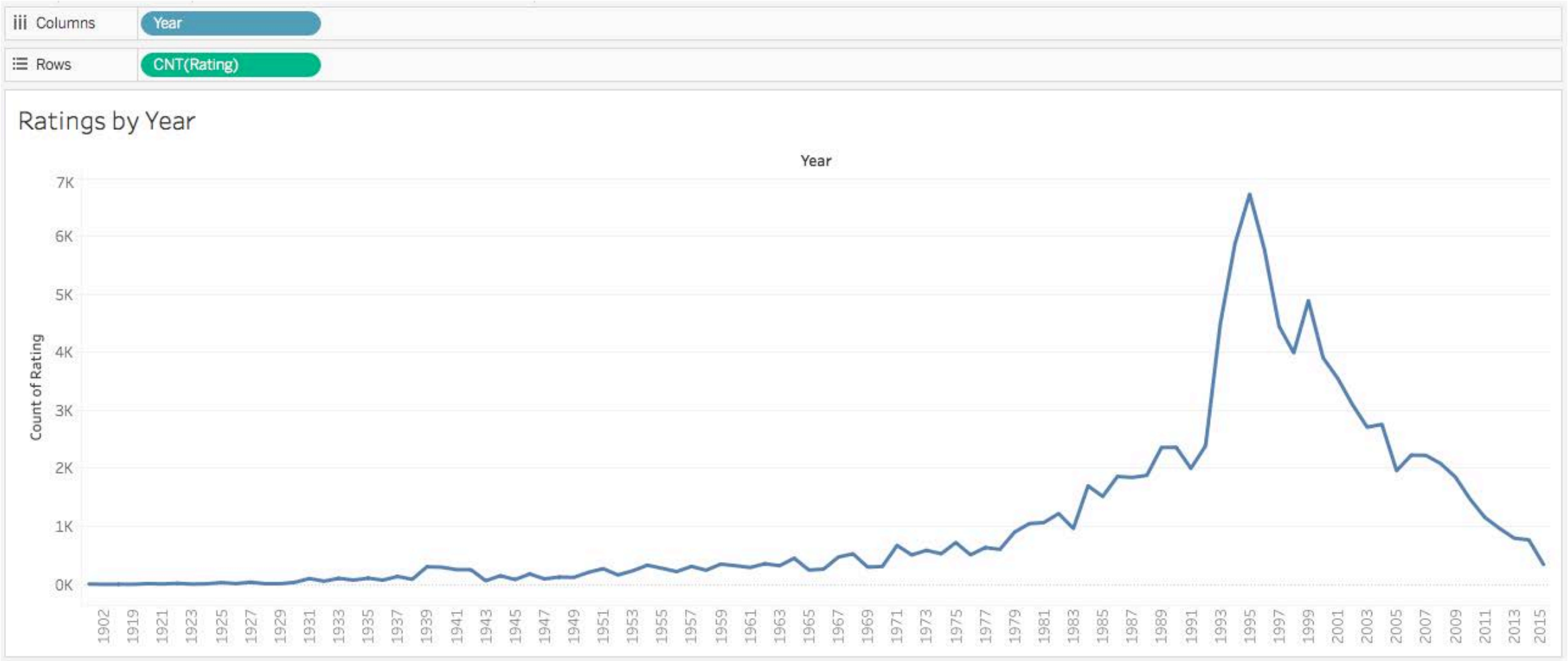
mean

COMMON DISTRIBUTIONS

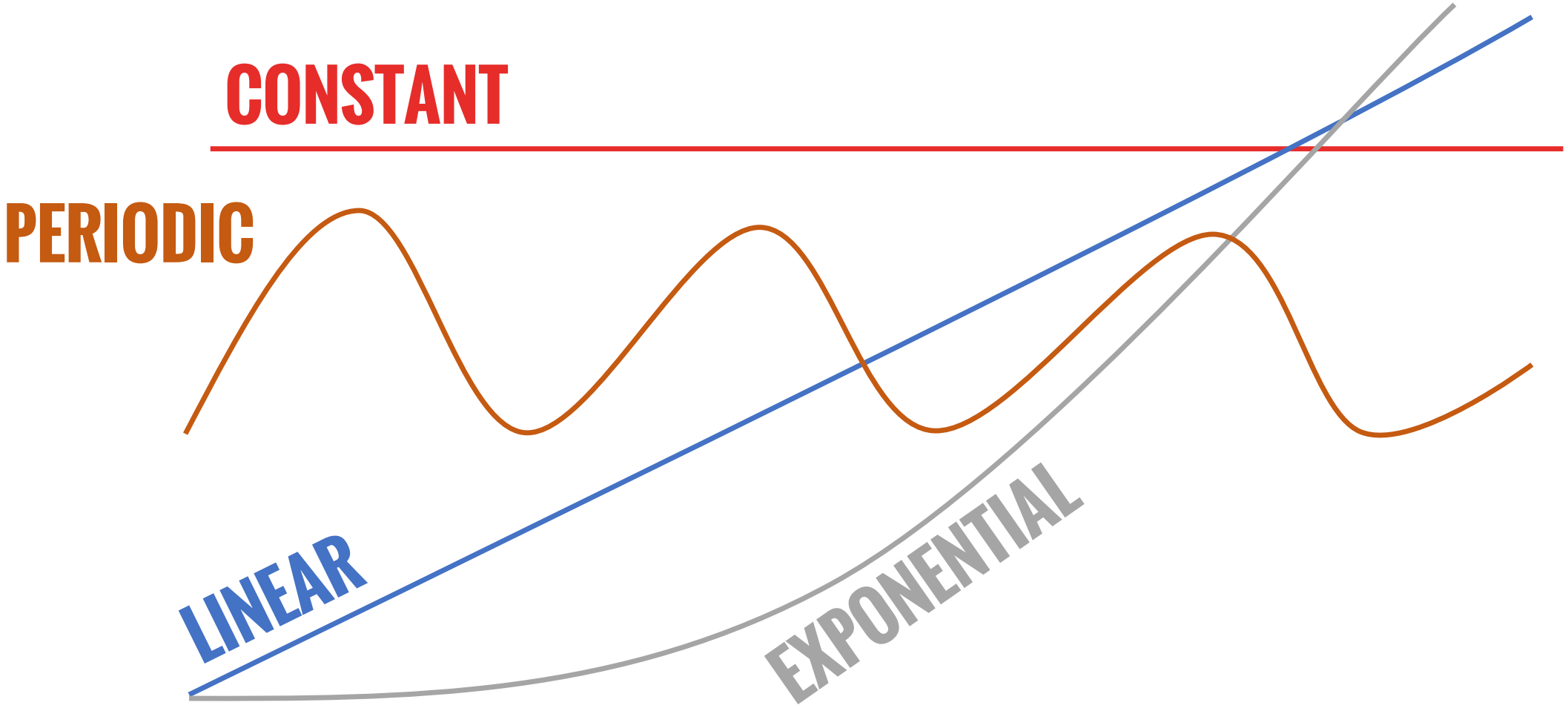


CLOUDERA

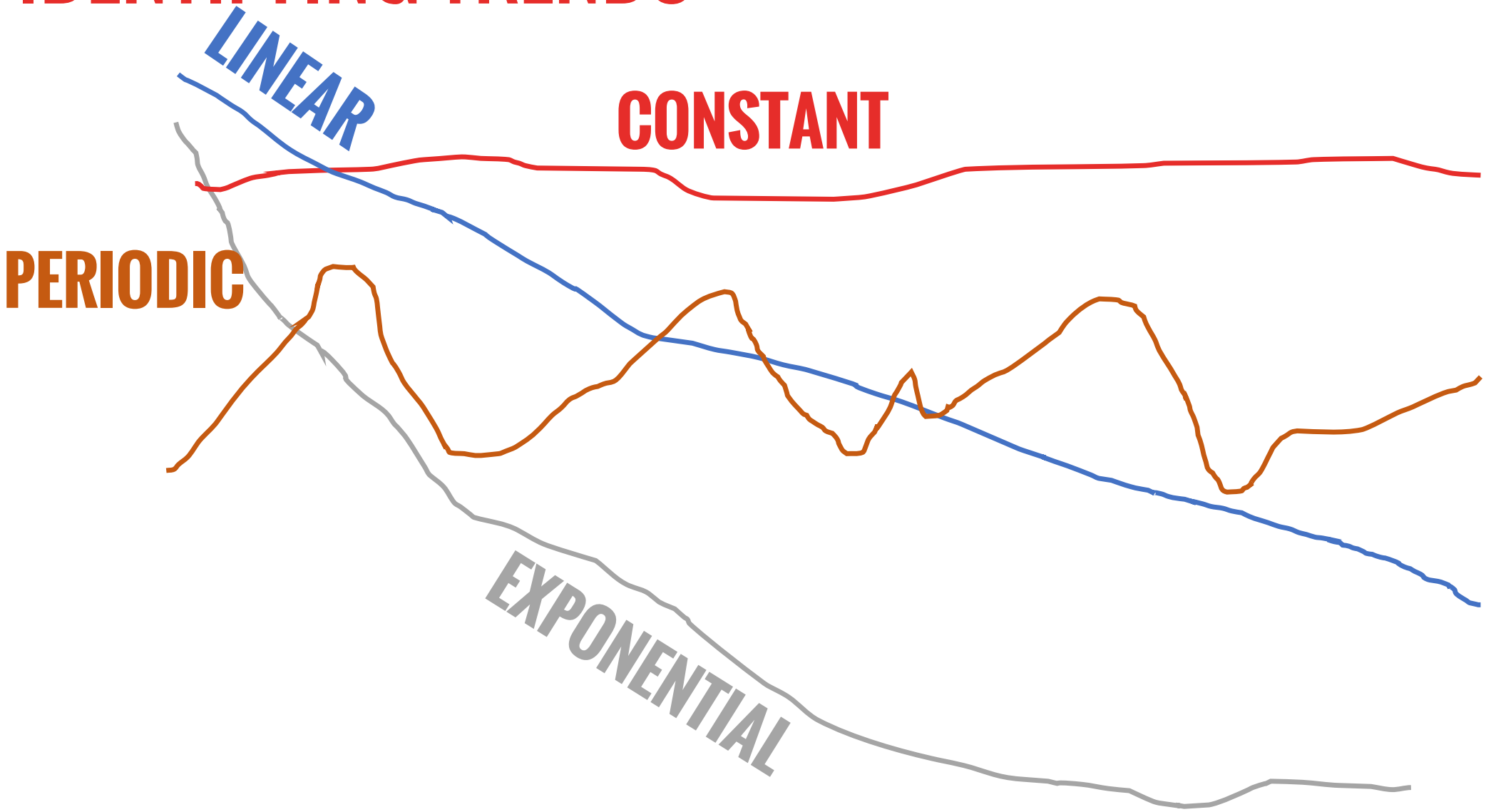
DISTRIBUTIONS OVER TIME



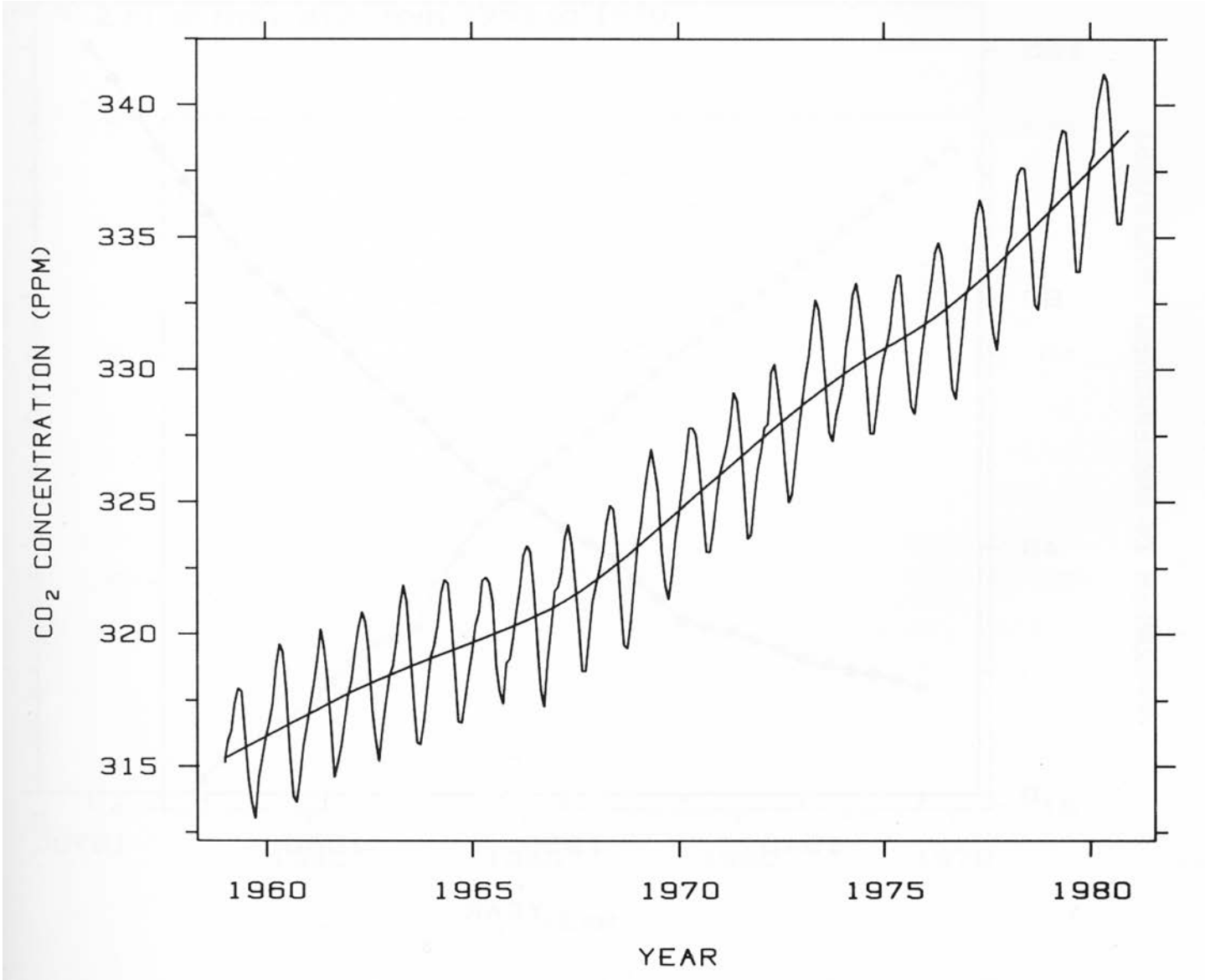
IDENTIFYING TRENDS



IDENTIFYING TRENDS

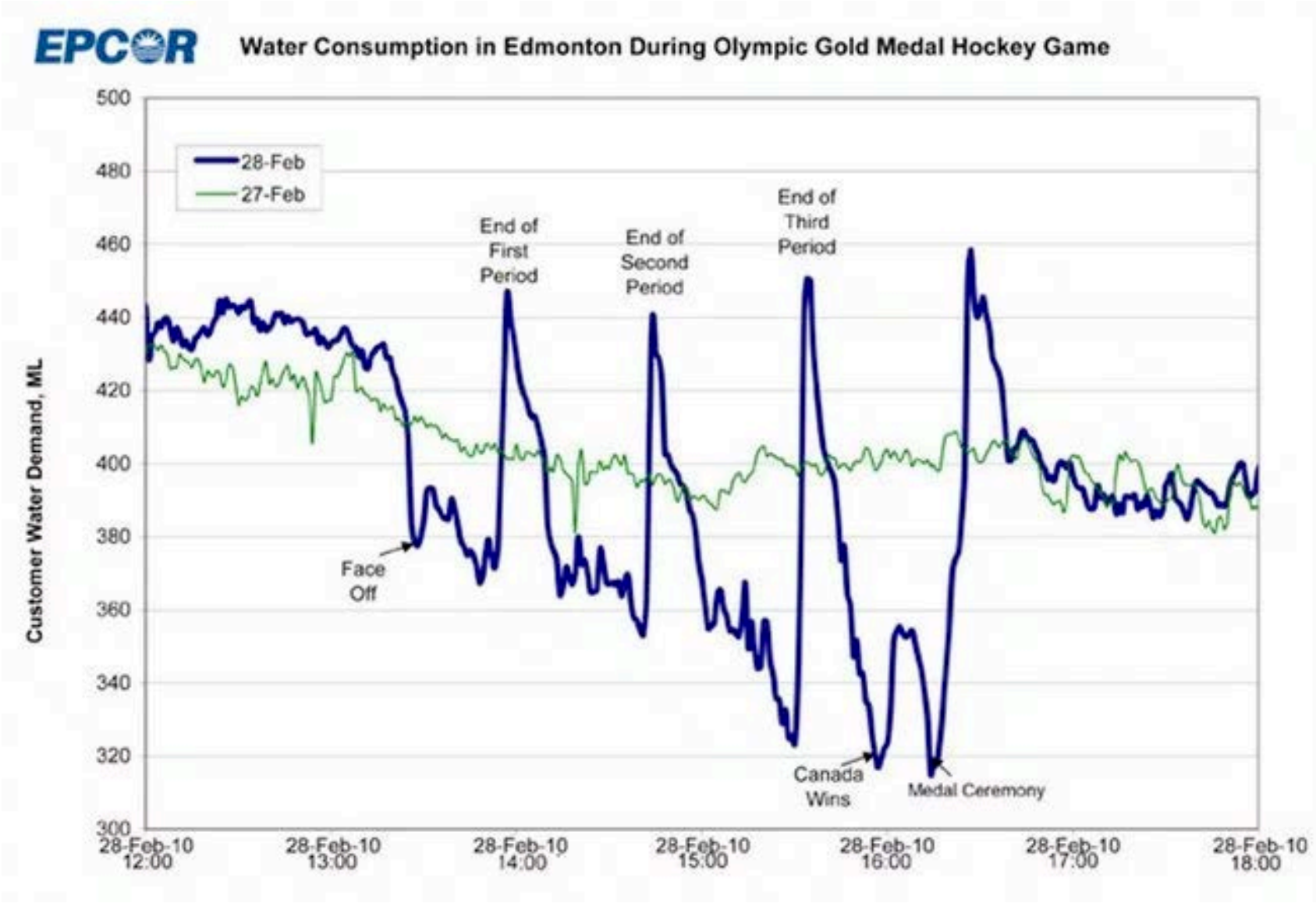


COMBINATIONS



YEARLY CO_2 CONCENTRATIONS
CLEVELAND 85

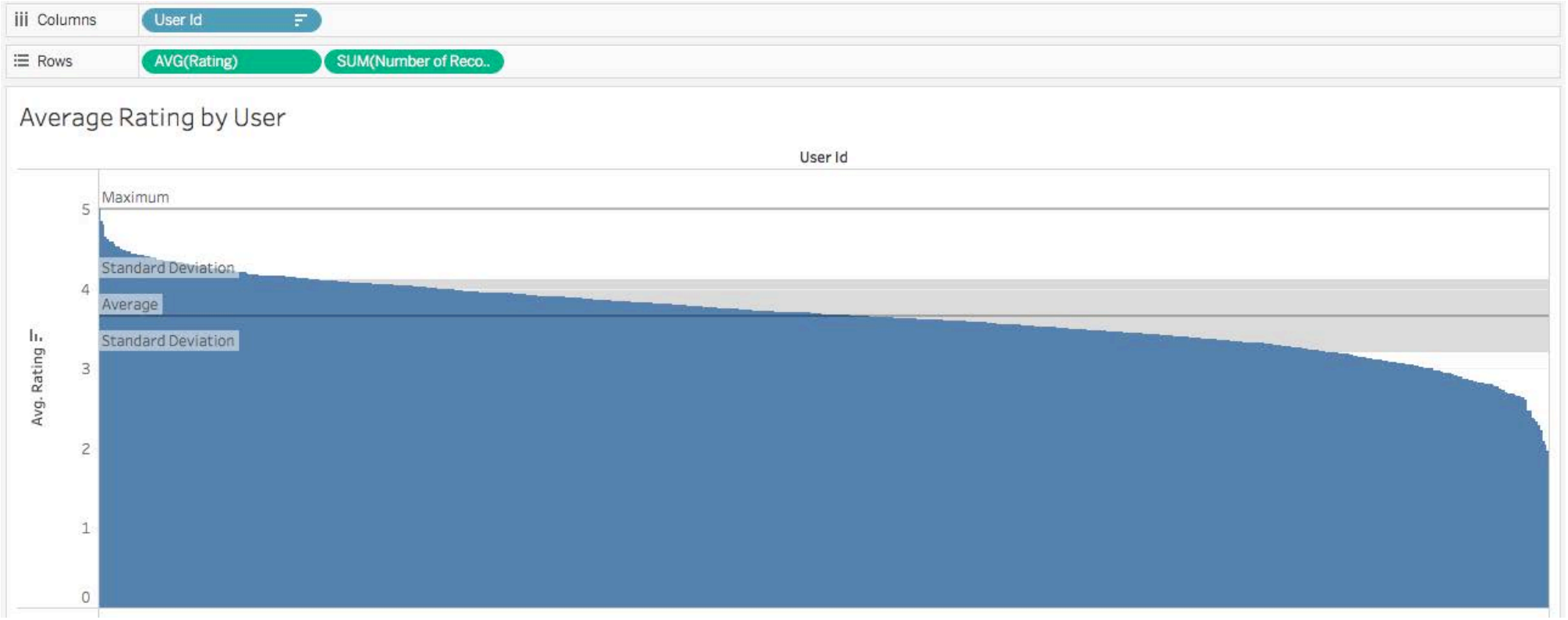
COMPARISON AGAINST A KNOWN BASELINE



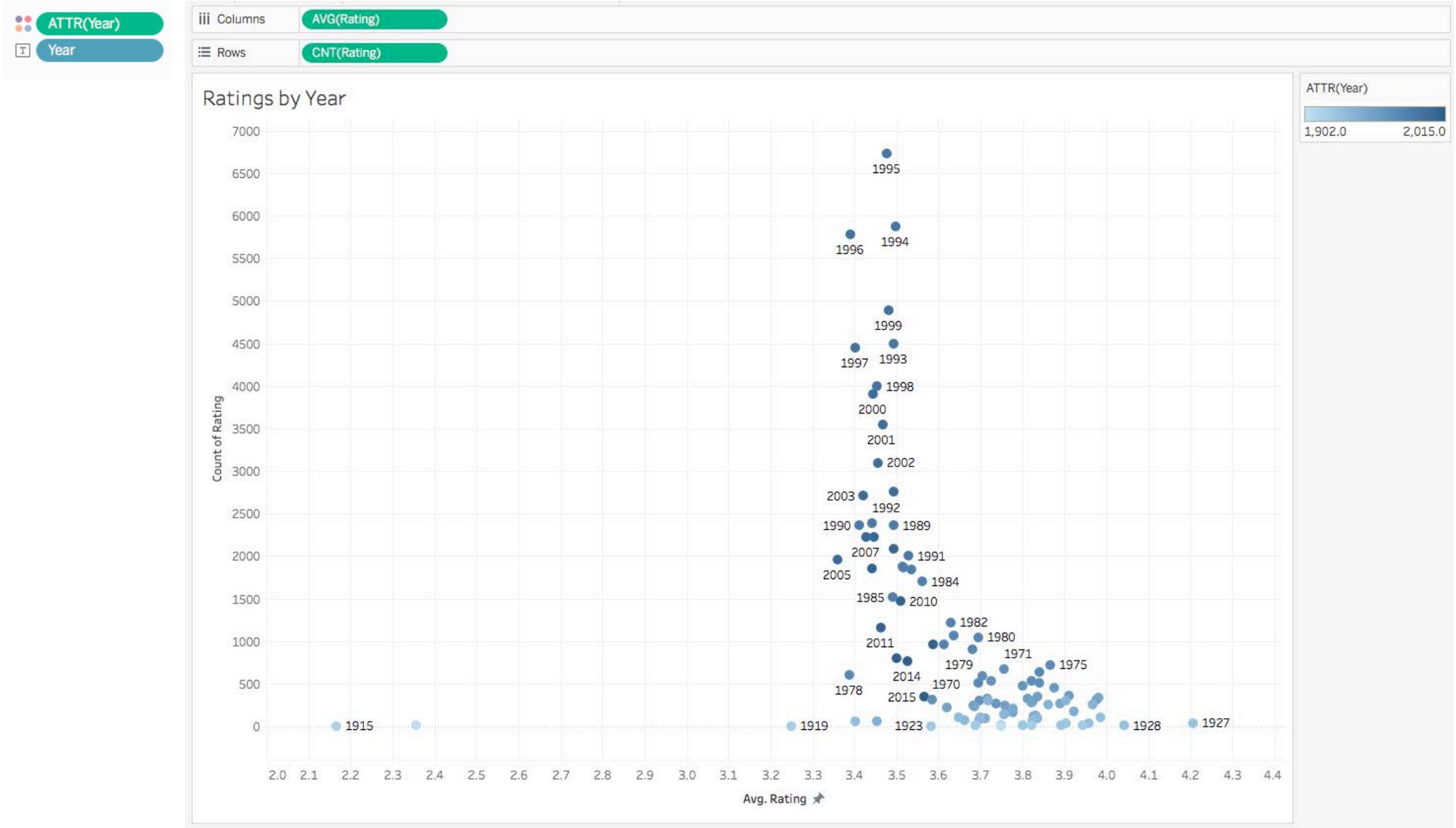
SOME FIRST STEPS TO START WITH

1. Plot the raw data
2. Plot simple statistics
3. Examine distributions
4. **Look at attributes together**

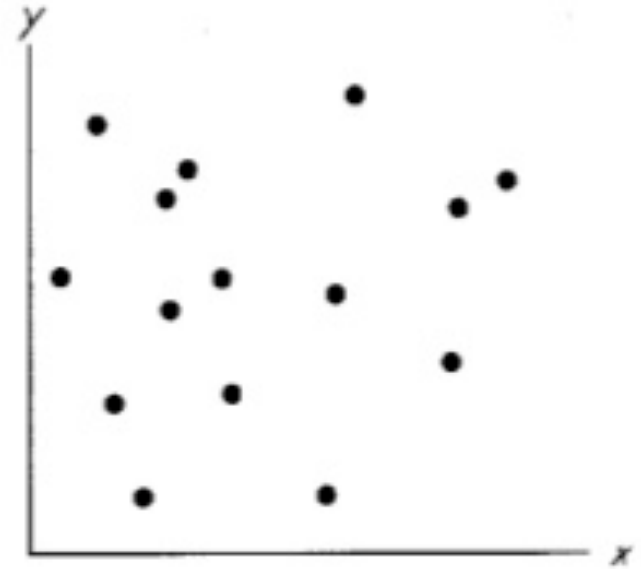
COMPARE MULTIPLE PLOTS



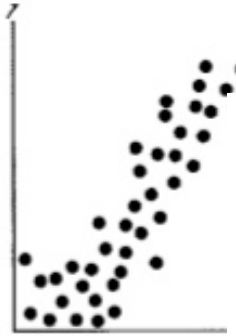
COMPARING TWO MEASURES



IDENTIFYING CORRELATIONs

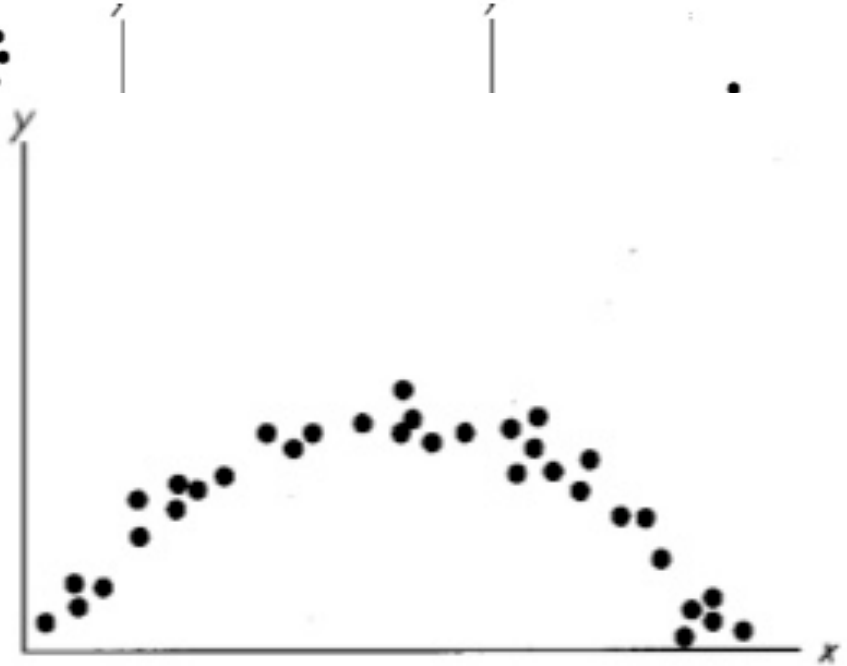


(g) No correlation between x and y



(a) Positive corre between x and y

(d) Negative cor between x and y



(h) Nonlinear correlation between x and y

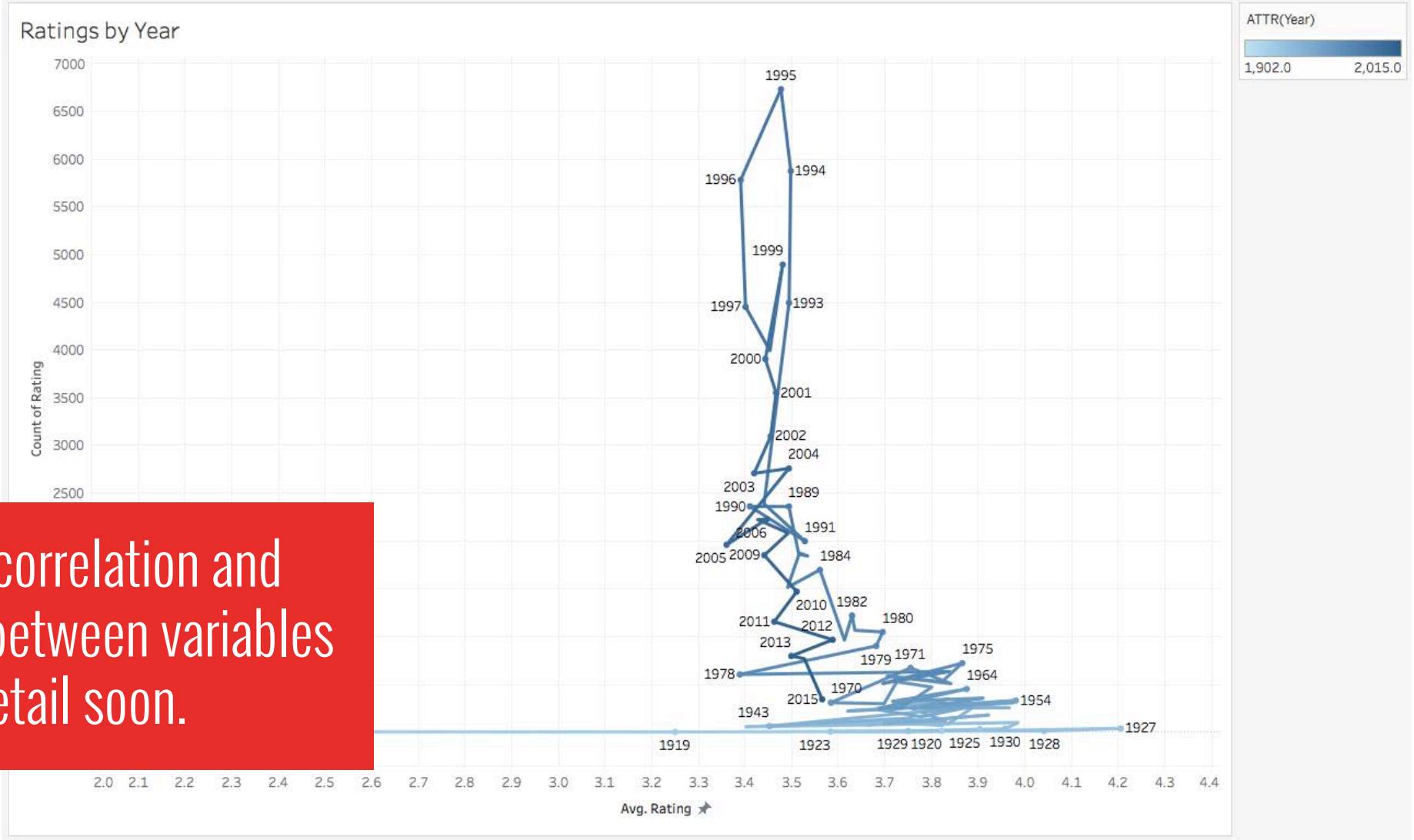
correlation between x and y

correlation between x and y

COMPARING TWO MEASURES

ATTR(Year)
Year

Columns AVG(Rating)
Rows CNT(Rating)

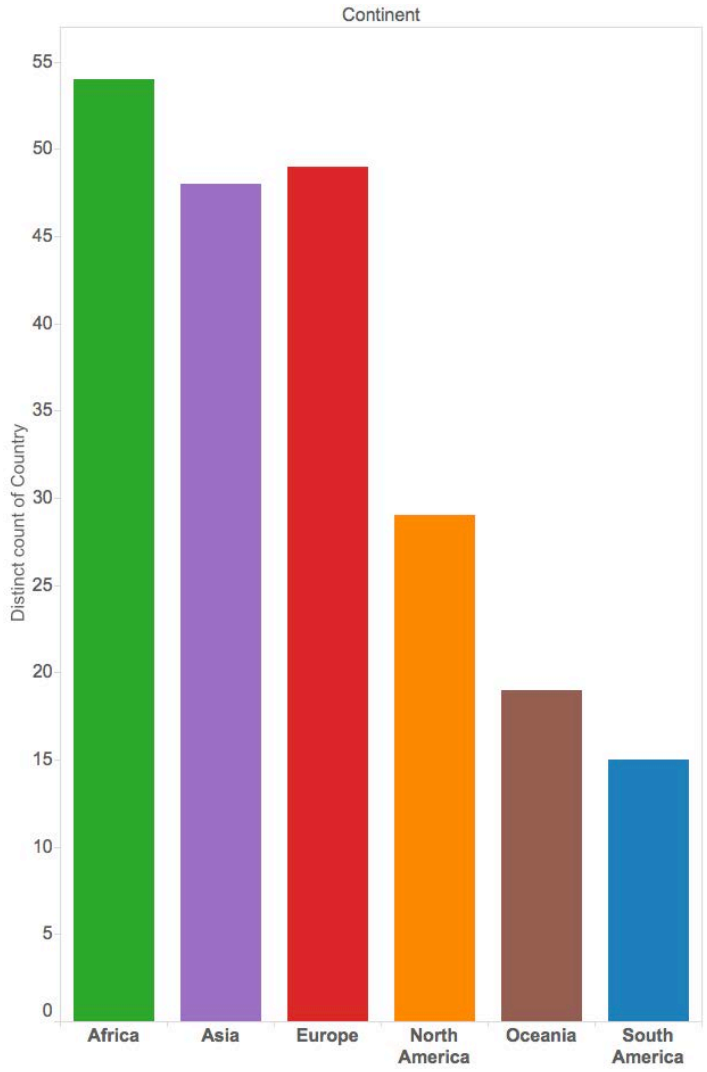


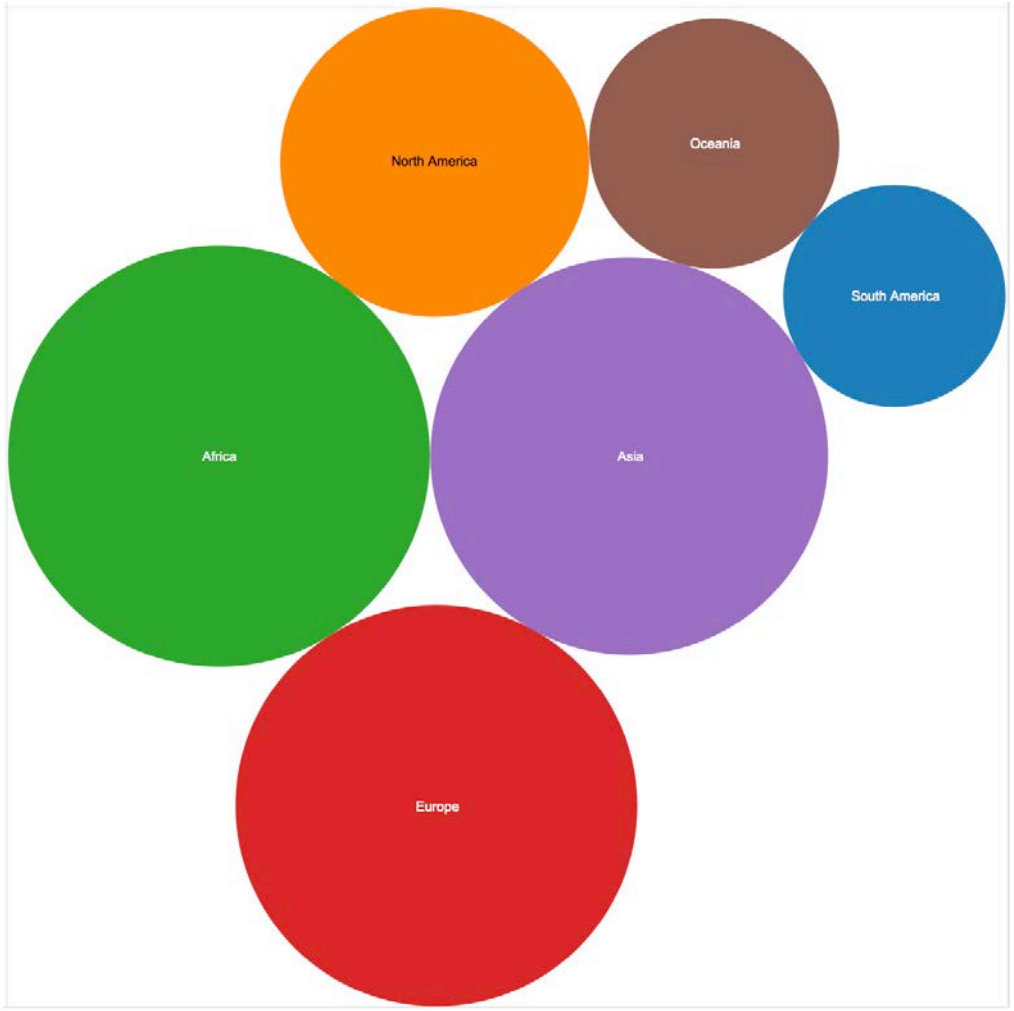
ANY HYPOTHESES?

We will revisit correlation and Relationships between variables In a lot more detail soon.

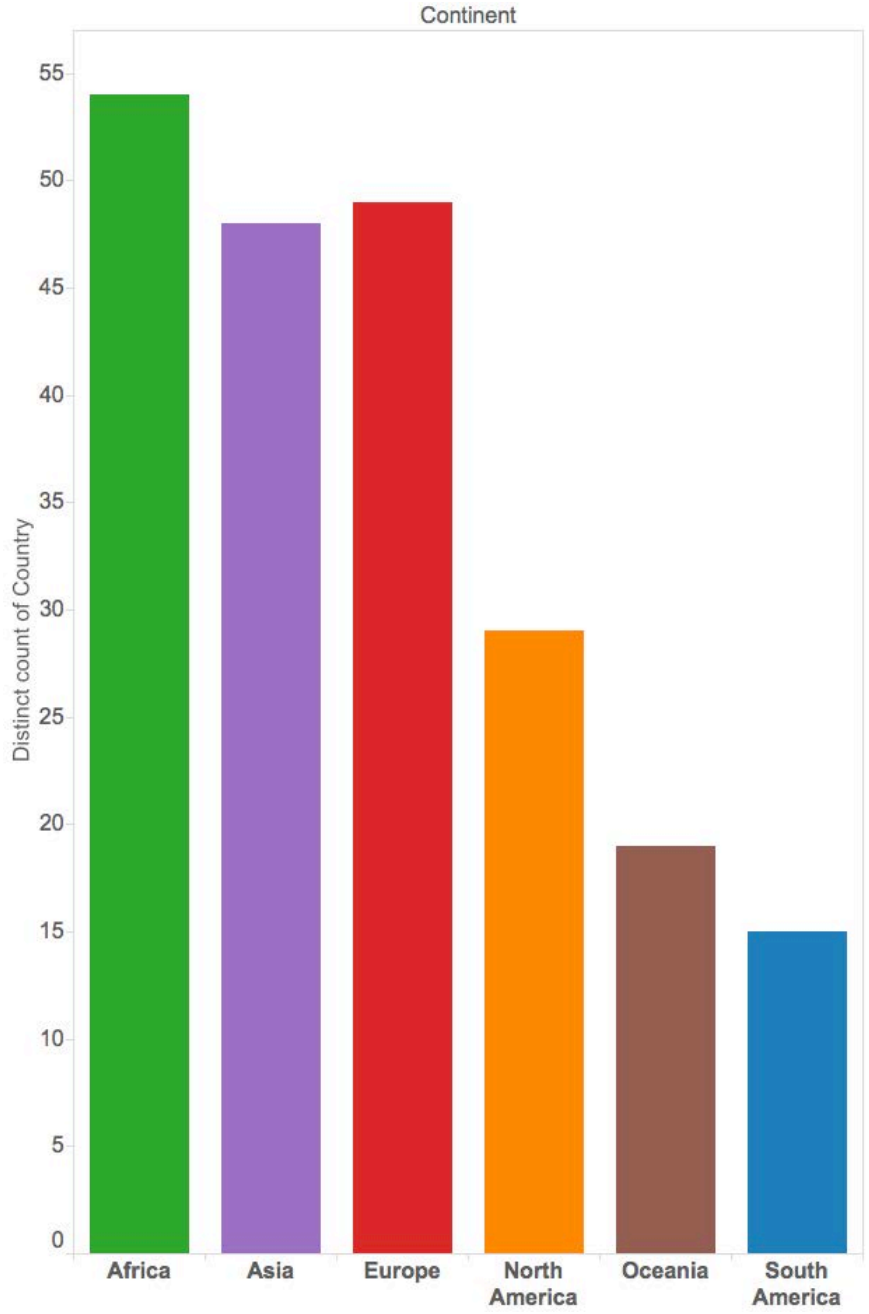
SOME COMMON VISUAL MAPPING MISTAKES

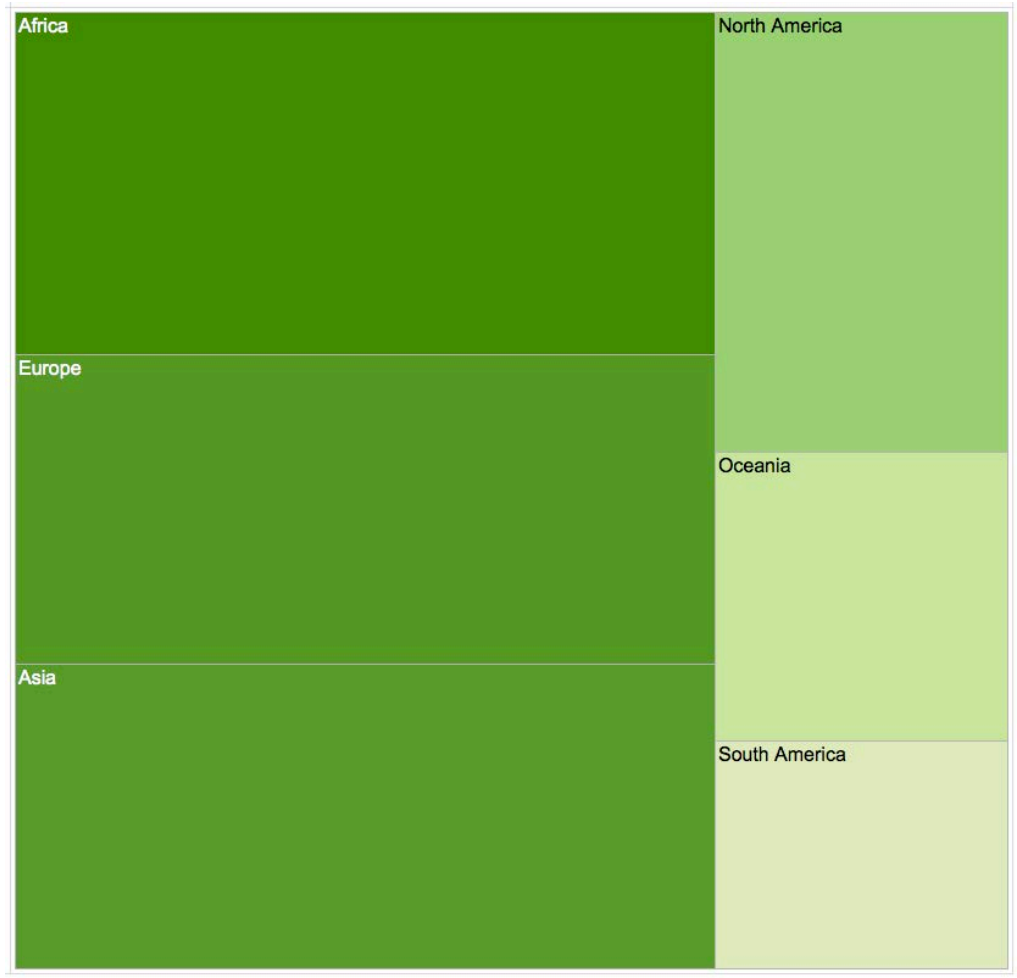
COMPARING QUANTITIES



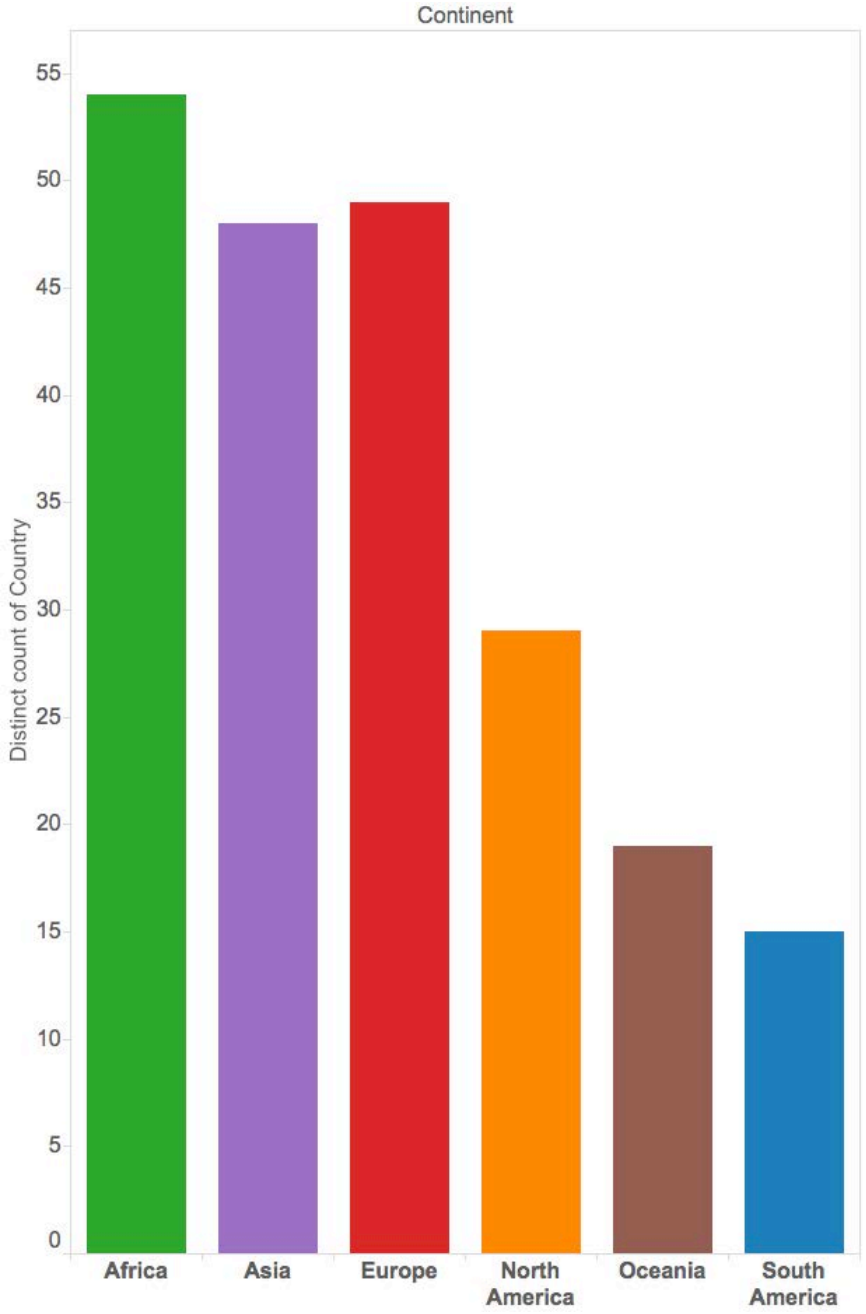

North America
Oceania
South America
Africa
Asia
Europe

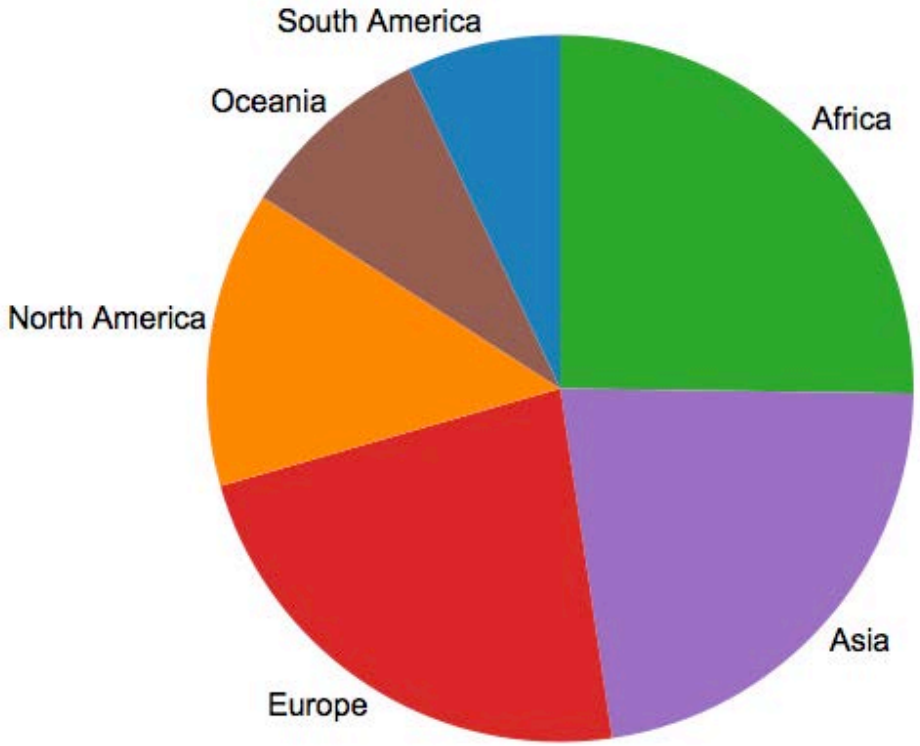
VS


Continent
Distinct count of Country
55
50
45
40
35
30
25
20
15
10
5
0
Africa
Asia
Europe
North America
Oceania
South America

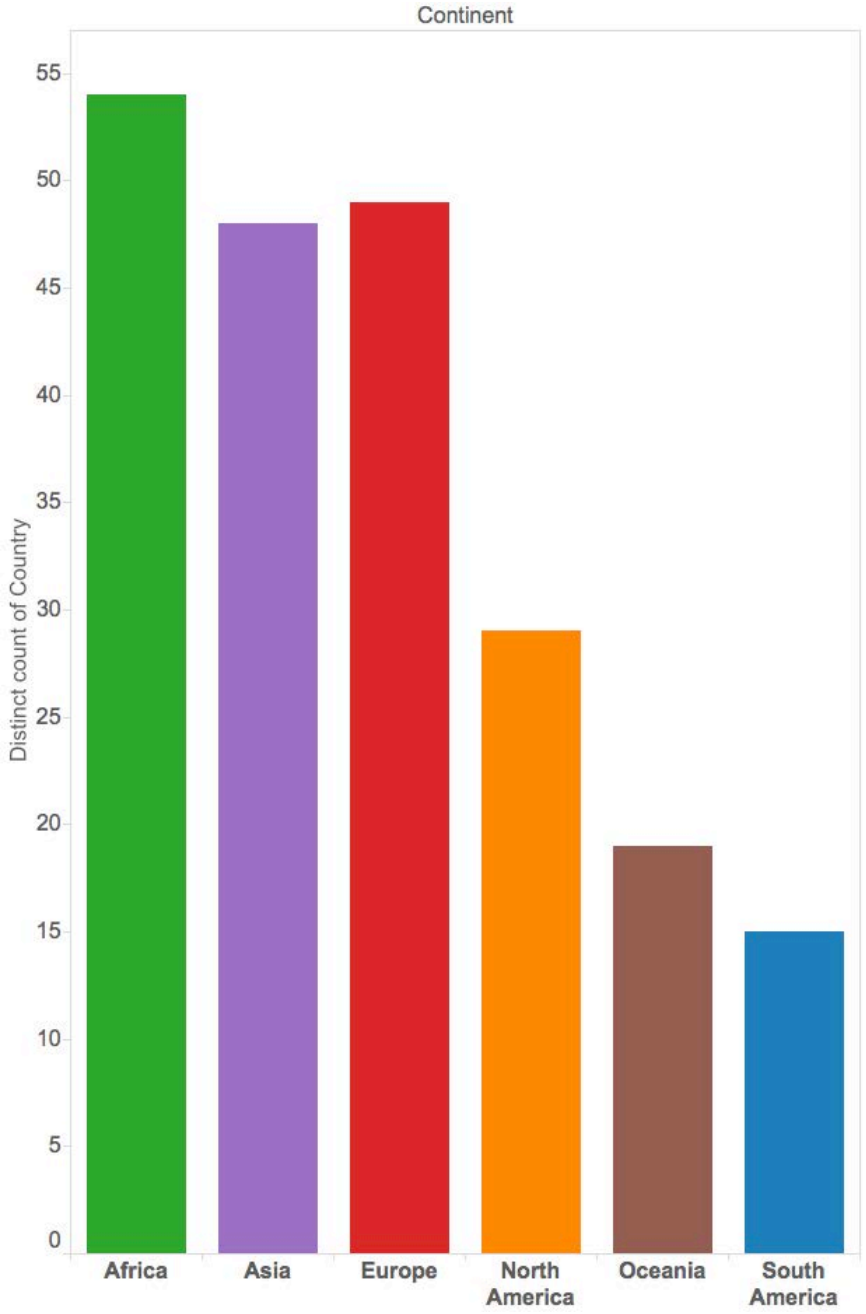

Africa
North America
Europe
Oceania
Asia
South America

VS

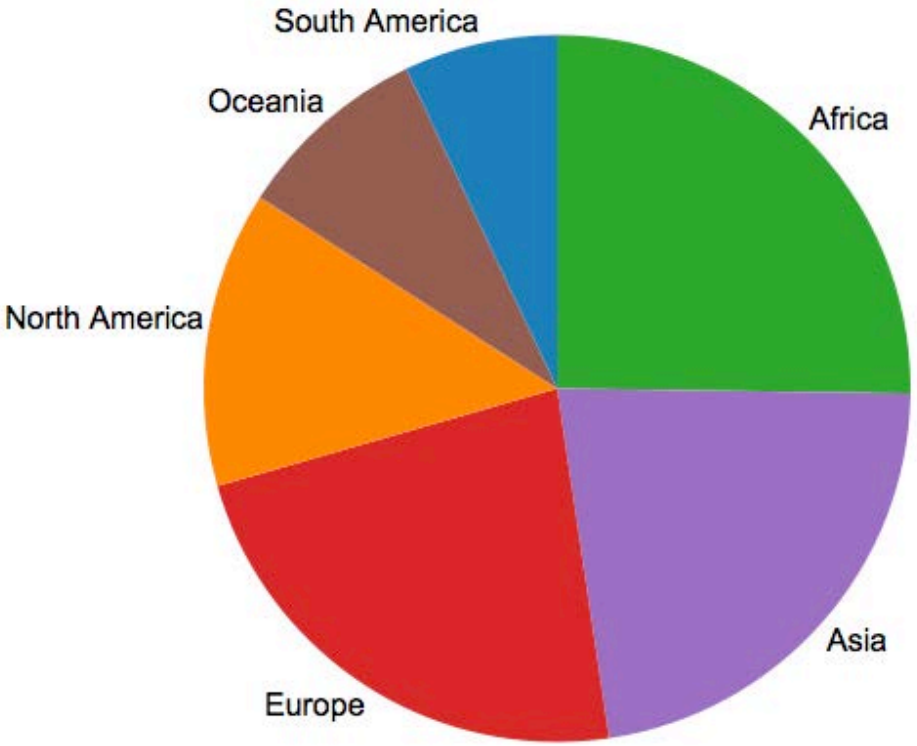

Continent
Distinct count of Country
55
50
45
40
35
30
25
20
15
10
5
0
Africa
Asia
Europe
North America
Oceania
South America


South America
Oceania
Africa
North America
Asia
Europe

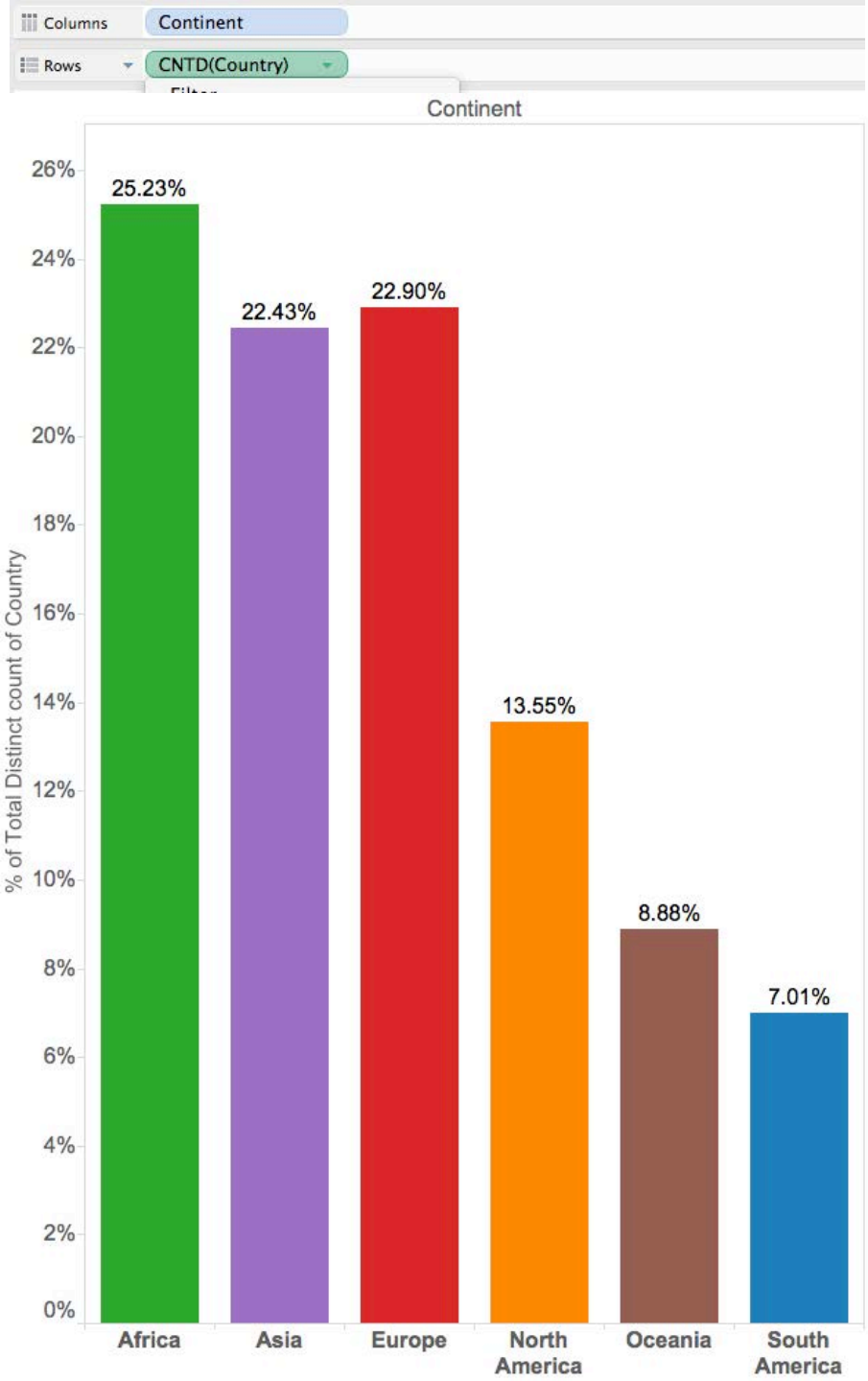
VS

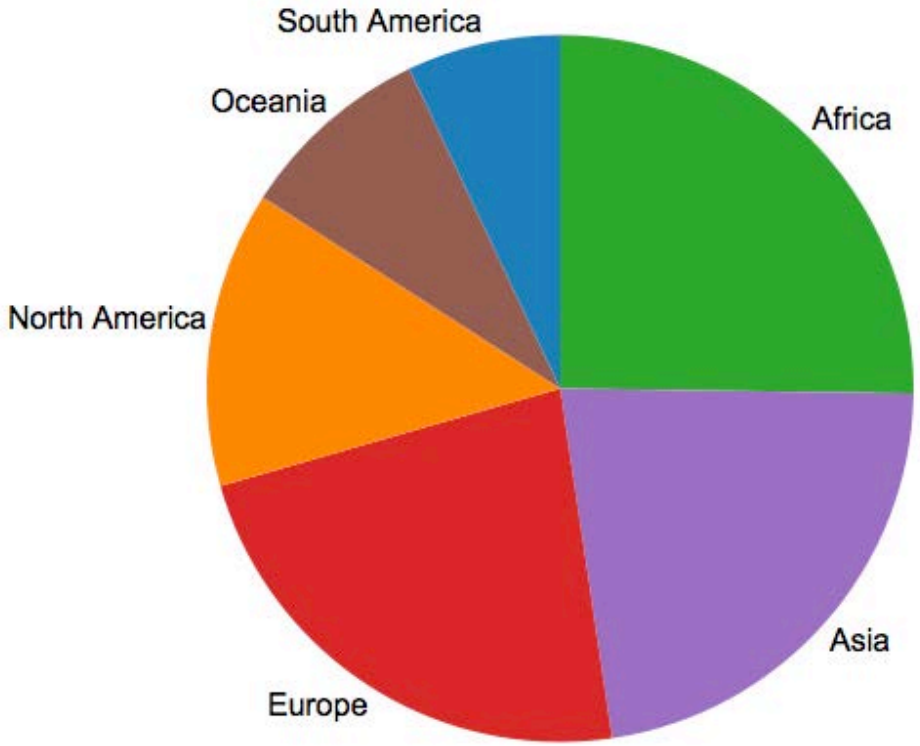

Continent
Distinct count of Country
55
50
45
40
35
30
25
20
15
10
5
0
Africa
Asia
Europe
North America
Oceania
South America

PARTS-OF-A-WHOLE RELATIONSHIPS

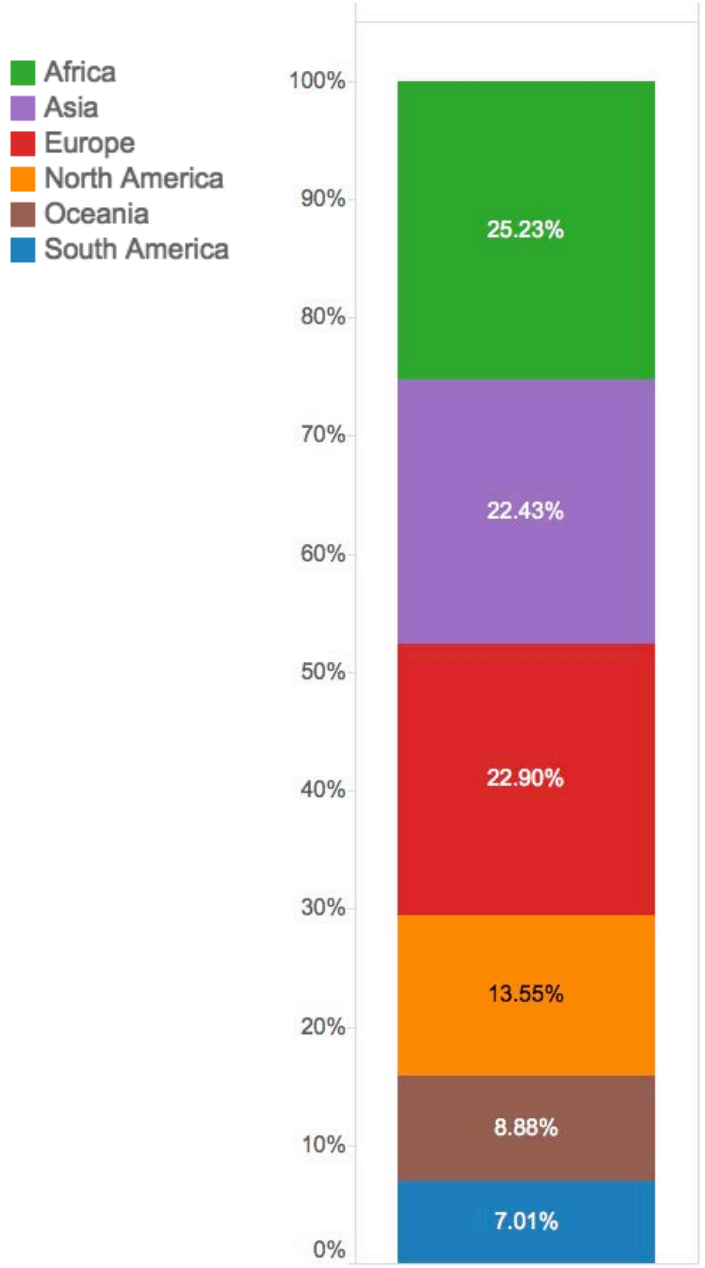

South America
Oceania
Africa
North America
Asia
Europe

VS

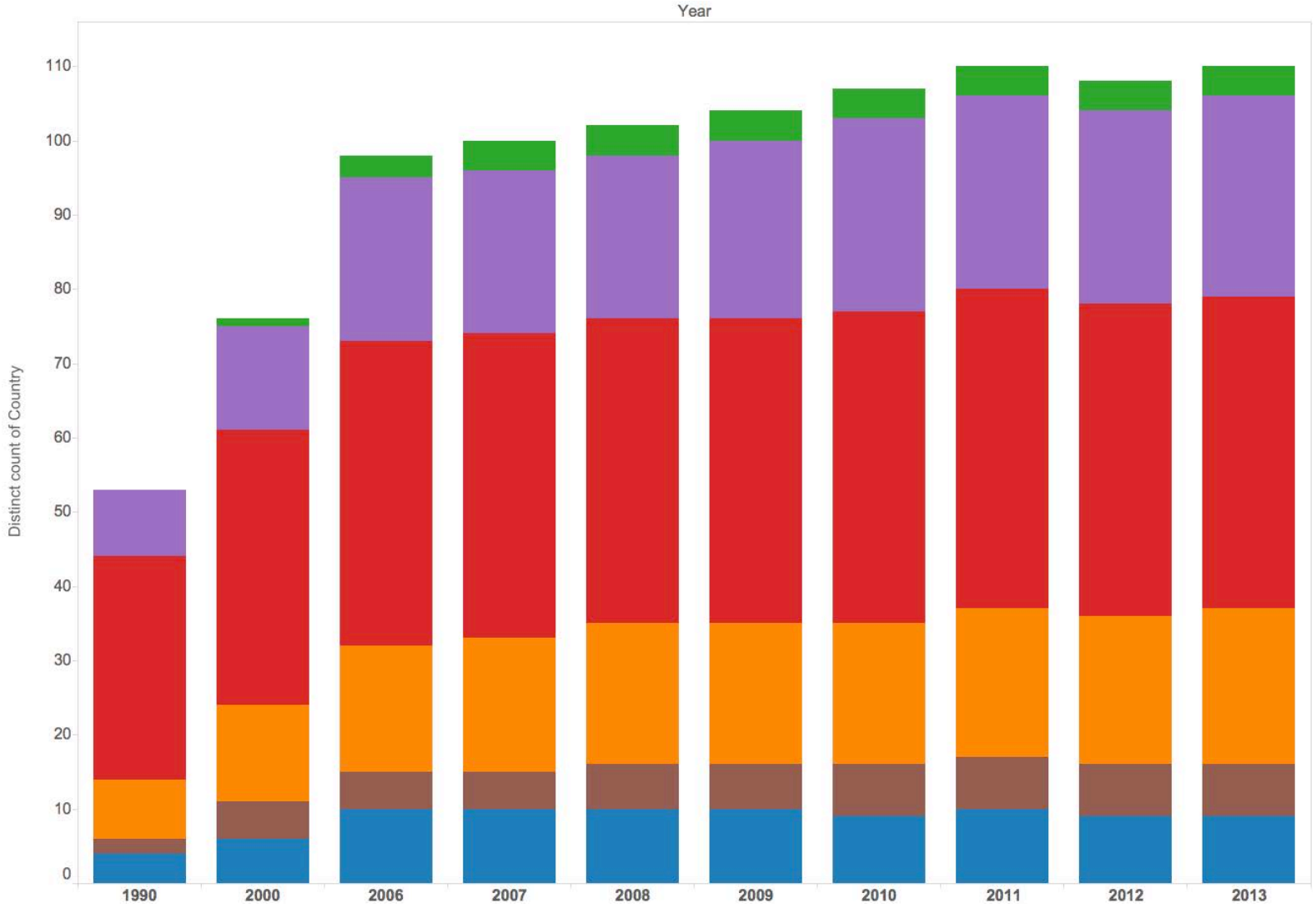

Columns
Continent
Rows
CNTD(Country)
Continent
% of Total Distinct count of Country
26%
24%
22%
20%
18%
16%
14%
12%
10%
8%
6%
4%
2%
0%
25.23%
22.43%
22.90%
13.55%
8.88%
7.01%
Africa
Asia
Europe
North America
Oceania
South America


South America
Oceania
Africa
North America
Asia
Europe

VS

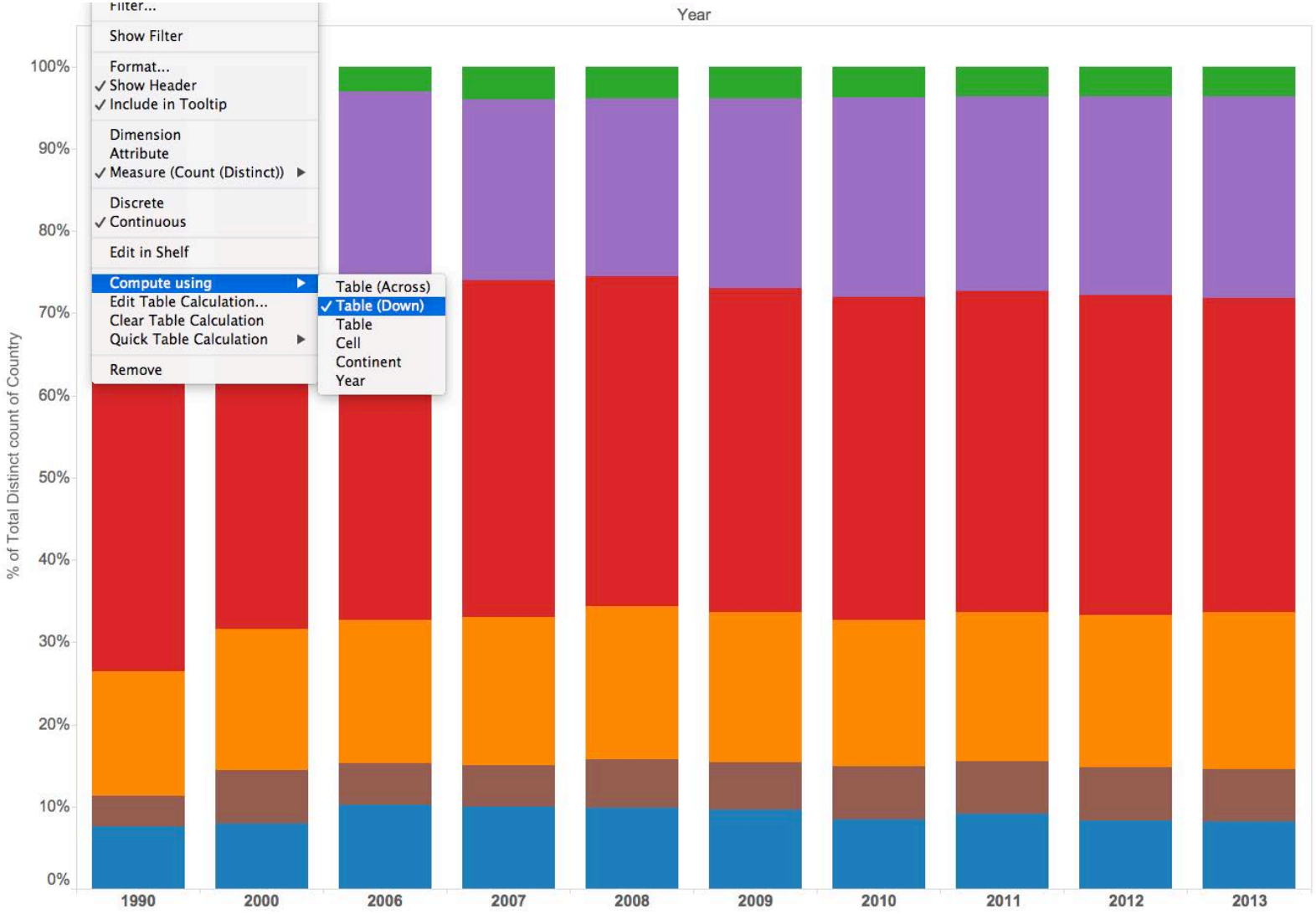

Africa
Asia
Europe
North America
Oceania
South America
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
25.23%
22.43%
22.90%
13.55%
8.88%
7.01%

OF COUNTRIES W/ LIFE EXPECTANCY > 70 YEARS

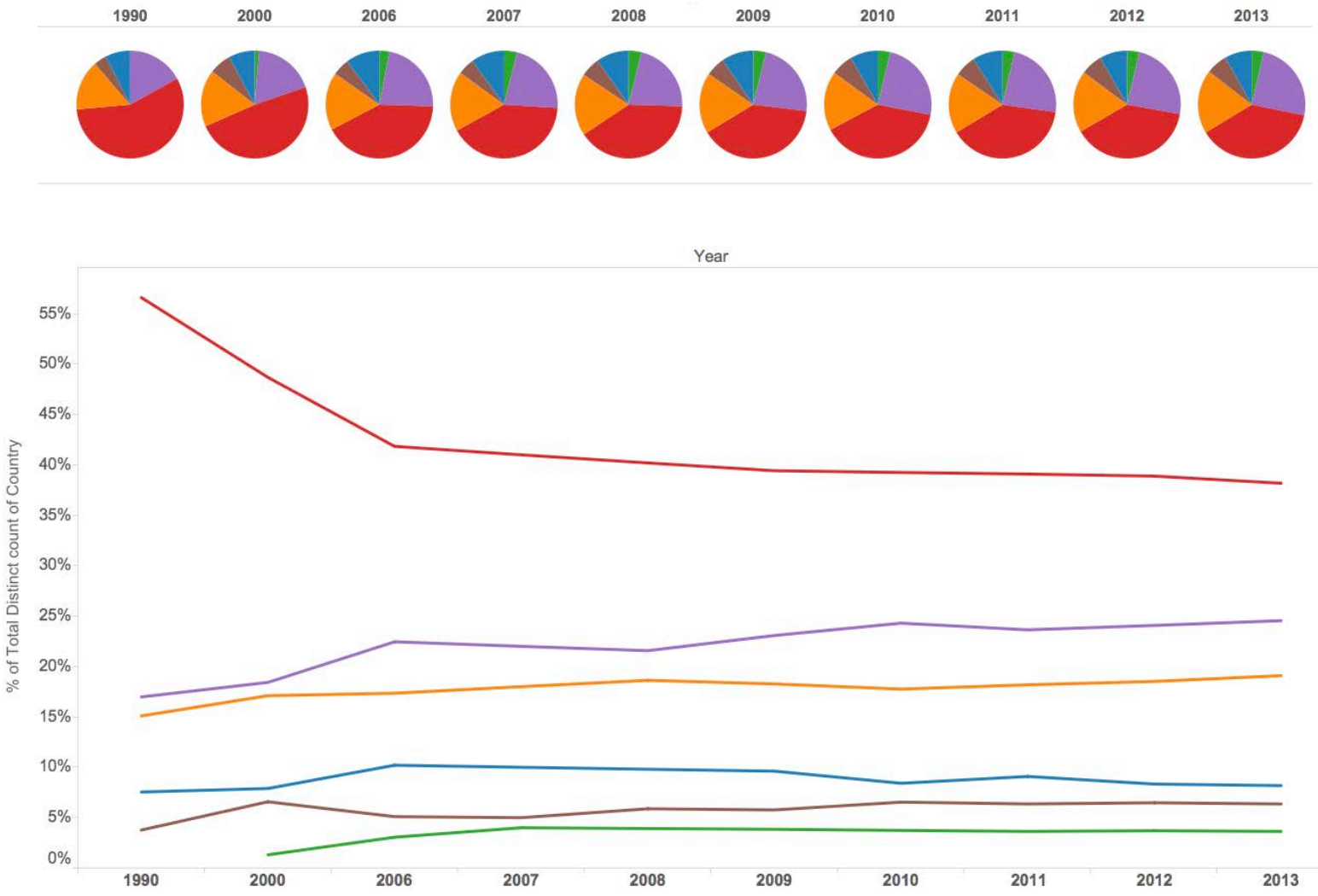


OOPS!

OF COUNTRIES W/ LIFE EXPECTANCY > 70 YEARS

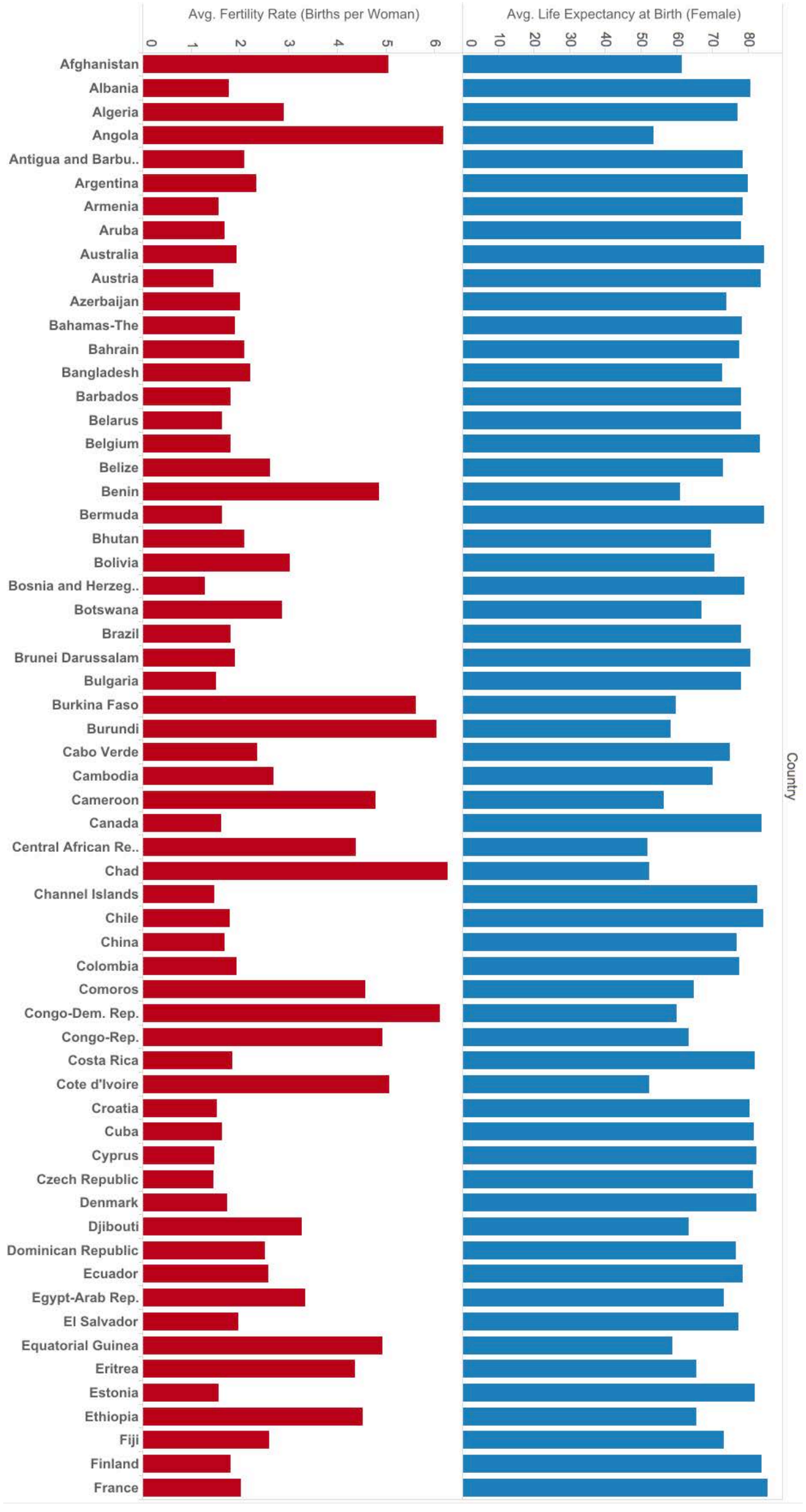


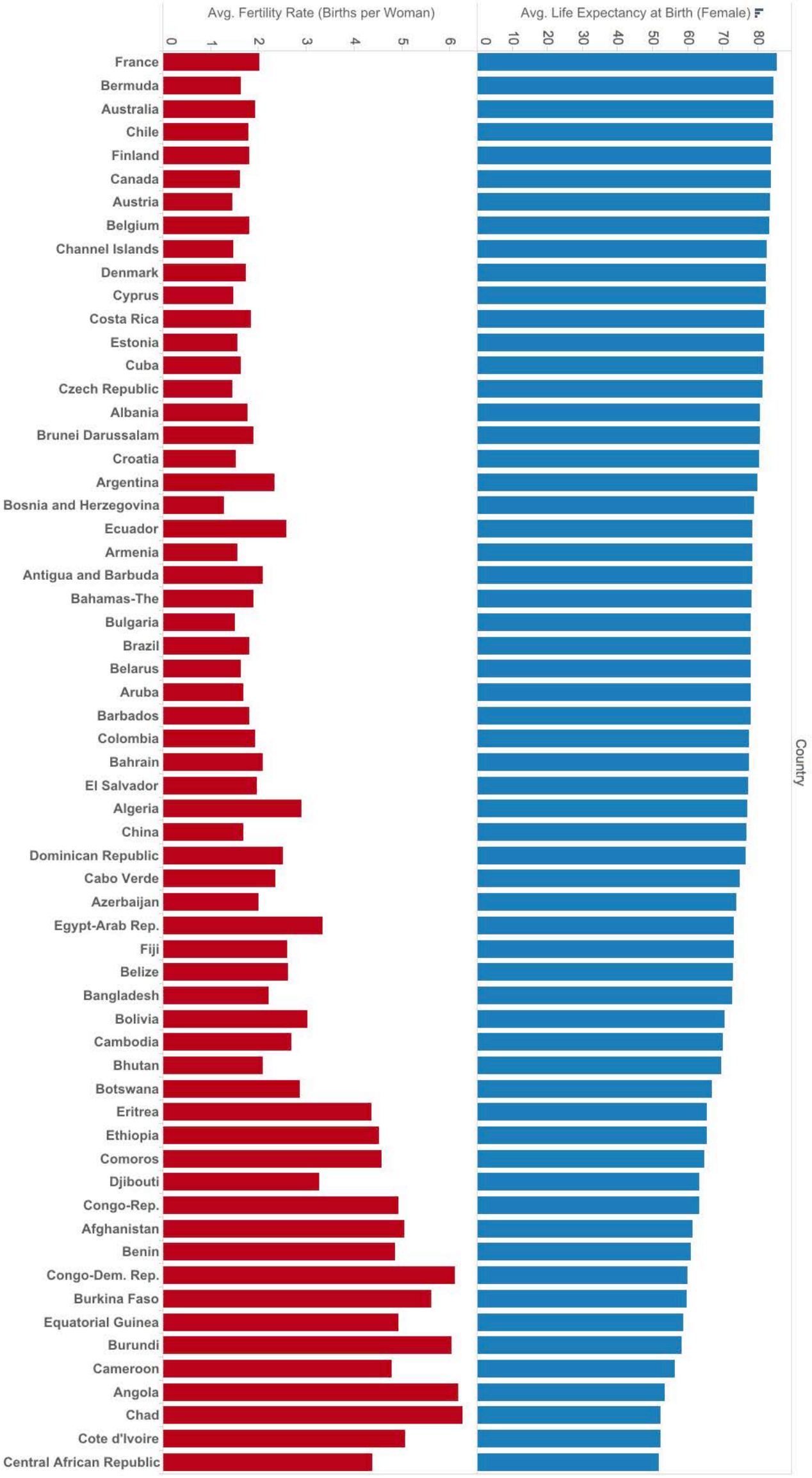
OF COUNTRIES W/ LIFE EXPECTANCY > 70 YEARS

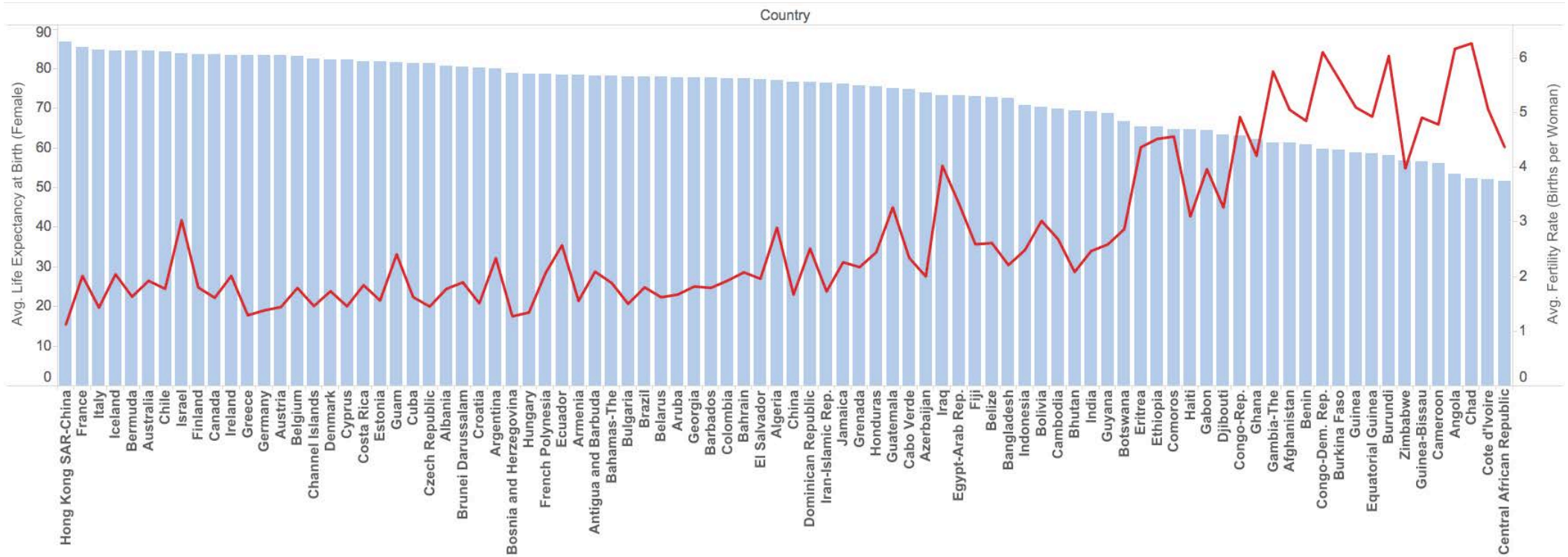


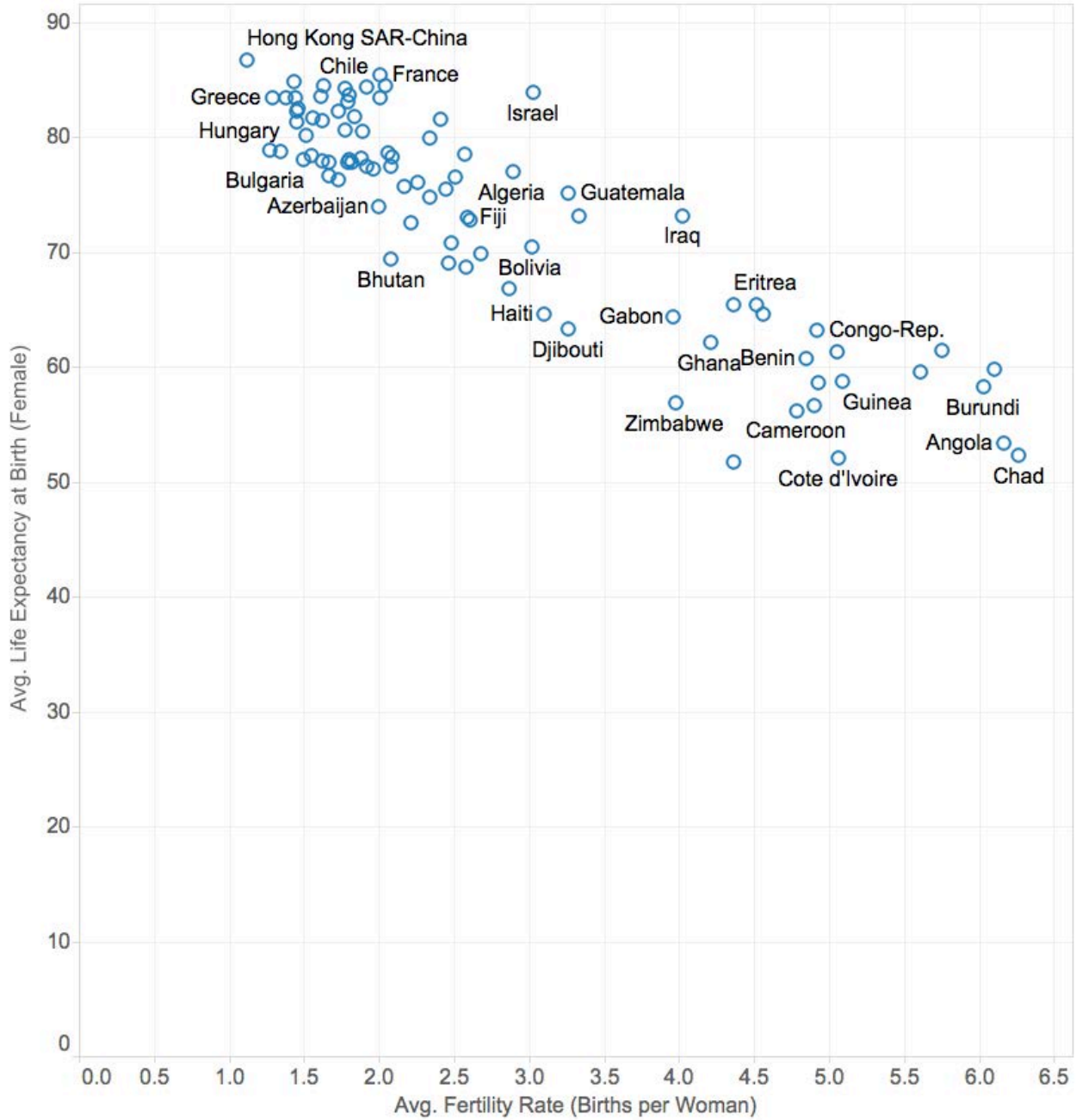
SHOWING RELATIONSHIPS BETWEEN MEASURES

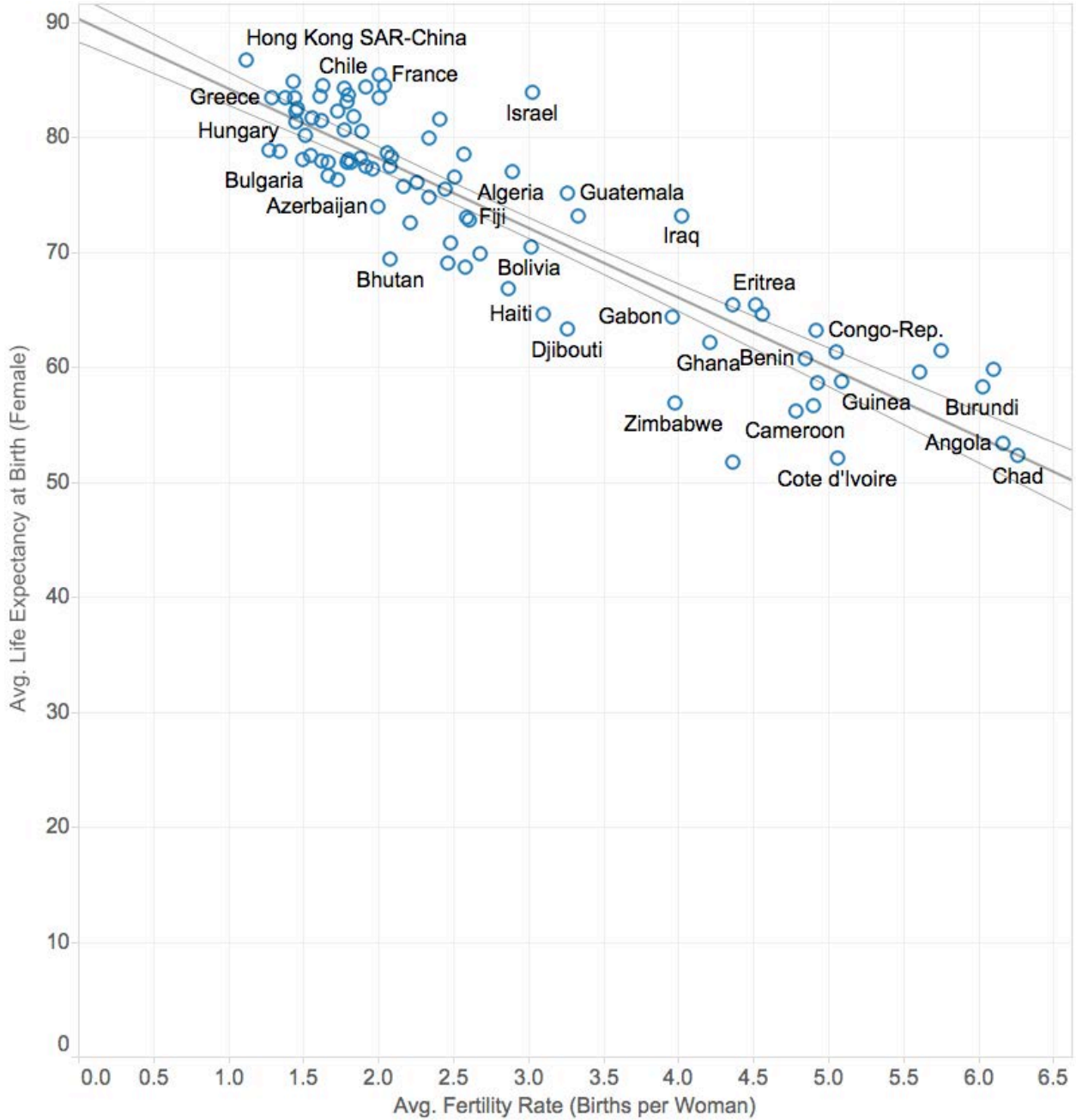
(CORRELATIONS)

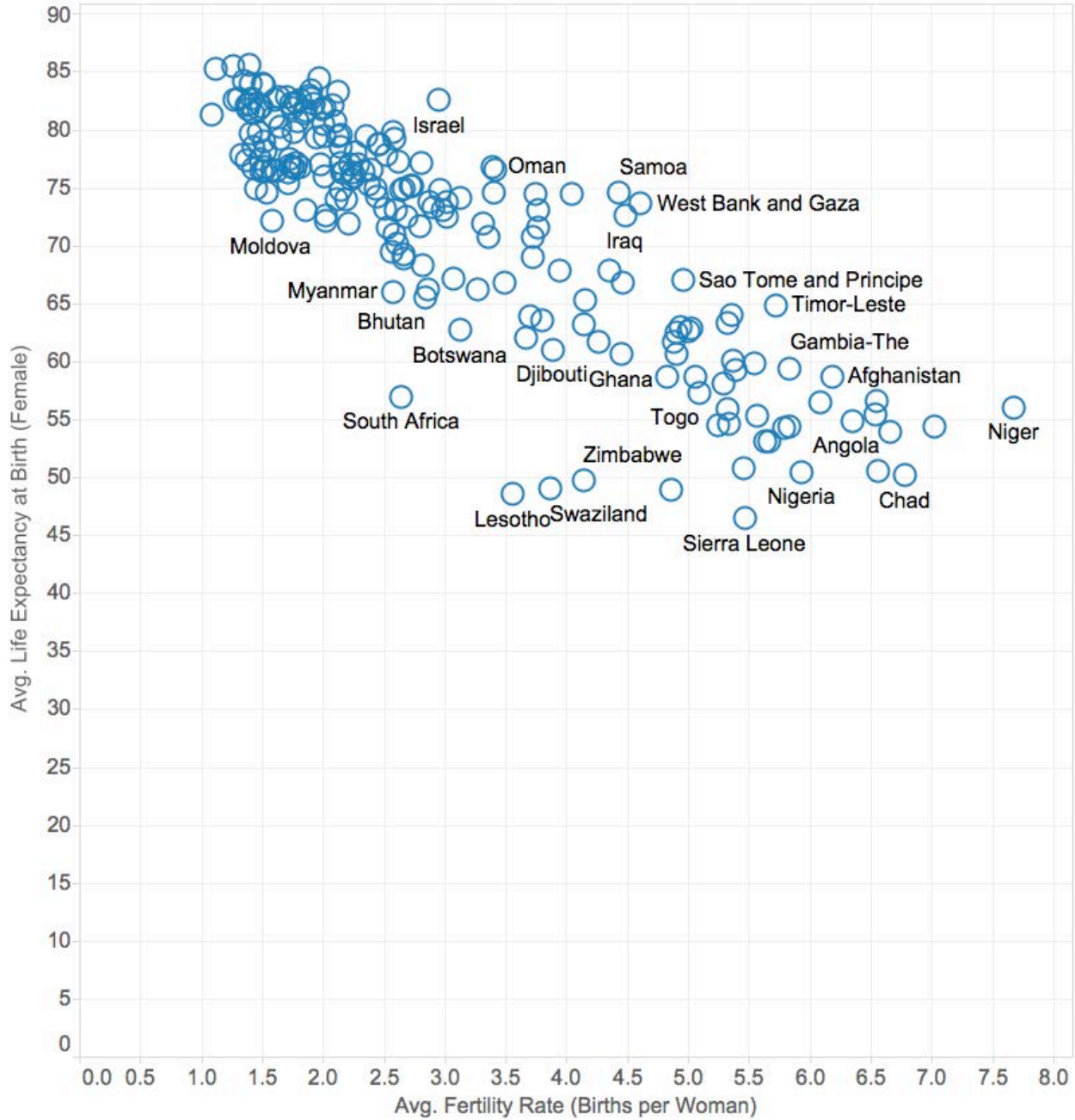

Avg. Fertility Rate (Births per Woman)
0
1
2
3
4
5
6
Avg. Life Expectancy at Birth (Female)
0
10
20
30
40
50
60
70
80
Afghanistan
Albania
Algeria
Angola
Antigua and Barbu..
Argentina
Armenia
Aruba
Australia
Austria
Azerbaijan
Bahamas-The
Bahrain
Bangladesh
Barbados
Belarus
Belgium
Belize
Benin
Bermuda
Bhutan
Bolivia
Bosnia and Herzeg..
Botswana
Brazil
Brunei Darussalam
Bulgaria
Burkina Faso
Burundi
Cabo Verde
Cambodia
Cameroon
Canada
Central African Re..
Chad
Channel Islands
Chile
China
Colombia
Comoros
Congo-Dem. Rep.
Congo-Rep.
Costa Rica
Cote d'Ivoire
Croatia
Cuba
Cyprus
Czech Republic
Denmark
Djibouti
Dominican Republic
Ecuador
Egypt-Arab Rep.
El Salvador
Equatorial Guinea
Eritrea
Estonia
Ethiopia
Fiji
Finland
France
Country

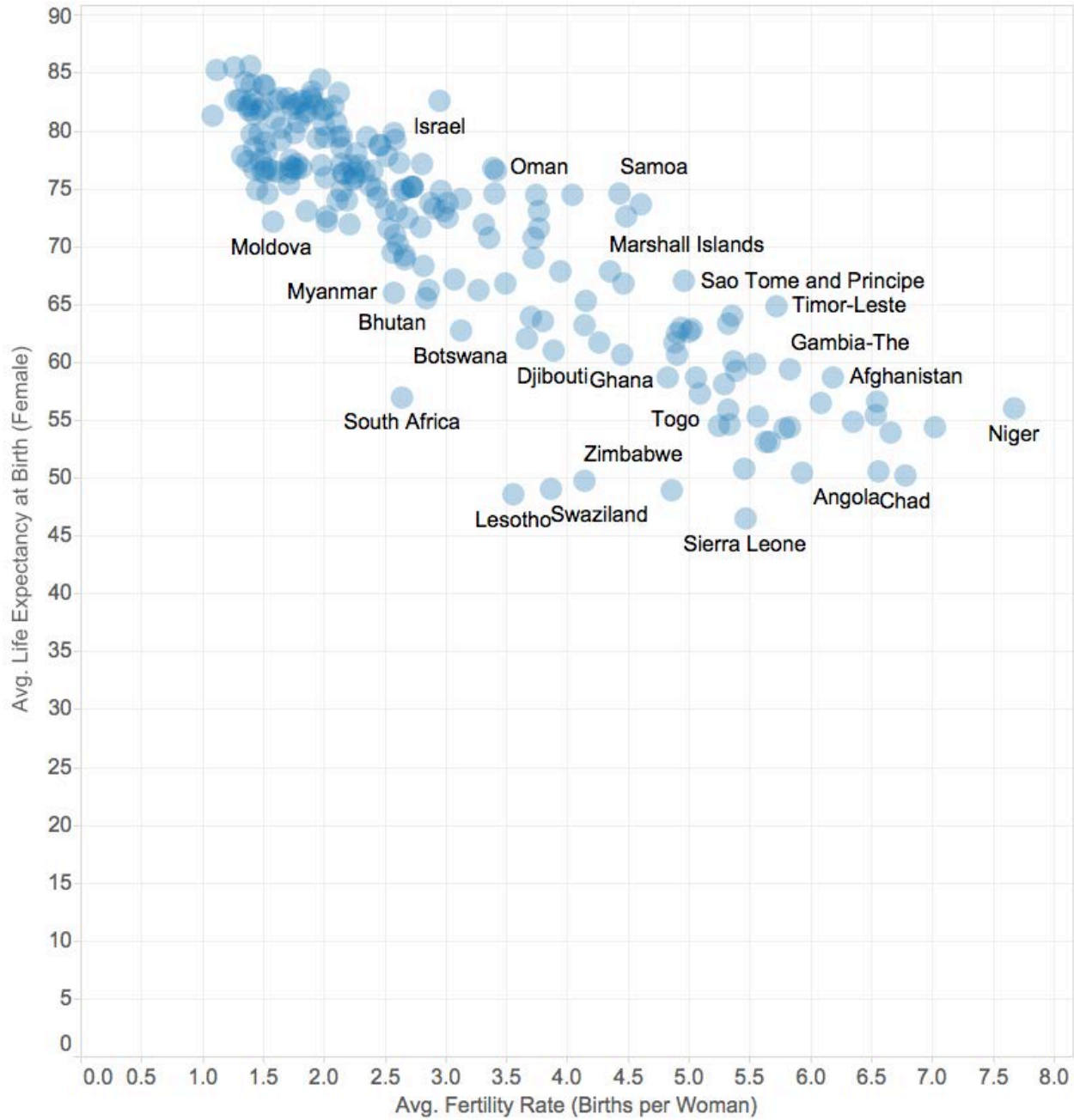

Avg. Fertility Rate (Births per Woman)
0
1
2
3
4
5
6
Avg. Life Expectancy at Birth (Female)
0
10
20
30
40
50
60
70
80
France
Bermuda
Australia
Chile
Finland
Canada
Austria
Belgium
Channel Islands
Denmark
Cyprus
Costa Rica
Estonia
Cuba
Czech Republic
Albania
Brunei Darussalam
Croatia
Argentina
Bosnia and Herzegovina
Ecuador
Armenia
Antigua and Barbuda
Bahamas-The
Bulgaria
Brazil
Belarus
Aruba
Barbados
Colombia
Bahrain
El Salvador
Algeria
China
Dominican Republic
Cabo Verde
Azerbaijan
Egypt-Arab Rep.
Fiji
Belize
Bangladesh
Bolivia
Cambodia
Bhutan
Botswana
Eritrea
Ethiopia
Comoros
Djibouti
Congo-Rep.
Afghanistan
Benin
Congo-Dem. Rep.
Burkina Faso
Equatorial Guinea
Burundi
Cameroon
Angola
Chad
Cote d'Ivoire
Central African Republic
Country

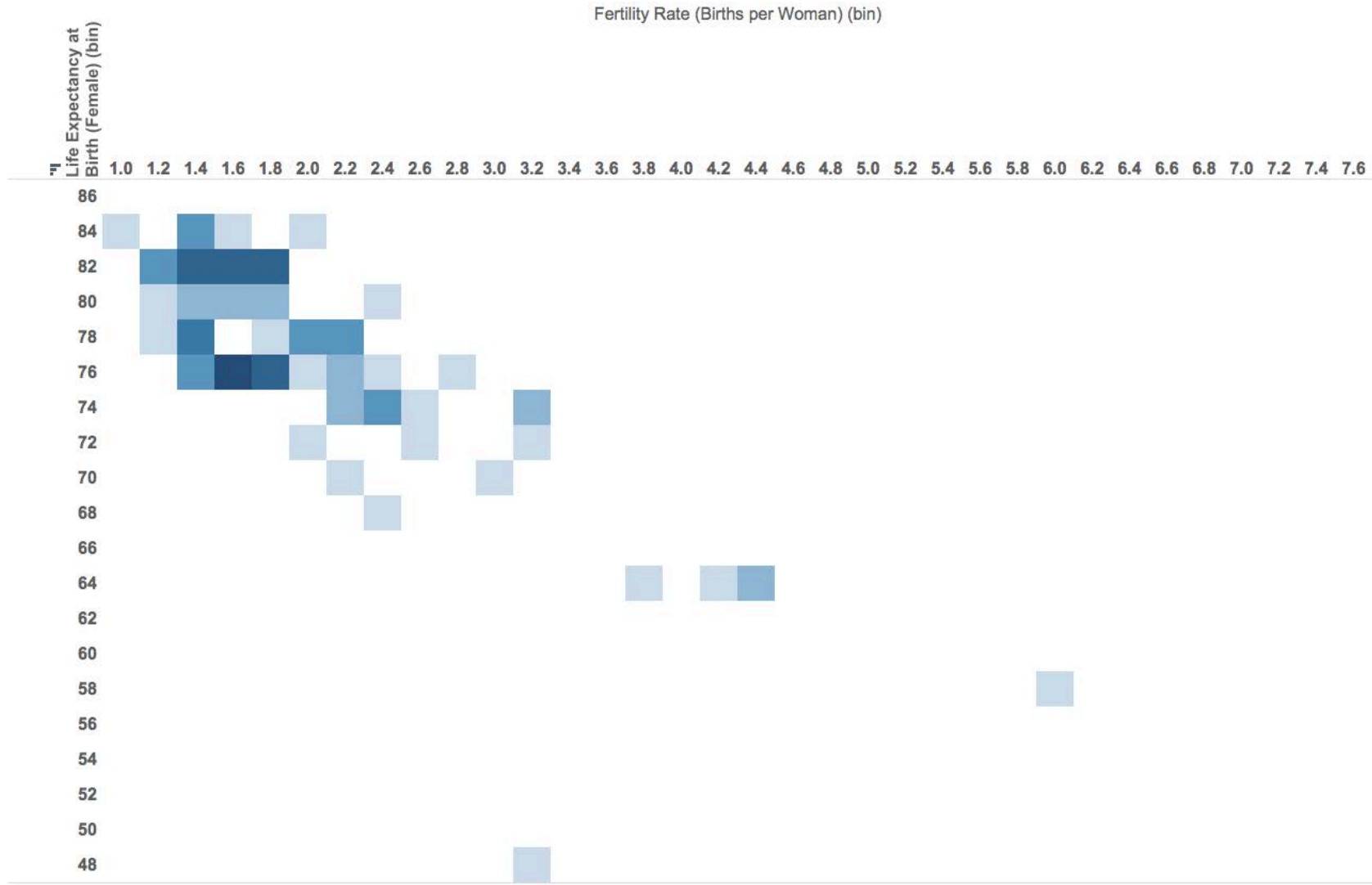

Country
Avg. Life Expectancy at Birth (Female)
Avg. Fertility Rate (Births per Woman)
0
10
20
30
40
50
60
70
80
90
0
1
2
3
4
5
6
Hong Kong SAR-China
France
Italy
Iceland
Bermuda
Australia
Chile
Israel
Finland
Canada
Ireland
Greece
Germany
Austria
Belgium
Channel Islands
Denmark
Cyprus
Costa Rica
Estonia
Guam
Cuba
Czech Republic
Albania
Brunei Darussalam
Croatia
Argentina
Bosnia and Herzegovina
Hungary
French Polynesia
Ecuador
Armenia
Antigua and Barbuda
Bahamas-The
Bulgaria
Brazil
Belarus
Aruba
Georgia
Barbados
Colombia
Bahrain
El Salvador
Algeria
China
Dominican Republic
Iran-Islamic Rep.
Jamaica
Grenada
Honduras
Guatemala
Cabo Verde
Azerbaijan
Iraq
Egypt-Arab Rep.
Fiji
Belize
Bangladesh
Indonesia
Bolivia
Cambodia
Bhutan
India
Guyana
Botswana
Eritrea
Ethiopia
Comoros
Haiti
Gabon
Djibouti
Congo-Rep.
Ghana
Gambia-The
Afghanistan
Benin
Congo-Dem. Rep.
Burkina Faso
Guinea
Equatorial Guinea
Burundi
Zimbabwe
Guinea-Bissau
Cameroon
Angola
Chad
Cote d'Ivoire
Central African Republic


Hong Kong SAR-China
Chile
France
Greece
Israel
Hungary
Bulgaria
Algeria
Guatemala
Azerbaijan
Fiji
Iraq
Bolivia
Bhutan
Eritrea
Haiti
Gabon
Congo-Rep.
Djibouti
Ghana
Benin
Guinea
Burundi
Zimbabwe
Cameroon
Angola
Cote d'Ivoire
Chad
Avg. Life Expectancy at Birth (Female)
Avg. Fertility Rate (Births per Woman)
0
10
20
30
40
50
60
70
80
90
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5


Hong Kong SAR-China
Chile
France
Greece
Israel
Hungary
Bulgaria
Algeria
Guatemala
Azerbaijan
Fiji
Iraq
Bhutan
Bolivia
Eritrea
Haiti
Gabon
Congo-Rep.
Djibouti
Ghana
Benin
Guinea
Burundi
Zimbabwe
Cameroon
Angola
Cote d'Ivoire
Chad
Avg. Life Expectancy at Birth (Female)
Avg. Fertility Rate (Births per Woman)
0
10
20
30
40
50
60
70
80
90
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5


Avg. Life Expectancy at Birth (Female)
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
0.0 0.5 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 7.5 8.0
Avg. Fertility Rate (Births per Woman)
Israel
Oman
Samoa
West Bank and Gaza
Iraq
Moldova
Myanmar
Sao Tome and Principe
Timor-Leste
Bhutan
Gambia-The
Botswana
Djibouti
Ghana
Afghanistan
South Africa
Togo
Niger
Angola
Zimbabwe
Nigeria
Chad
Lesotho
Swaziland
Sierra Leone


Avg. Life Expectancy at Birth (Female)
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
Israel
Oman
Samoa
Moldova
Marshall Islands
Sao Tome and Principe
Myanmar
Timor-Leste
Bhutan
Gambia-The
Botswana
Djibouti
Ghana
Afghanistan
South Africa
Togo
Niger
Zimbabwe
Angola
Chad
Lesotho
Swaziland
Sierra Leone
0.0 0.5 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 7.5 8.0
Avg. Fertility Rate (Births per Woman)


Fertility Rate (Births per Woman) (bin)
Life Expectancy at Birth (Female) (bin)
1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8 3.0 3.2 3.4 3.6 3.8 4.0 4.2 4.4 4.6 4.8 5.0 5.2 5.4 5.6 5.8 6.0 6.2 6.4 6.6 6.8 7.0 7.2 7.4 7.6
86
84
82
80
78
76
74
72
70
68
66
64
62
60
58
56
54
52
50
48

QUESTIONS?