Why do we need to visualize ? The Anscombe's Quartet

$X^{(1)}$	$Y^{(1)}$
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.24
7.00	4.82
5.00	5.68

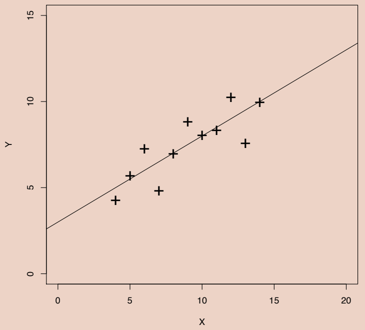
$N = 11$ samples
Mean of $X = 9.0$
Mean of $Y = 7.5$

Correlation $= 0.816$

Why do we need to visualize ? The Anscombe's Quartet

$X^{(1)}$	$Y^{(1)}$
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.24
7.00	4.82
5.00	5.68

Scatter plot



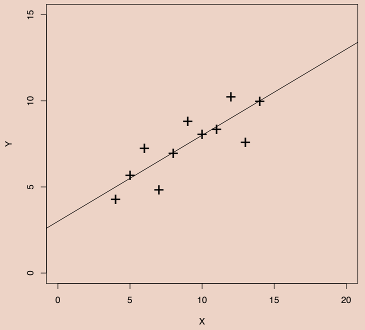
$N = 11$ samples
Mean of X = 9.
Mean of Y = 7
Intercept = 3
Slope = 0.5
Res. stdev = 1.237
Correlation = 0.816

Why do we need to visualize ? The Anscombe's Quartet

$X^{(1)}$	$Y^{(1)}$
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.24
7.00	4.82
5.00	5.68

$N = 11$ samples
Mean of $X = 9.$
Mean of $Y = 7$
Intercept = 3
Slope = 0.5
Res. stdev = 1.237
Correlation = 0.816

Scatter plot



1. The data set "behaves like" a linear curve with some scatter;
2. There is no justification for a more complicated model (e.g., quadratic);
3. There are no outliers;
4. The vertical spread of the data appears to be of equal height irrespective of the X-value; this indicates that the data are equally-precise throughout and so a "regular" (that is, equi-weighted) fit is appropriate.

Why do we need to visualize ? The Anscombe's Quartet

$X^{(1)}$	$Y^{(1)}$
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.24
7.00	4.82
5.00	5.68

$N = 11$ samples
Mean of $X = 9.0$
Mean of $Y = 7.5$
Intercept $= 3$
Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

$X^{(2)}$	$Y^{(2)}$
10.00	9.14
8.00	8.14
13.00	8.74
9.00	8.77
11.00	9.26
14.00	8.10
6.00	6.13
4.00	3.10
12.00	9.13
7.00	7.26
5.00	4.74

$N = 11$ samples
Mean of $X = 9.0$
Mean of $Y = 7.5$
Intercept $= 3$
Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

$X^{(3)}$	$Y^{(3)}$
10.00	7.46
8.00	6.77
13.00	12.74
9.00	7.11
11.00	7.81
14.00	8.84
6.00	6.08
4.00	5.39
12.00	8.15
7.00	6.42
5.00	5.73

$N = 11$ samples
Mean of $X = 9.0$
Mean of $Y = 7.5$
Intercept $= 3$
Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

$X^{(4)}$	$Y^{(4)}$
8.00	6.58
8.00	5.76
8.00	7.71
8.00	8.84
8.00	8.47
8.00	7.04
8.00	5.25
19.00	12.50
8.00	5.56
8.00	7.91
8.00	6.89

$N = 11$ samples
Mean of $X = 9.0$
Mean of $Y = 7.5$
Intercept $= 3$
Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

Why do we need to visualize ? The Anscombe's Quartet

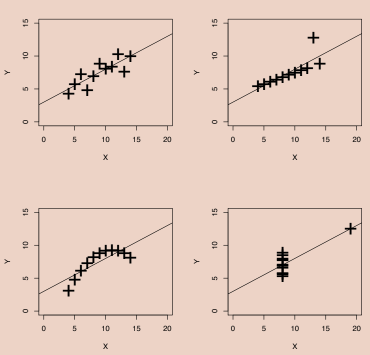
$X^{(1)}$	$Y^{(1)}$
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.24
7.00	4.82
5.00	5.68

$X^{(2)}$	$Y^{(2)}$

$X^{(3)}$	$Y^{(3)}$

$X^{(4)}$	$Y^{(4)}$

Scatter plot



$N = 11$ samples
Mean of $X = 9.$
Mean of $Y = 7$
Intercept $= 3$
Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

Why do we need to visualize ? The Anscombe's Quartet

$X^{(1)}$	$Y^{(1)}$
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.24
7.00	4.82
5.00	5.68

$N = 11$ samples
Mean of $X = 9$
Mean of $Y = 7$
Intercept $= 3$
Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

$X^{(2)}$	$Y^{(2)}$

Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

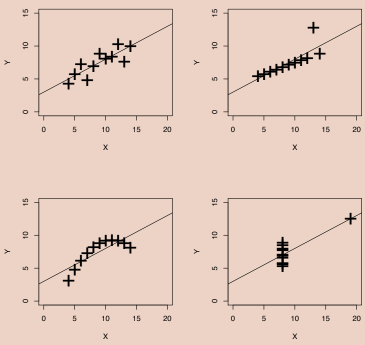
$X^{(3)}$	$Y^{(3)}$

Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

$X^{(4)}$	$Y^{(4)}$

Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

Scatter plot



1. data set 1 is clearly linear with some scatter.
2. data set 2 is clearly quadratic.
3. data set 3 clearly has an outlier.
4. data set 4 is obviously the victim of a poor experimental design with a single point far removed from the bulk of the data "wagging the dog".

Problem statement

- All analysis we perform rely on (sometimes implicit) assumptions. If these assumptions do not hold, the analysis will be a complete non-sense.
- Checking these assumptions is not always easy and sometimes, it may even be difficult to list all these assumptions and formally state them.

  A visualization can help to check these assumptions.
- Visual representation resort to our cognitive faculties to check properties.

  The visualization is meant to let us detect expected and unexpected behavior with respect to a given model.

Using the “right” representations

- The problem is to represent on a limited space, typically a screen with a fixed resolution, a meaningful information about the behavior of an application or system.
- $\rightsquigarrow$ need to aggregate data and be aware of what information loss this incurs.
- Every visualization emphasizes some characteristics and hides others. Being aware of the underlying models helps choosing the right representation.