KU LEUVEN

Getting to Know Your Data

Jesse Davis

Outline

- Computing simple statistics

- Simple visualizations

KU LEUVEN

Motivation

Exploring data is vitally important as it helps you think about

- What technique to use
- What challenges you may (will) encounter
- What flaws or problems there are with data
- What results to expect

KU LEUVEN

What Should You Look For?

- Good to look at simple statistics of
  - Number of variables
  - Size of data
  - Missing values
  - Class skew
- For each attribute, look at
  - **Discrete:** number of possible values, are they ordered, frequency of each value, etc.
  - **Numeric:** range, mean, min, max, etc.

KU LEUVEN

Summary Statistics

- **Mean:** $\mu_X = \frac{\sum_i x_i}{n}$
- **Mode:** Most common value
- **Median:** Value v s.t. half the values above v, half below v
- **Variance:** $\sigma^2 = \frac{\sum_i (x_i - \mu_X)^2}{n}$
- **Quartile:** Sort X
  - Q1: Value at position 0.25n
  - Q3: Value at position 0.75n
  - IQR: Q3 – Q1

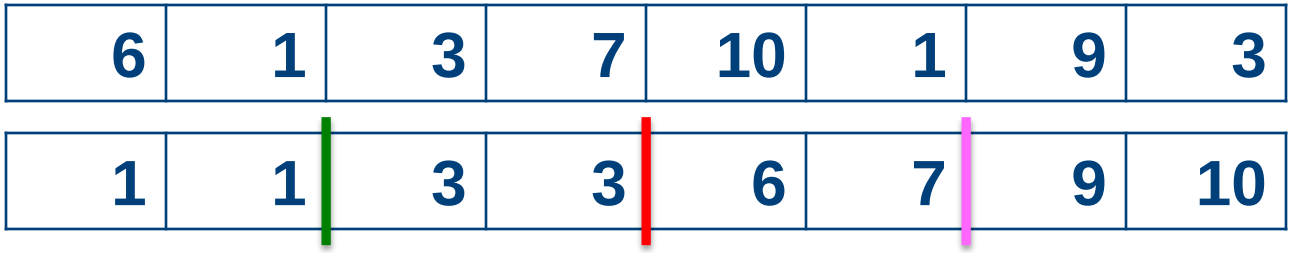
Quiz

6	1	3	7	10	1	9	3

What is the:

- Mean:
- Mode:
- Median:
- Q1:
- Q3:

Quiz



What is the:

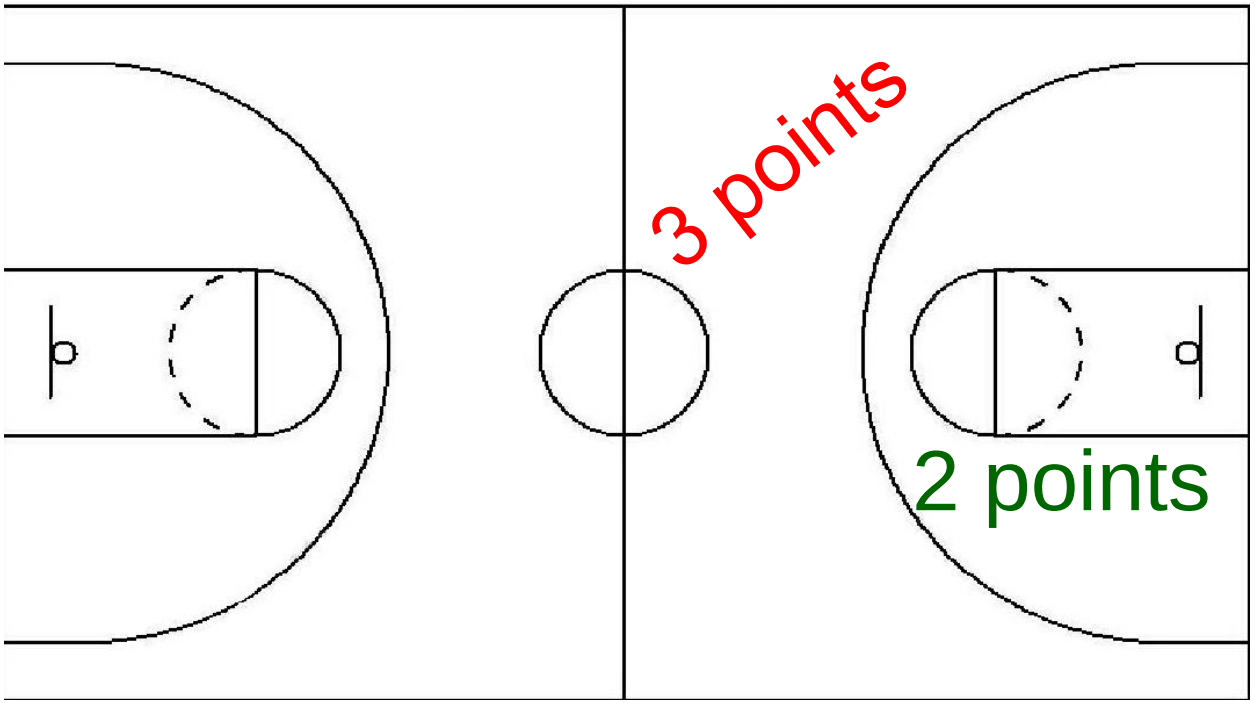
- Mean: 5
- Mode: 1, 3
- Median: 4.5
- Q1: 2
- Q3: 8

Weighted Averages

- Mean assumes that each data point is of equal importance
- Often some data points are more important
  - Estimate more reliable
  - More valuable
  - Etc.
- Weighted average: $\mu_X = \frac{\sum_i w_i x_i}{\sum_i w_i}$

KU LEUVEN

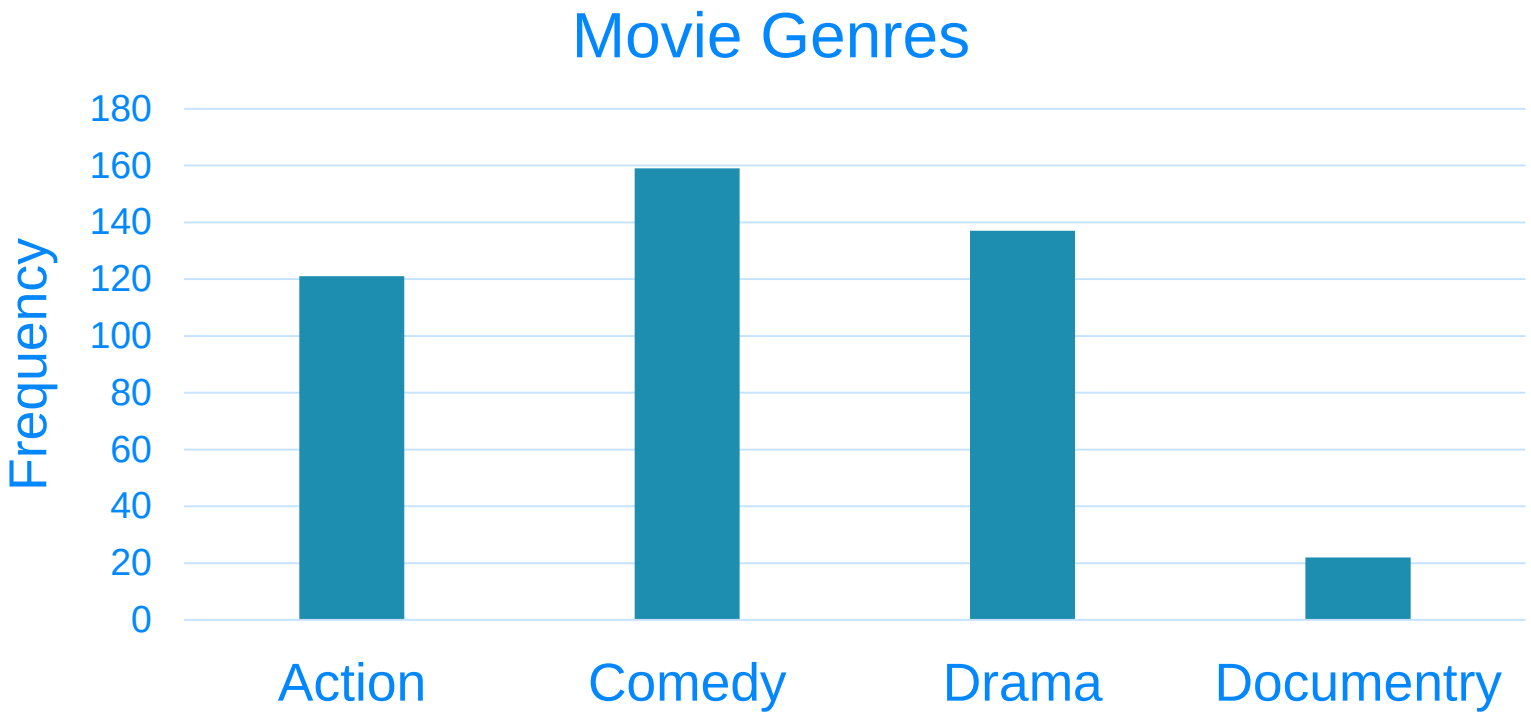
Example: Basketball



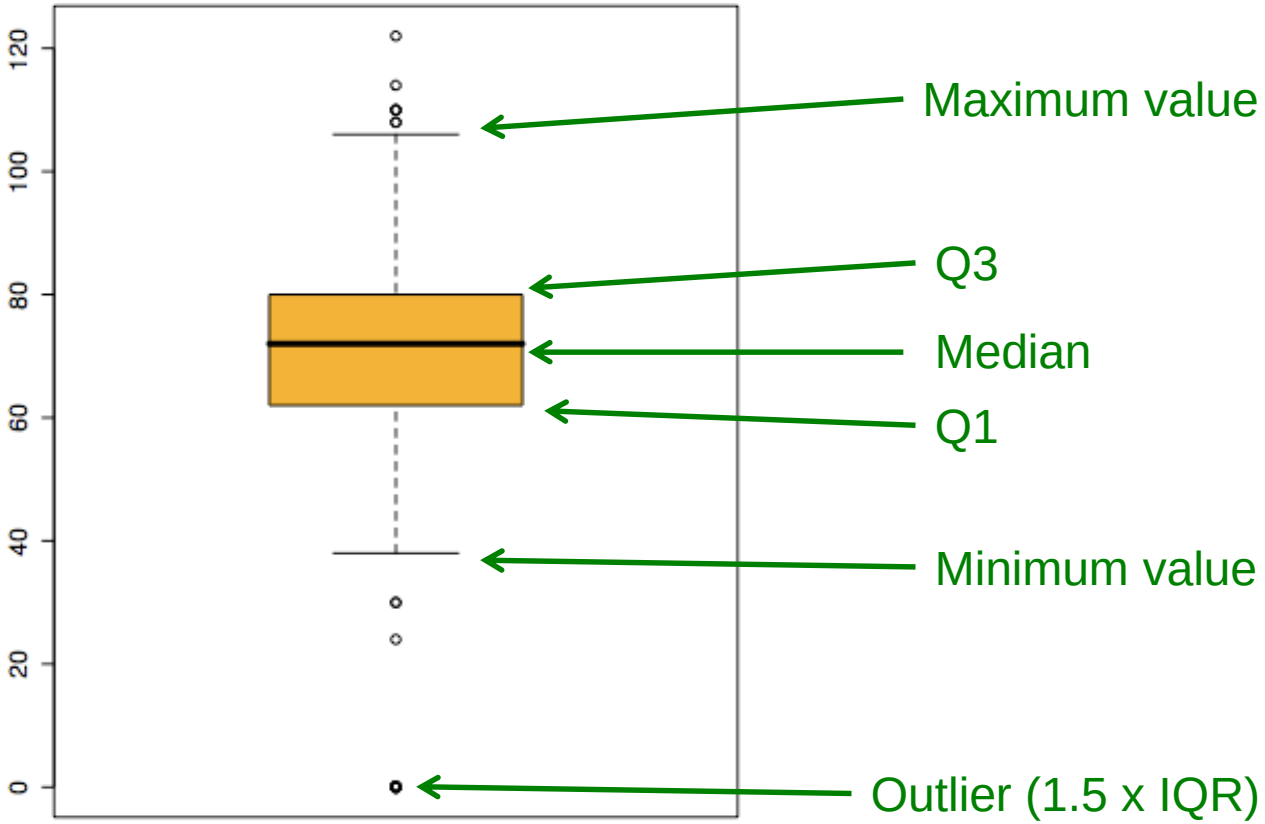
$$FG\ \% = \frac{2p + 3p}{2pa + 3pa}$$

$$eFG\ \% = \frac{2p + 1.5 * 3p}{2pa + 3pa}$$

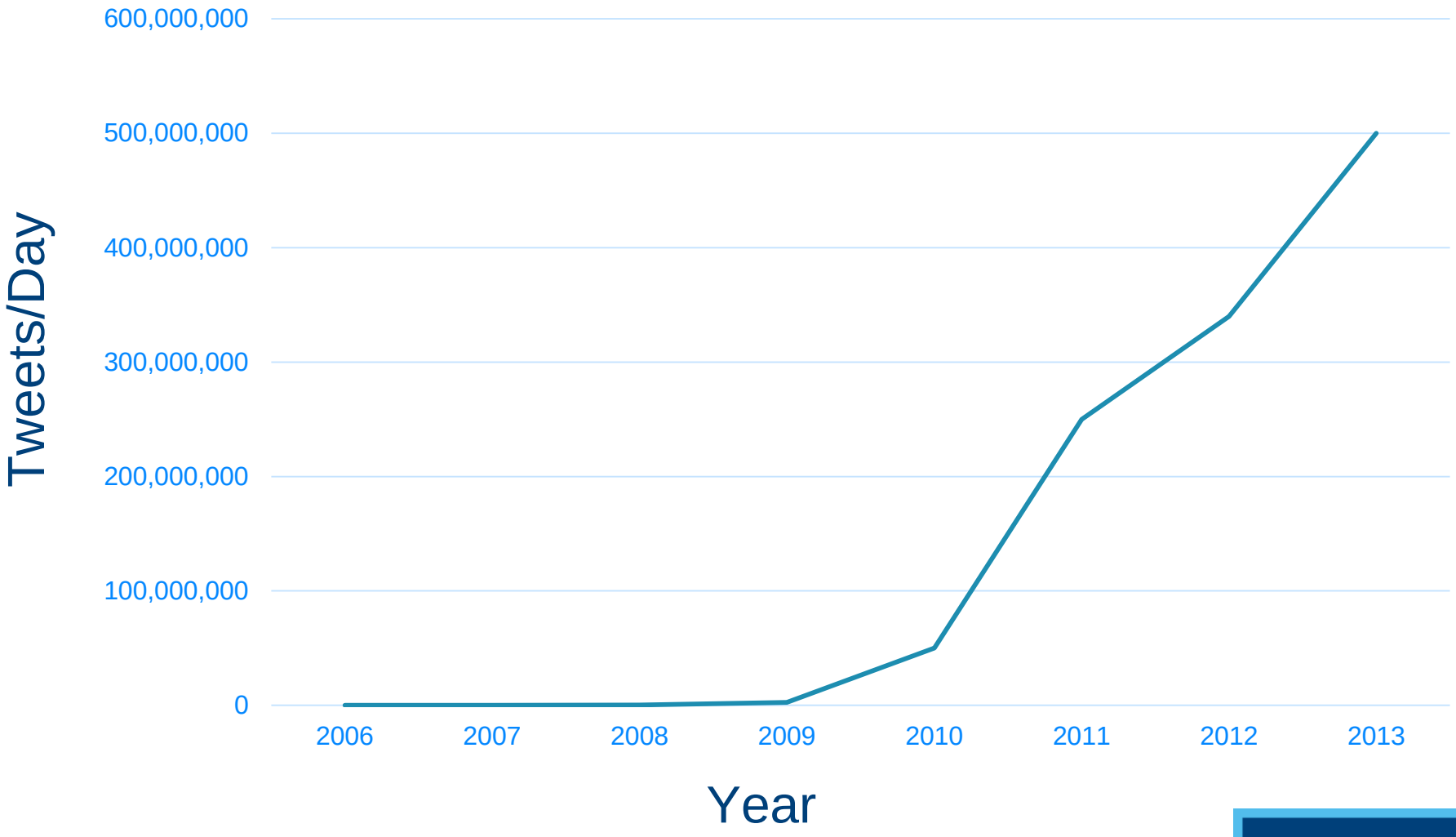
Histograms


Movie Genres
Frequency
180
160
140
120
100
80
60
40
20
0
Action
Comedy
Drama
Documentry

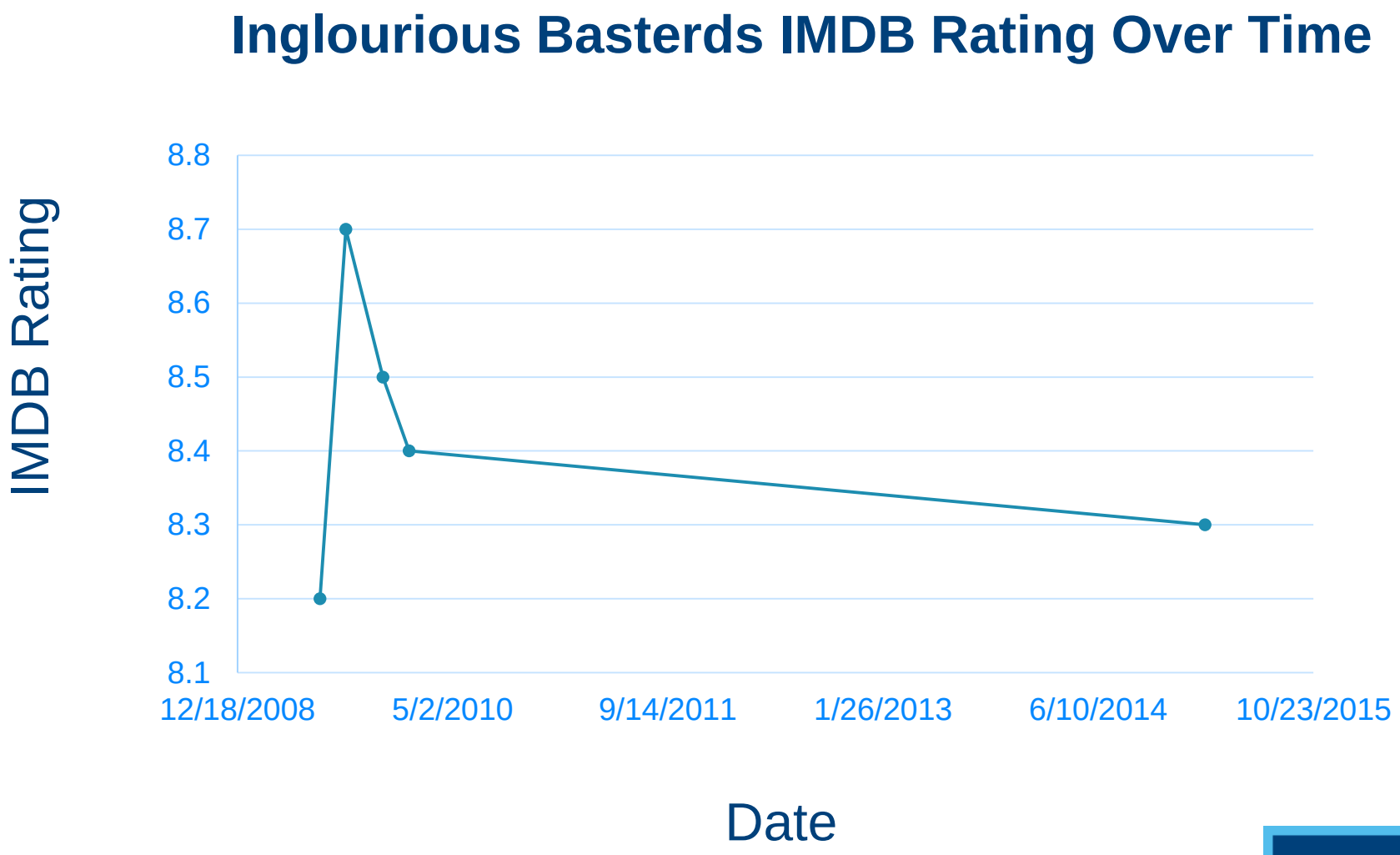
Boxplots


0
20
40
60
80
100
120
Maximum value
Q3
Median
Q1
Minimum value
Outlier (1.5 x IQR)

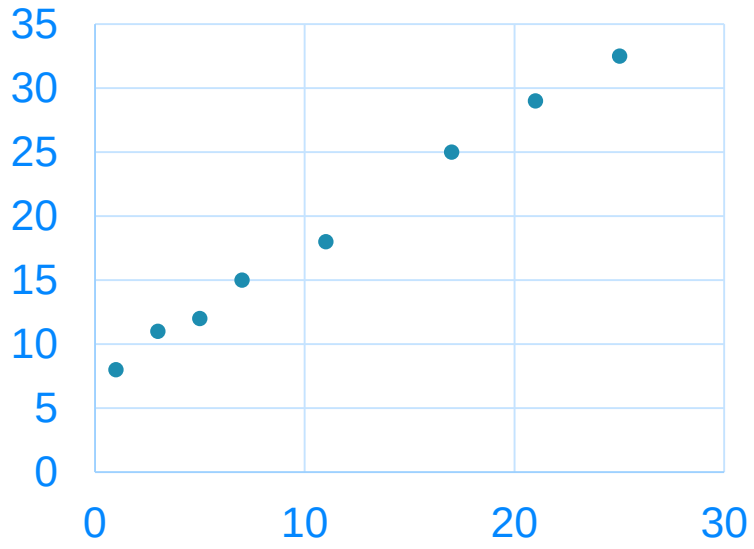
Time Series

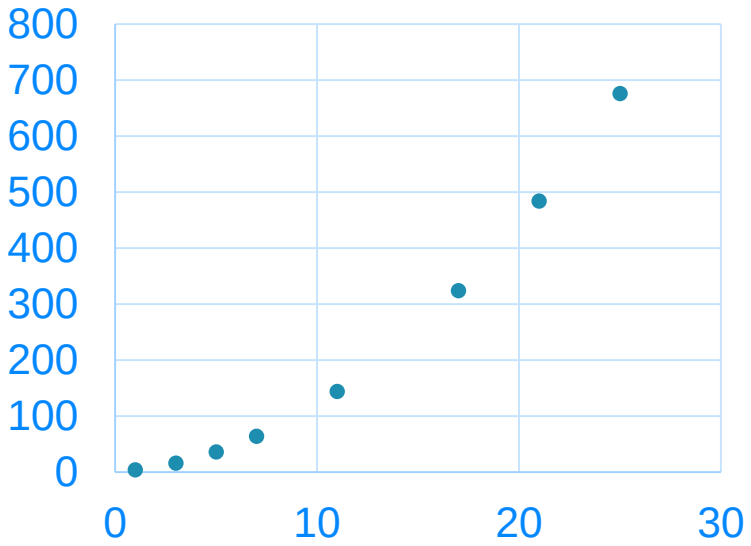

600,000,000
500,000,000
400,000,000
300,000,000
200,000,000
100,000,000
0
Tweets/Day
2006
2007
2008
2009
2010
2011
2012
2013
Year

Time Series


Inglourious Basterds IMDB Rating Over Time
IMDB Rating
8.8
8.7
8.6
8.5
8.4
8.3
8.2
8.1
12/18/2008
5/2/2010
9/14/2011
1/26/2013
6/10/2014
10/23/2015
Date

Scatterplots





KU LEUVEN

Spatial

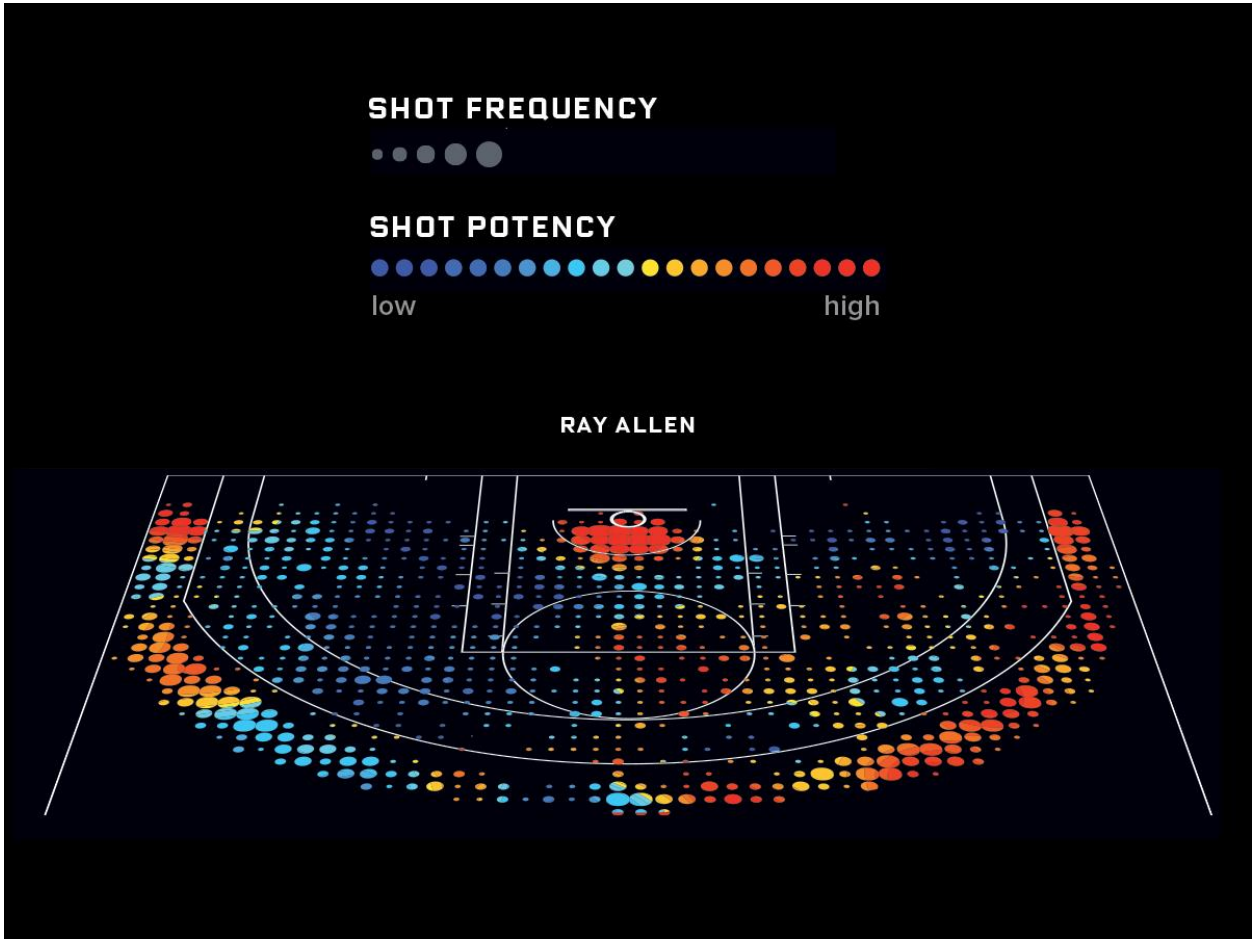


Image from: http://www.wired.com/2014/10/faster-higher-stronger/