Data Dwarf

University of California: Santa Barbara

2015 Senior Computer Science Capstone Project

Team Members:

Sean Spearman
Cody Brown
Ray Smets
Aimee Galang
Tim Shen

Data is Valuable

The Problem

1. Data’s format is hard to understand.

2. Finding relationships is tedious and not guaranteed.

The Problem

1. Data's format is hard to understand.

2. Finding relationships is tedious and not guaranteed.

	A	B	C	D	E	F
1	Date	Internet Users (per 100 People)	Rural Population (% of total population)	US Household Final Consumption Expenditure PPP (current international $)	US Unemployment Rate (% of Total Labor Force)	Age Dependency Ratio (% of Working-Age Population)
2	12/31/2013	84.2	18.723	1.14843E+13	7.35	50.39534096
3	12/31/2012	79.3	18.892	1.10831E+13	8.075	49.83963833
4	12/31/2011	69.72946076	19.06	1.06893E+13	8.933	49.38454487
5	12/31/2010	71.69	19.228	1.02022E+13	9.625	49.03995685
6	12/31/2009	71	19.394	9.847E+12	9.283	48.81001789
7	12/31/2008	74	19.562	1.00136E+13	5.8	48.689405
8	12/31/2007	75	19.731	9.7505E+12	4.617	48.67564802
9	12/31/2006	68.93119327	19.901	9.304E+12	4.608	48.76268507
10	12/31/2005	67.96805292	20.072	8.7941E+12	5.083	48.94280961
11	12/31/2004	64.75825648	20.243	8.26E+12	5.542	49.21455249
12	12/31/2003	61.69711712	20.417	7.7655E+12	5.992	49.56972092
13	12/31/2002	58.78540388	20.591	7.3841E+12	5.783	49.98365968
14	12/31/2001	49.08083159	20.766	7.1031E+12	4.742	50.42393238
15	12/31/2000	43.07916264	20.943	6.7924E+12	3.967	50.86247379
16	12/31/1999	35.84872446	21.258	6.307E+12	4.217	51.29067669
17	12/31/1998	30.09319659	21.623	5.903E+12	4.5	51.70052723
18	12/31/1997	21.61640097	21.992	5.5607E+12	4.942	52.06392098
19	12/31/1996	16.41935296	22.364	5.2681E+12	5.408	52.34669698
20	12/31/1995	9.237088297	22.743	4.9842E+12	5.592	52.52296438
21	12/31/1994	4.862780635	23.125	4.741E+12	6.1	52.57795283
22	12/31/1993	2.271673294	23.512	4.471E+12	6.908	52.51600601
23	12/31/1992	1.724202539	23.903	4.2157E+12	7.492	52.35597463
24	12/31/1991	1.163193726	24.299	3.9602E+12	6.85	52.12947135
25	12/31/1990	0.784728502	24.7	3.8256E+12	5.617	51.86500605
26	12/31/1989	n.a.	24.911	3.5928E+12	5.258	51.57014711
27	12/31/1988	n.a.	25.058	3.3469E+12	5.492	51.25318383
28	12/31/1987	n.a.	25.207	3.0921E+12	6.175	50.9462661
29	12/31/1986	n.a.	25.356	2.8984E+12	7	50.68849122
30	12/31/1985	n.a.	25.506	2.7226E+12	7.192	50.51191344
31	12/31/1984	n.a.	25.656	2.4981E+12	7.508	50.431508
32	12/31/1983	n.a.	25.806	2.2865E+12	9.6	50.45404956
33	12/31/1982	n.a.	25.958	2.0739E+12	9.708	50.59283254
34	12/31/1981	n.a.	26.11	1.9375E+12	7.617	50.85913991
35	12/31/1980	n.a.	26.262	1.7546E+12	7.175	51.26088467
36	12/31/1979	n.a.	26.308	n.a.	n.a.	51.80449825

Datasets sourced from *World Bank Cross Country Data* and *ILOSTAT Database* through *Quandl*

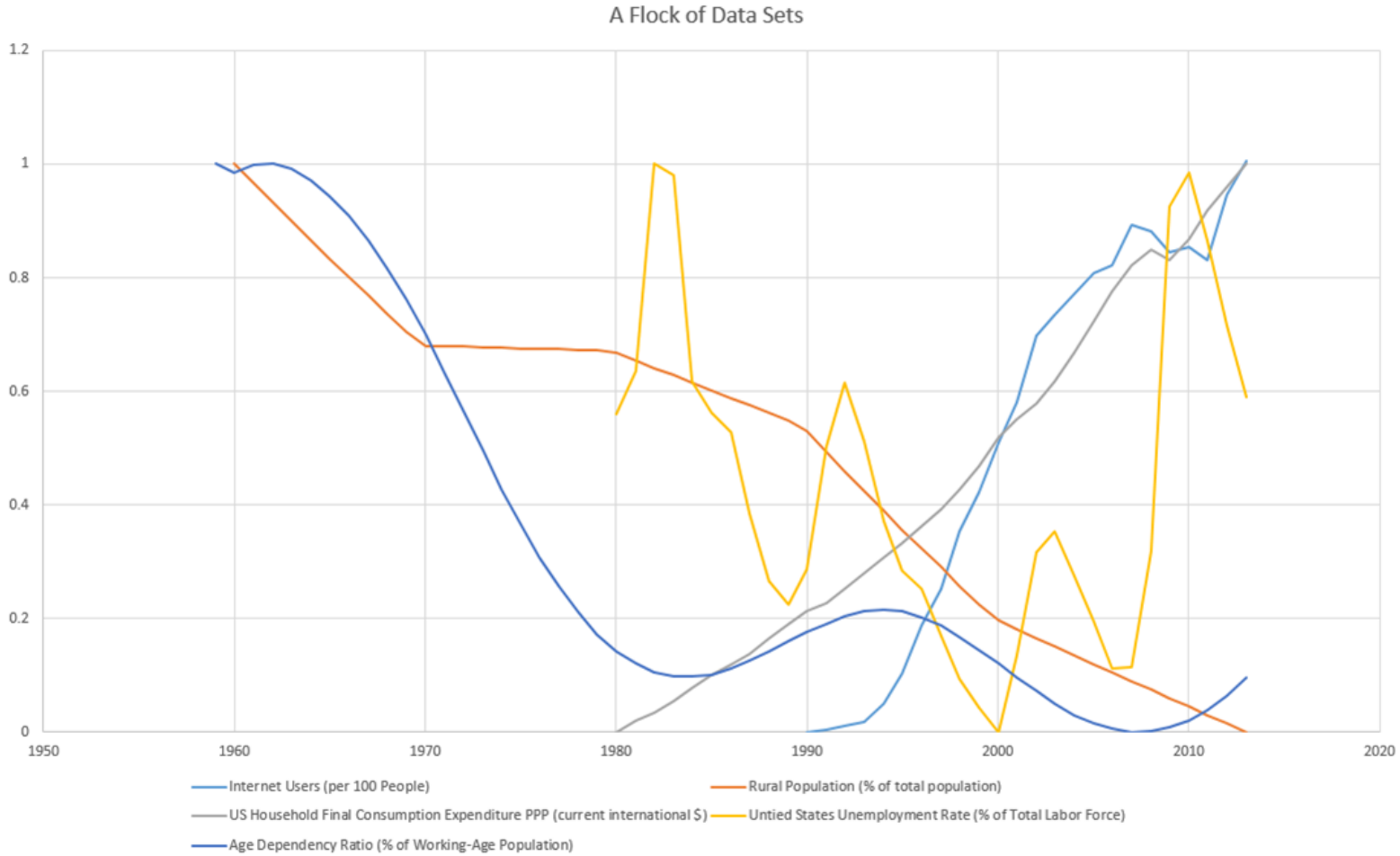
	A	B	C	D	E	F
1	Date	Internet Users (per 100 People)	Rural Population (% of total population)	US Household Final Consumption Expenditure PPP (current international $)	US Unemployment Rate (% of Total Labor Force)	Age Dependency Ratio (% of Working-Age Population)
2	12/31/2013	84.2	18.723	1.14843E+13	7.35	50.39534096
3	12/31/2012	79.3	18.892	1.10831E+13	8.075	49.83963833
4	12/31/2011	69.72946076	19.06	1.06893E+13	8.933	49.3845[illegible]487
5	12/31/2010	71.69	19.228	1.02022E+13	9.625	49.03995[illegible]585
6	12/31/2009	71	19.394	9.847E+12	9.283	48.8[illegible]001789
7	12/31/2008	74	19.562	1.00136E+13	5.8	48.689405
8	12/31/2007	75	19.731	9.7505E+12	4.617	48.[illegible]64802
9	12/31/2006	68.93119327	19.901	9.304E+12	4.608	48.76268507
10	12/31/2005	67.96805292	20.072	8.7941E+12	5.083	48.94280961
11	12/31/2004	64.75825648	20.243	8.26E+12	5.542	49.21455249
12	12/31/2003	61.69711712	20.417	7.7655E+12	5.992	49.56972092
13	12/31/2002	58.78540388	20.591	7.3841E+12	5.783	49.98365968
14	12/31/2001	49.08083159	20.766	7.1031E+12	4.742	50.42393238
15	12/31/2000	43.07916264	20.943	6.7924E+12	3.967	50.86247379
16	12/31/1999	35.84872446	21.258	6.307E+12	4.217	51.29067669
17	12/31/1998	30.09319659	21.623	5.903E+12	4.5	51.70052723
18	12/31/1997	21.61640097	21.992	5.5607E+12	4.942	52.06392098
19	12/31/1996	16.41935296	22.364	5.2681E+12	5.408	52.34669698
20	12/31/1995	9.237088297	22.743	4.9842E+12	5.592	52.52296438
21	12/31/1994	4.862780635	23.125	4.741E+12	6.1	52.57795283
22	12/31/1993	2.271673294	23.512	4.471E+12	6.908	52.51600601
23	12/31/1992	1.724202539	23.903	4.2157E+12	7.492	52.35597463
24	12/31/1991	1.163193726	24.299	3.9602E+12	6.85	52.12947135
25	12/31/1990	0.784728502	24.7	3.8256E+12	5.617	51.86500605
26	12/31/1989	n.a.	24.911	3.5928E+12	5.258	51.57014711
27	12/31/1988	n.a.	25.058	3.3469E+12	5.492	51.25318383
28	12/31/1987	n.a.	25.207	3.0921E+12	6.175	50.9462661
29	12/31/1986	n.a.	25.356	2.8984E+12	7	50.68849122
30	12/31/1985	n.a.	25.506	2.7226E+12	7.192	50.51191344
31	12/31/1984	n.a.	25.656	2.4981E+12	7.508	50.431508
32	12/31/[illegible]983	n.a.	25.806	2.2865E+12	9.6	50.45404956
33	12/31/1982	n.a.	25.958	2.0739E+12	9.708	50.59283254
34	12/31/198[illegible]	n.a.	26.11	[illegible]9375E+12	7.617	50.85913991
35	12/31/1980	n.a.	26.262	[illegible]7546E+12	7.175	51.26088467
36	12/31/19[illegible]	n.a.	26.308	n.a.	n.a.	51.80449825

Datasets sourced from *World Bank Cross Country Data* and *ILOSTAT Database* through *Quandl*

The Problem

1. Data's format is hard to understand.

2. Finding relationships is tedious and not guaranteed.


A Flock of Data Sets
1.2
1
0.8
0.6
0.4
0.2
0
1950
1960
1970
1980
1990
2000
2010
2020
Internet Users (per 100 People)
Rural Population (% of total population)
US Household Final Consumption Expenditure PPP (current international $)
Untied States Unemployment Rate (% of Total Labor Force)
Age Dependency Ratio (% of Working-Age Population)

Correlation Indicates Potentially Meaningful Relationships

DataDwarf

Technology

Front End

Back End

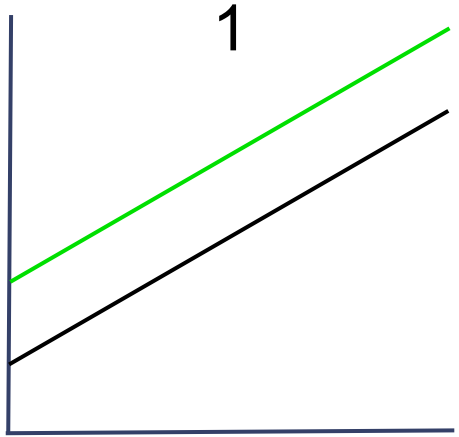
Data Source

Teamwork

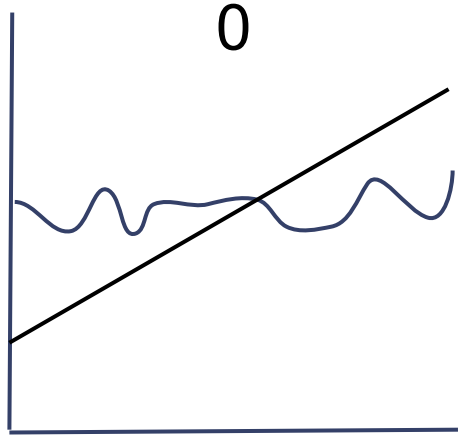
Correlation Value Usage

- **Input:** 2 datasets
- **Output:** one number
- **Want to know:** how similar they are

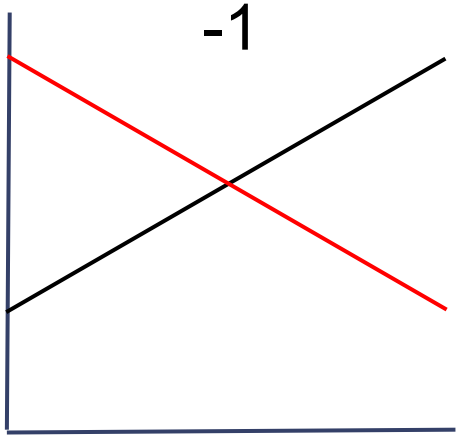
Correlation Range



Directly correlated



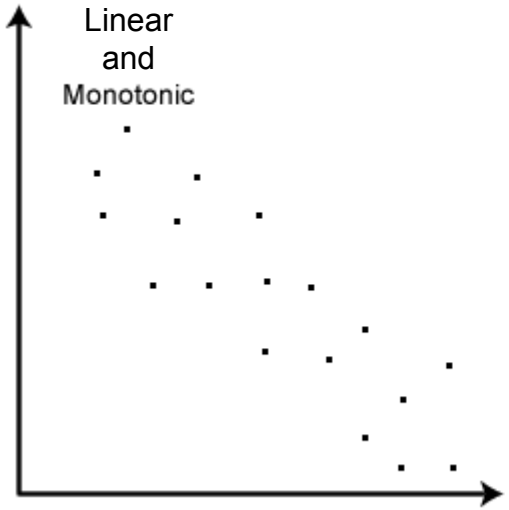
Not correlated



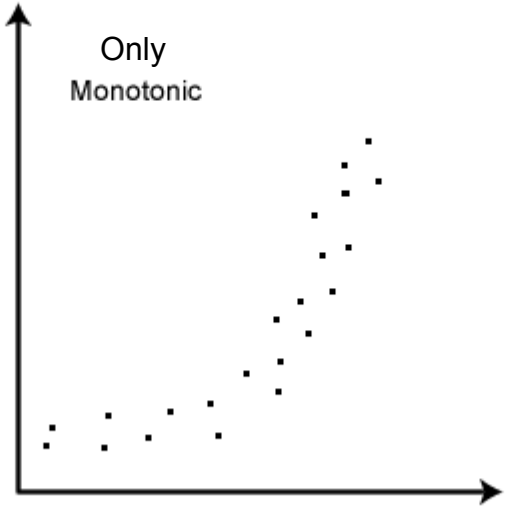
Inversely correlated

Pearson vs (Spearman or Kendall)

use Pearson

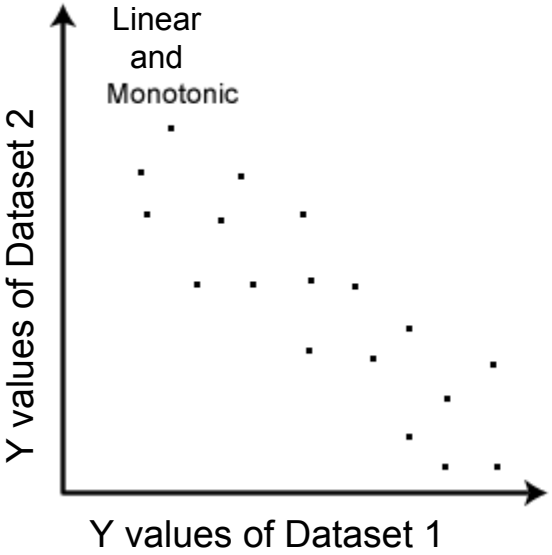


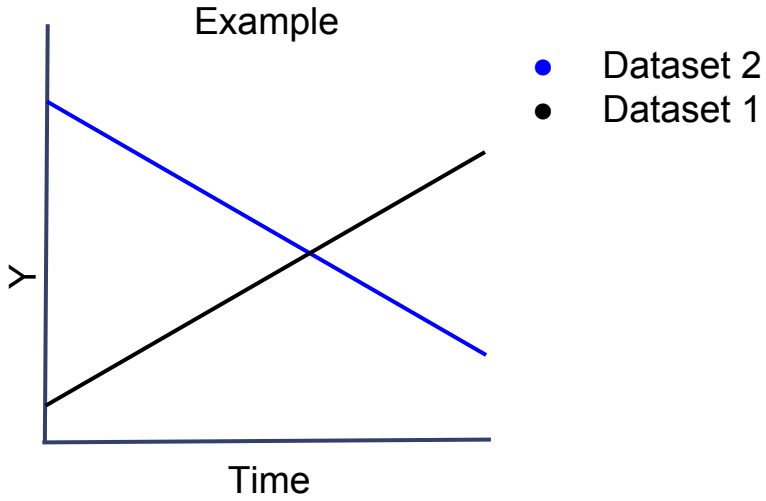
use Spearman or Kendall



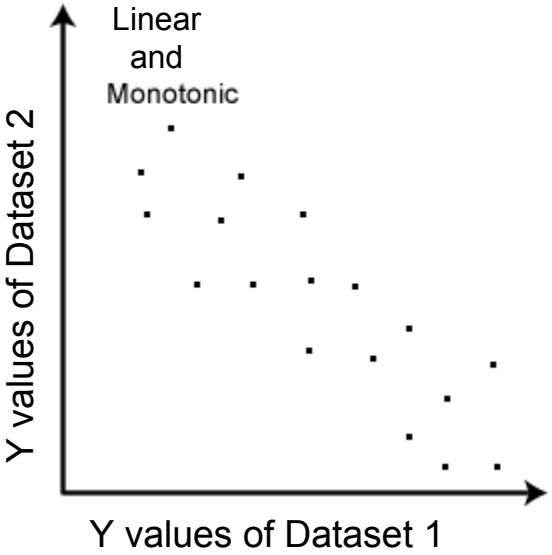
https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

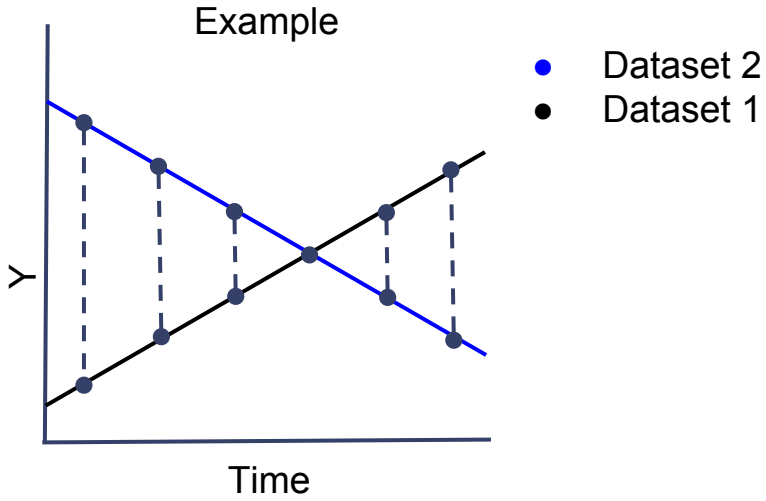
Pearson


Linear
and
Monotonic
Y values of Dataset 2
Y values of Dataset 1

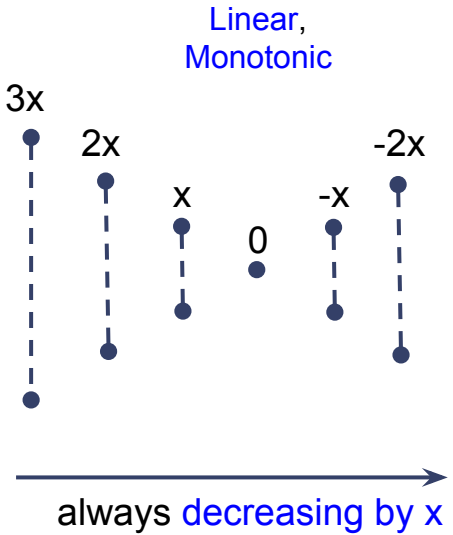

Example
Dataset 2
Dataset 1
Y
Time

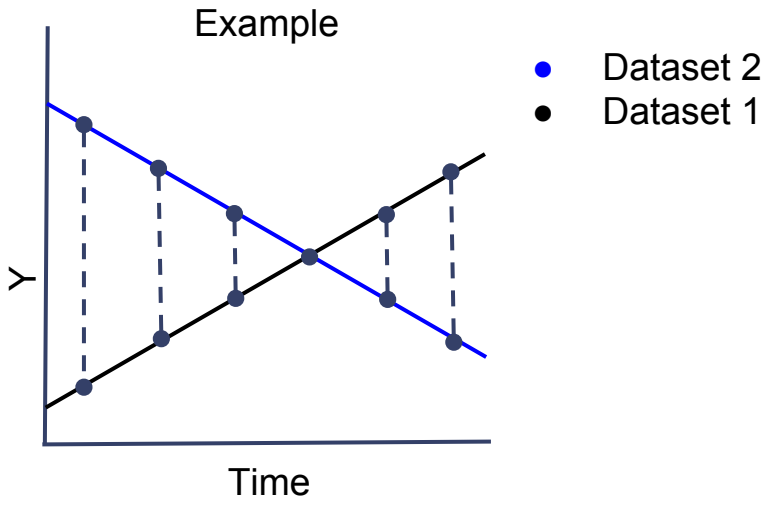
Pearson


Linear
and
Monotonic
Y values of Dataset 2
Y values of Dataset 1

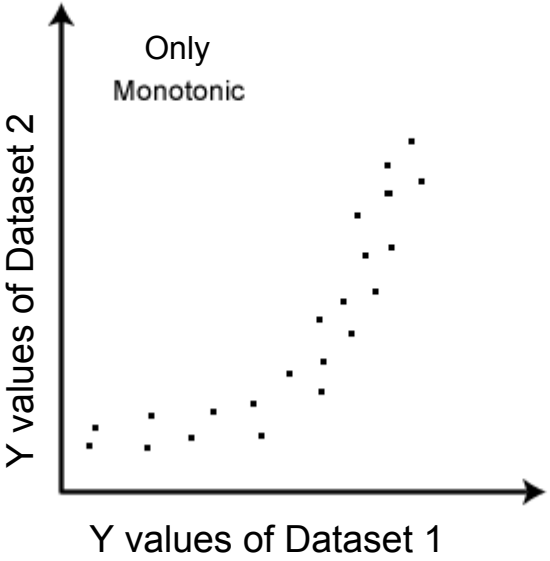

Example
Dataset 2
Dataset 1
Y
Time

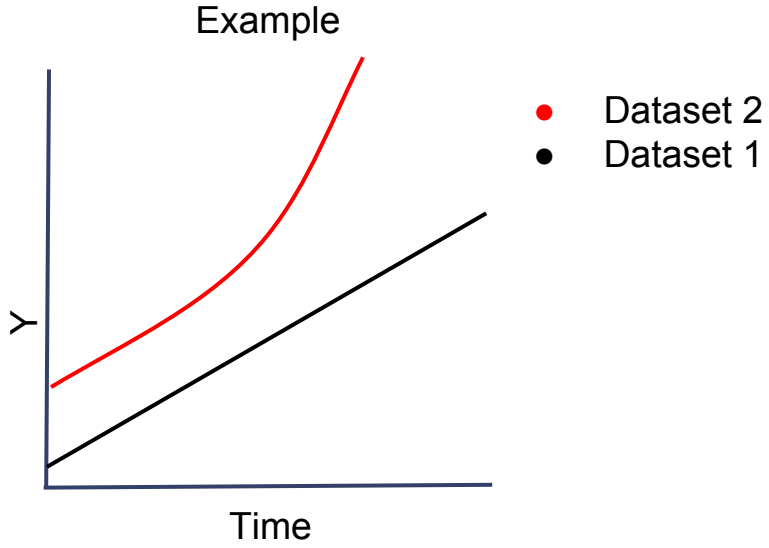
Pearson


Linear,
Monotonic
3x
2x
x
0
-x
-2x
always decreasing by x

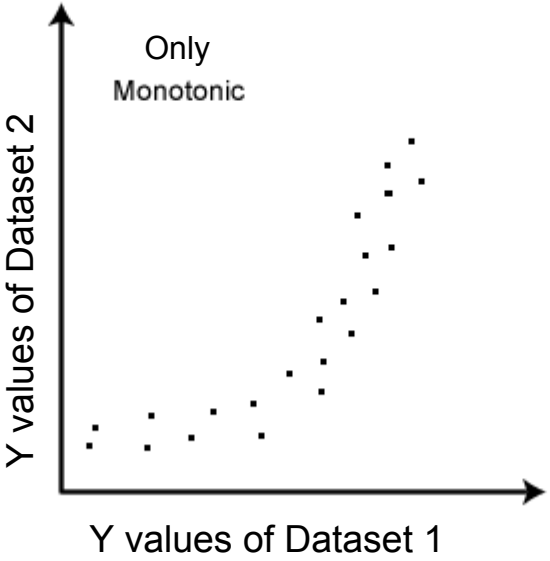

Example
Y
Time
Dataset 2
Dataset 1

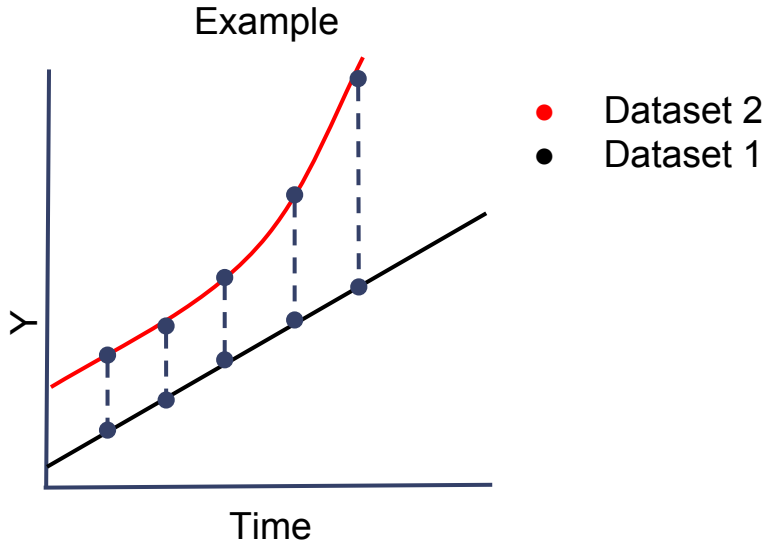
Spearman and Kendall



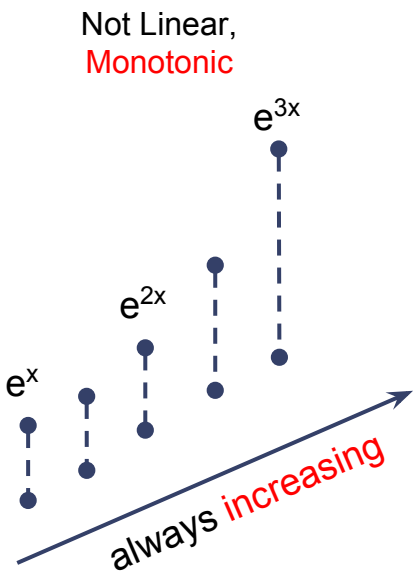


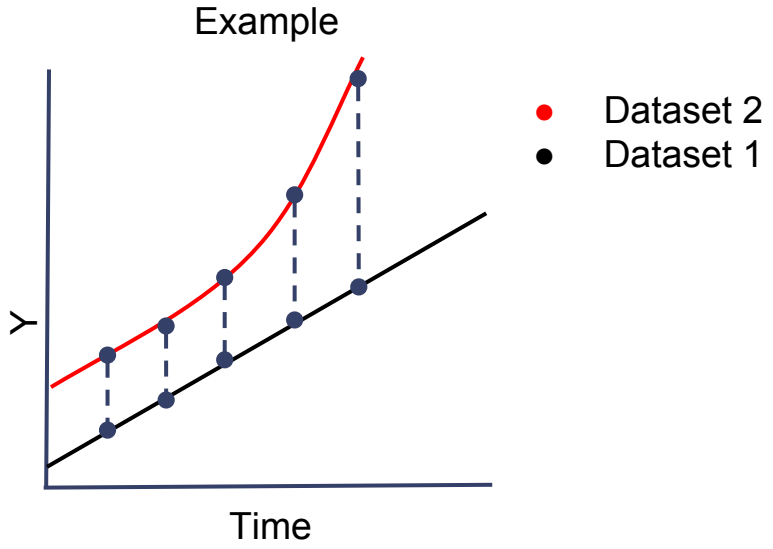
Spearman and Kendall



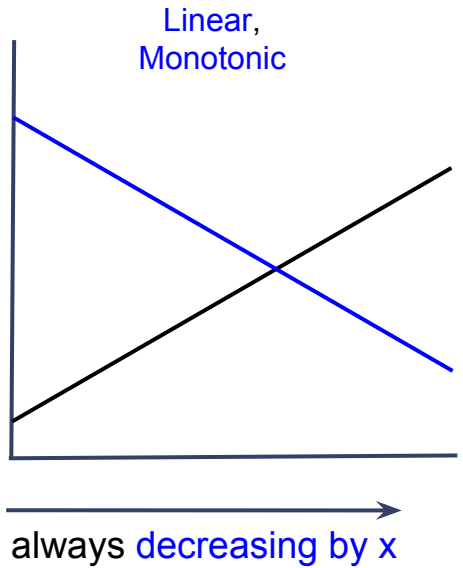


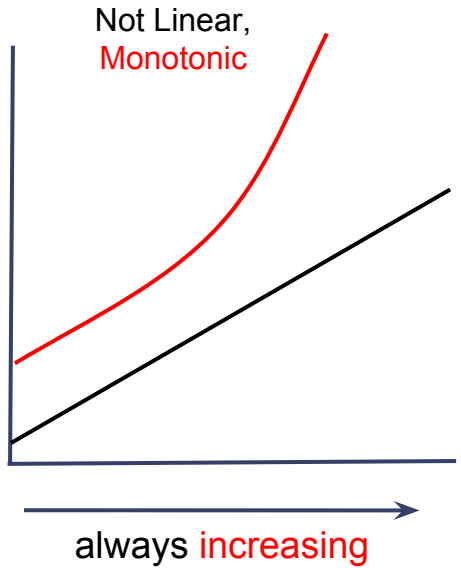
Spearman and Kendall





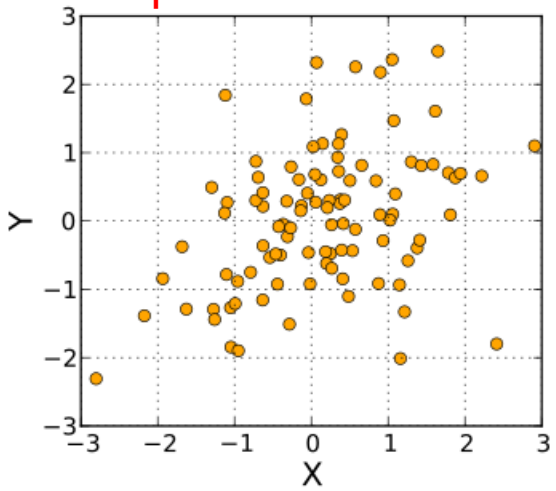
Pearson vs (Spearman and Kendall)



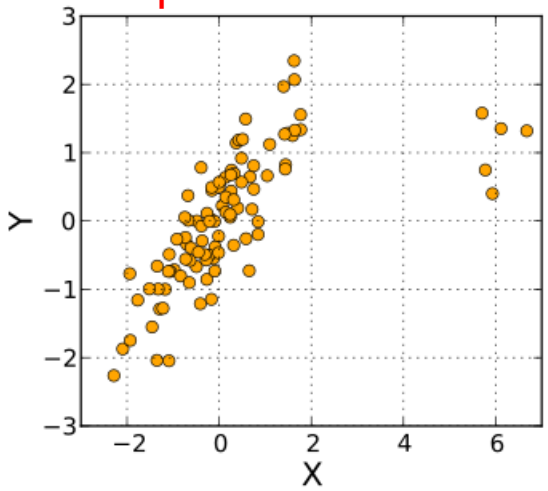


Pearson vs Spearman

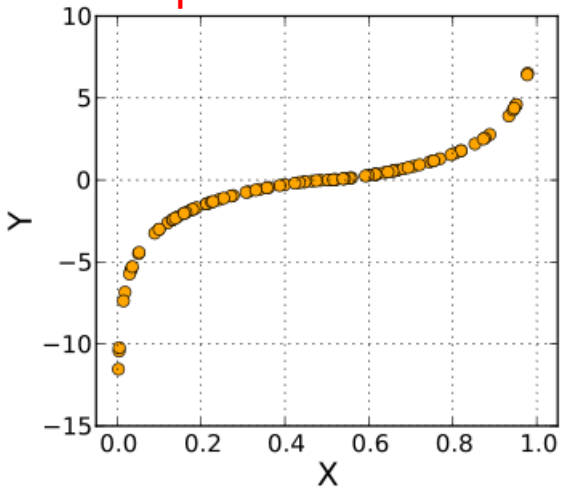
Pearson = 0.37
Spearman = 0.35



Pearson = 0.67
Spearman = 0.84

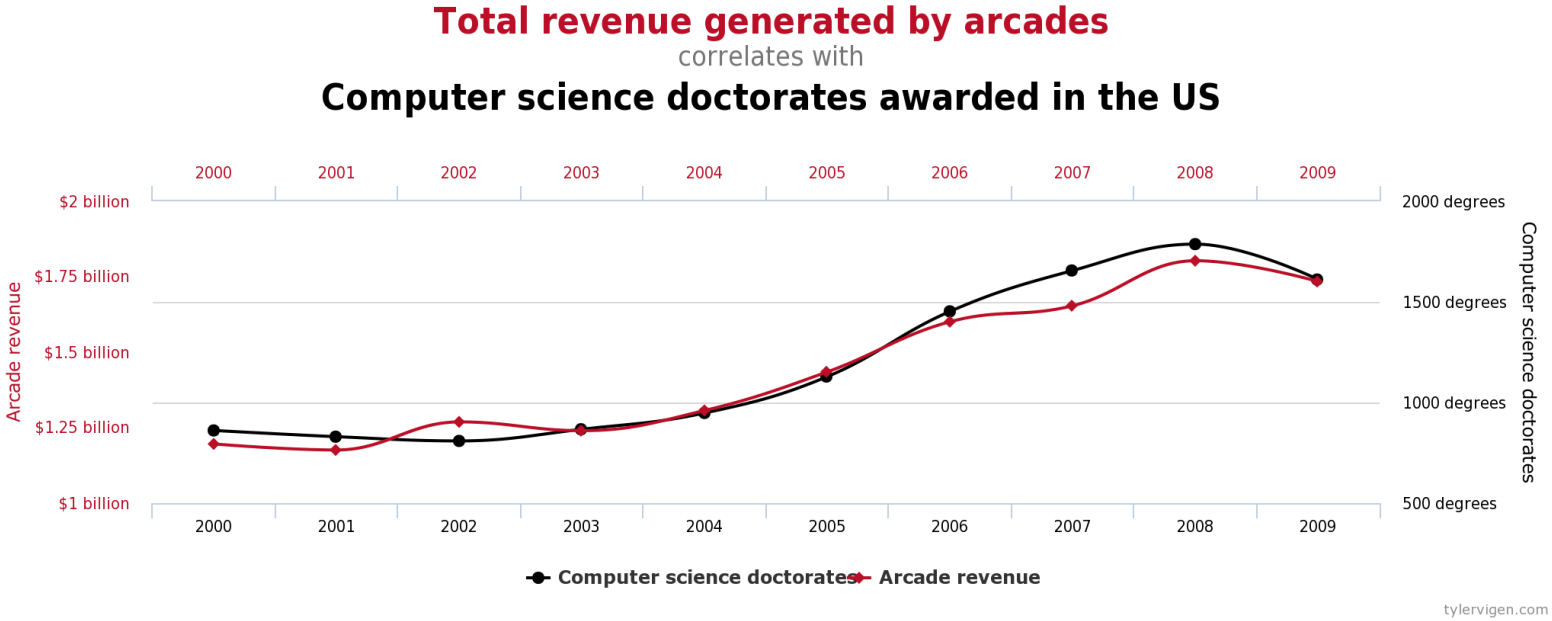


Pearson = 0.88
Spearman = 1



http://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Spurious Correlation Example



Demo

Bring your own data (BYOD) to DataDwarf.io