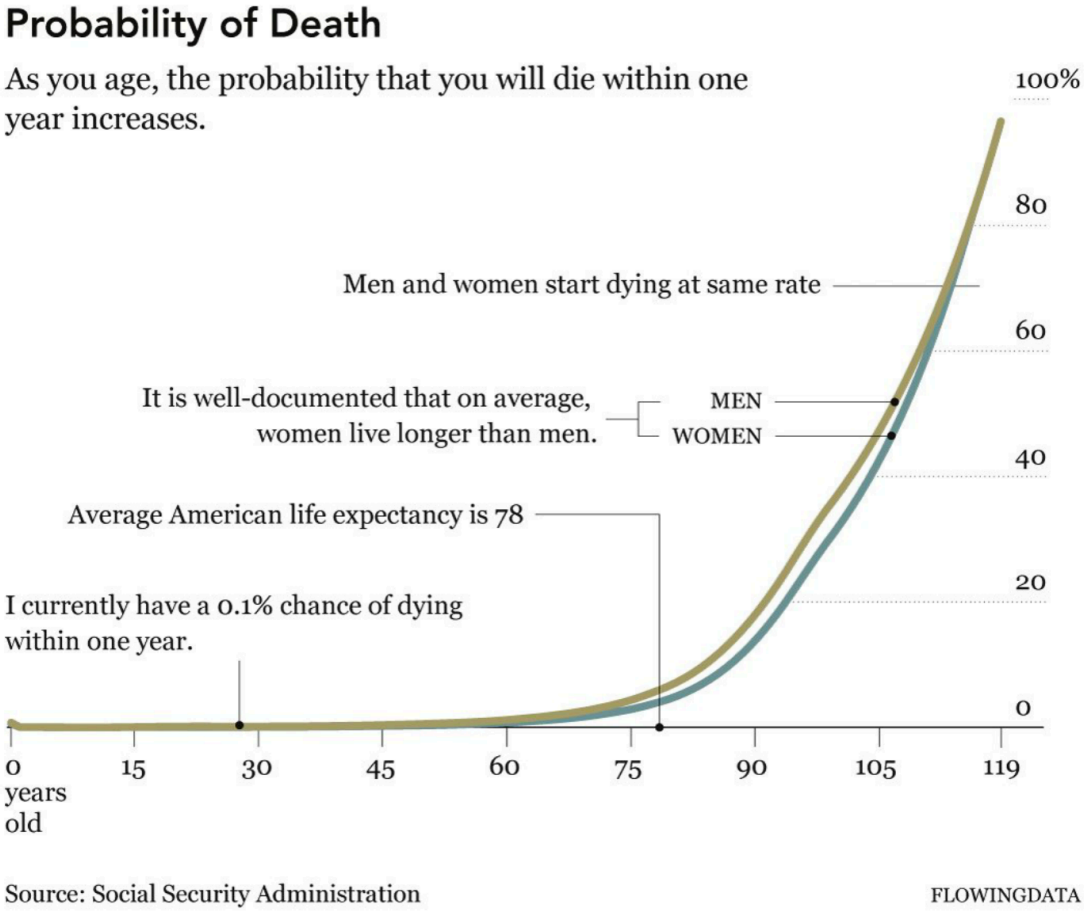
Data Visualization: Messages in Data

Karl Ho
School of Economic, Political and Policy Sciences
University of Texas at Dallas

Data Story:



Source: Yau 2011

Numbers

1. Numbers vs. ?
2. Zeros and ones
3. Integers (0, 1, 2, 100) vs. Non-integers (2.5, 3.1416)
4. Positive and negative numbers

Quantitative vs. Qualitative Data

1. Numbers vs. Labels
2. Quantity vs. Quality
3. Ordinal, Interval, Ratio vs. Nominal
4. e.g. Yes/No--> Qualitative
5. e.g. How much--> Quantitative

Quantitative vs. Qualitative Data

1. Higher quantity means higher quality?
2. Higher quality leads to higher quantity?

Basics of data organization

1. Variables and observations
    1. Alternative terms: Fields and cases
2. Rows for observations , columns for variables
3. Names and labels
4. Table vs. query

What to visualize in data?

1. Data Generating Process
2. Property
3. Distribution
4. Pattern
5. Differences
6. Relationship

Time series data

1. Nature
    1. Temporal dependency: non-stationarity autocorrelation
    2. Periodicity: seasonality, cycle
2. Zeros -> events?
3. Scale linearity

Event count data

1. Nature
    1. Distribution
    2. Bounds
        1. No upper bounds
        2. One lower bound: zero
    3. Zeros
2. Continuous vs. discrete
3. Intervals vs. duration

Bertram M. Gross (1986)

"the world or my part of it is seen as an ongoing stream of events in time . . . Facts and process are separated into discrete elements only by human analysis . . . Change-whether rapid or slow, hidden or open-is continuous."

Anscombe example (1973)

```
> anscombe
   x1 x2 x3 x4    y1   y2    y3    y4
1  10 10 10  8  8.04 9.14  7.46  6.58
2   8  8  8  8  6.95 8.14  6.77  5.76
3  13 13 13  8  7.58 8.74 12.74  7.71
4   9  9  9  8  8.81 8.77  7.11  8.84
5  11 11 11  8  8.33 9.26  7.81  8.47
6  14 14 14  8  9.96 8.10  8.84  7.04
7   6  6  6  8  7.24 6.13  6.08  5.25
8   4  4  4 19  4.26 3.10  5.39 12.50
9  12 12 12  8 10.84 9.13  8.15  5.56
10  7  7  7  8  4.82 7.26  6.42  7.91
11  5  5  5  8  5.68 4.74  5.73  6.89
```

Anscombe example (1973)

```
> summary(anscombe)
       x1             x2             x3             x4           y1              y2        
 Min.   : 4.0   Min.   : 4.0   Min.   : 4.0   Min.   : 8   Min.   : 4.260   Min.   :3.100  
 1st Qu.: 6.5   1st Qu.: 6.5   1st Qu.: 6.5   1st Qu.: 8   1st Qu.: 6.315   1st Qu.:6.695  
 Median : 9.0   Median : 9.0   Median : 9.0   Median : 8   Median : 7.580   Median :8.140  
 Mean   : 9.0   Mean   : 9.0   Mean   : 9.0   Mean   : 9   Mean   : 7.501   Mean   :7.501  
 3rd Qu.:11.5   3rd Qu.:11.5   3rd Qu.:11.5   3rd Qu.: 8   3rd Qu.: 8.570   3rd Qu.:8.950  
 Max.   :14.0   Max.   :14.0   Max.   :14.0   Max.   :19   Max.   :10.840   Max.   :9.260  
       y3              y4        
 Min.   : 5.39   Min.   : 5.250  
 1st Qu.: 6.25   1st Qu.: 6.170  
 Median : 7.11   Median : 7.040  
 Mean   : 7.50   Mean   : 7.501  
 3rd Qu.: 7.98   3rd Qu.: 8.190  
 Max.   :12.74   Max.   :12.500  
```

Anscombe example (1973)

```
Analysis of Variance Table

Response: y1
          Df Sum Sq Mean Sq F value  Pr(>F)
x1         1 27.510 27.5100   17.99 0.00217 **
Residuals  9 13.763  1.5292
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Analysis of Variance Table

Response: y2
          Df Sum Sq Mean Sq F value   Pr(>F)
x2         1 27.500 27.5000  17.966 0.002179 **
Residuals  9 13.776  1.5307
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Analysis of Variance Table

Response: y3
          Df Sum Sq Mean Sq F value   Pr(>F)
x3         1 27.470 27.4700  17.972 0.002176 **
Residuals  9 13.756  1.5285
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Analysis of Variance Table

Response: y4
          Df Sum Sq Mean Sq F value   Pr(>F)
x4         1 27.490 27.4900  18.003 0.002165 **
Residuals  9 13.742  1.5269
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

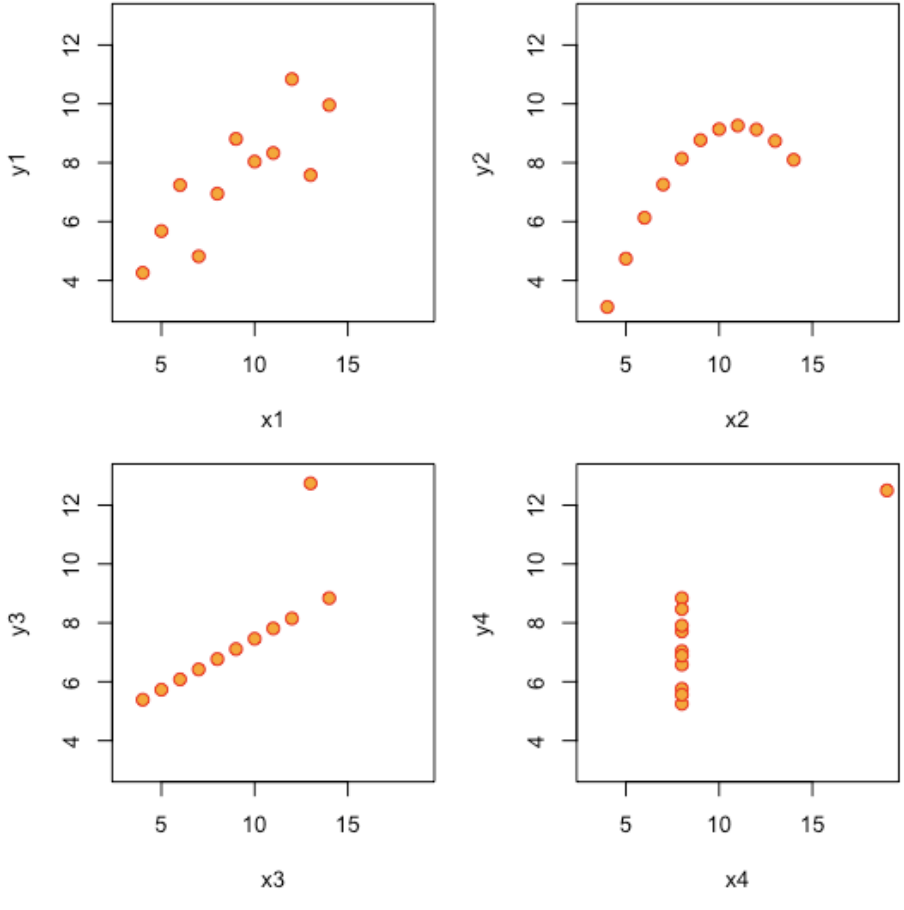
Anscombe example (1973)

	lm1	lm2	lm3	lm4
(Intercept)	3.0000909	3.000909	3.0024545	3.0017273
x1	0.5000909	0.500000	0.4997273	0.4999091

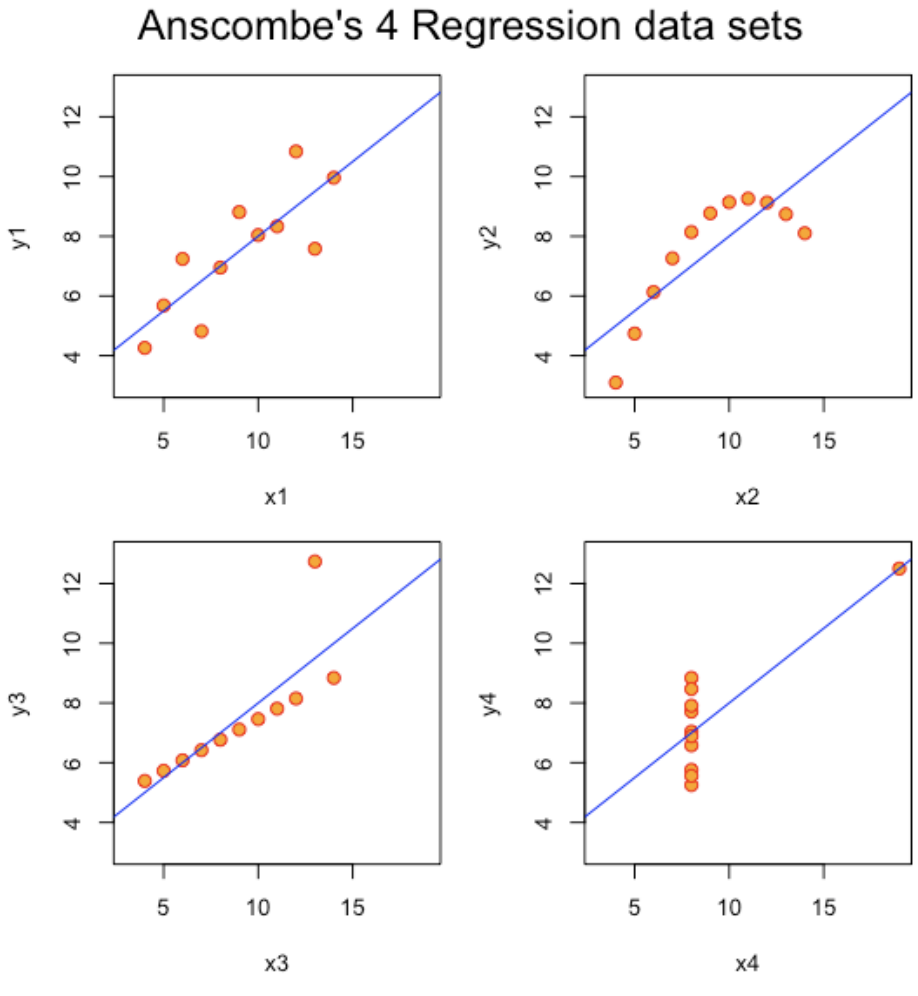
Anscombe example (1973)

```
$lm1
             Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 3.0000909  1.1247468 2.667348 0.025734051
x1          0.5000909  0.1179055 4.241455 0.002169629

$lm2
             Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 3.000909  1.1253024 2.666758 0.025758941
x2          0.500000  0.1179637 4.238590 0.002178816

$lm3
             Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 3.0024545  1.1244812 2.670080 0.025619109
x3          0.4997273  0.1178777 4.239372 0.002176305

$lm4
             Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 3.0017273  1.1239211 2.670763 0.025590425
x4          0.4999091  0.1178189 4.243028 0.002164602
```

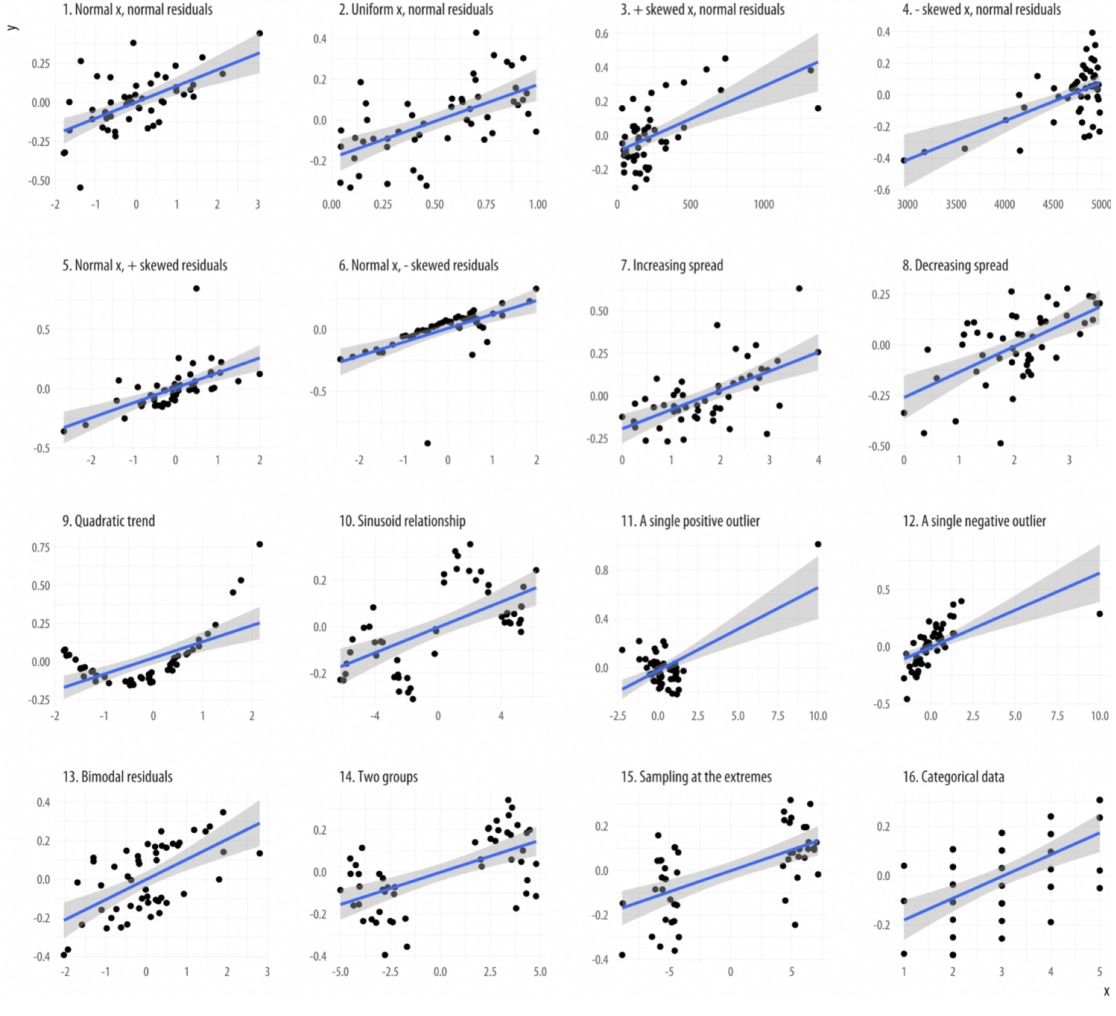
Tufte: Same relationship? (2001)



Tufte: Same relationship? (2001)



Jan Vanhove example (2016)



Elements of a Chart

1. Dimensionality
    1. How many dimensions are there?
2. Relationships
    1. Strength
    2. Fit
    3. Error bands
    4. Panels