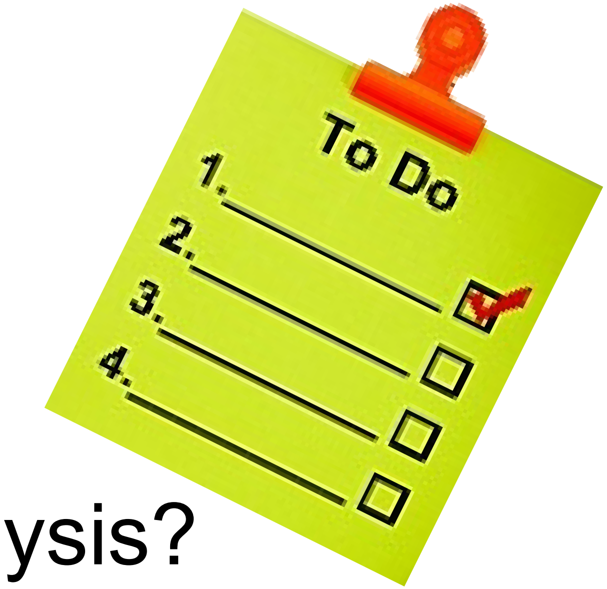
Research Methods

Exploratory Data Analysis

Dr. Sven Magg, Prof. Dr. Stefan Wermter

http://www.informatik.uni-hamburg.de/WTM/

Plan for today!



1. What is exploratory data analysis?
2. Descriptive Statistics for one variable
3. Measures for central tendency, shape and dispersion
4. Useful uni-variate visualisations
5. The Big Apple Experiment!

Exploratory Data Analysis

- Fundamental Model of Data: $y = f(x, \varepsilon)$
  - what factors strongly or weakly influence y and how do they combine?
  - is there evidence of important factors in ε, maybe hidden factors?
- Once we have evidence, we can use confirmatory studies to test whether x really is a causal factor influencing y

- EDA helps to
  - find the causal story hidden in the data ($\Rightarrow$ modelling)
  - understand phenomena and find structure in the data

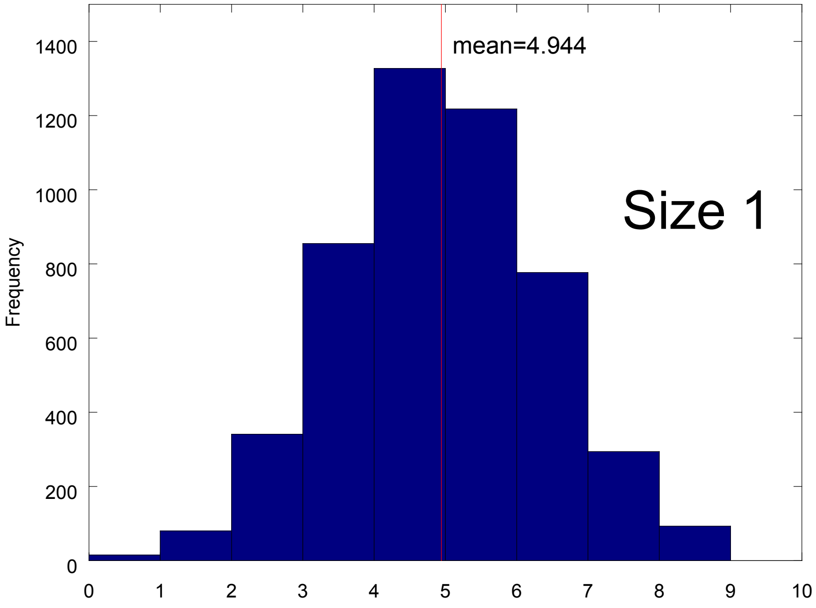
Exploratory Data Analysis

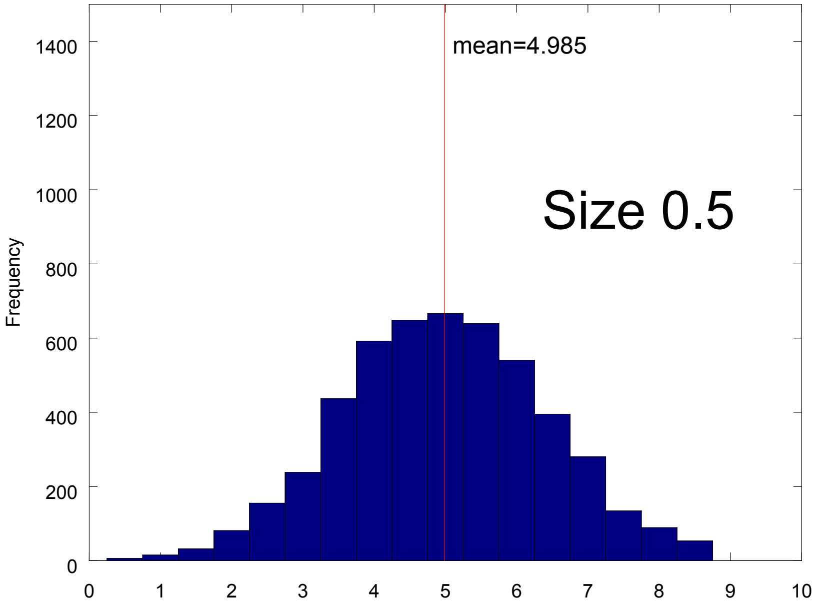
- Structure in data is evidence of causal influences
- EDA can uncover and clarify this structure
- "What do I see and what does it mean?"

- EDA needs lots of practice!
  - Depending on the start, you might end up on different paths of exploration
  - Misinterpretations can cost time
  - EDA is like archaeology: You find a stone and it might be a fossil or some petrified dirt. It needs a good eye to spot the difference!

Univariate visualisations

- Frequency histogram
  - Display relative frequency of values in data
  - works with all data scales
  - Values are “binned” into a number of bins





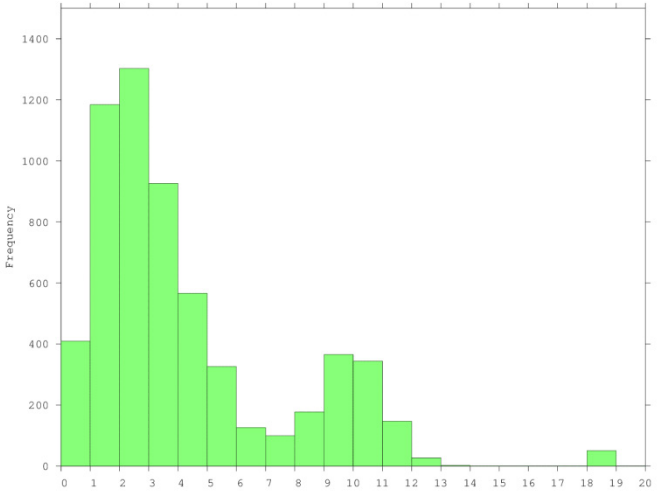
Descriptive Statistics

- Measures of central tendency: Mean, median, mode
- Mean: Arithmetic mean
- Median
  - Element that splits a ordered sample in two equal parts
  - With even number of elements (3,4,5,6), either,
    - Take average of nearest neighbours, i.e. 4.5, or
    - select one of the nearest neighbours, i.e. 4
- What's better, median or mean?
  - Outliers: Mean is sensitive, median is robust
  - Possible solution: Trimmed mean (trim top and bottom of ordered list, calculate mean of the rest)
  - Median can be used with ordinal data!

Measures of central tendency

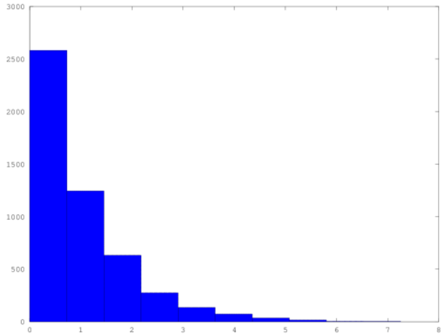
- Mode
  - Most common value in a distribution (mode(1,2,2,6) = 2)
  - With continuous data, often not useful
  - With binned data, denotes bin with most values
  - Often used to denote number of areas with high frequency
- In a symmetric, unimodal distribution, all 3 are the same

- Example:
  - Bi-modal or tri-modal distribution
  - Mean expected to be right of median
  - Where is the calculated mode?

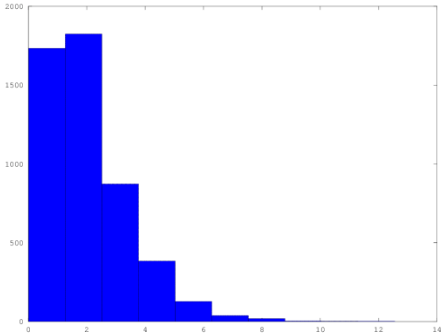


Measures of shape

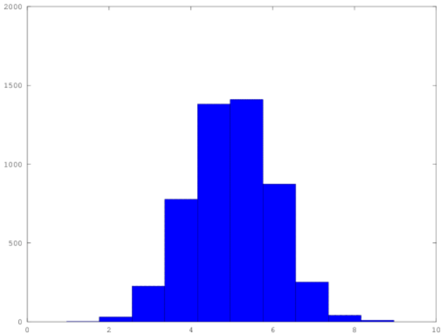
- Skew
  - Measures lack of symmetry
  - Positively skewed: frequent values on left with tail to the right
  - Negatively skewed: frequent values on right with tail to the left



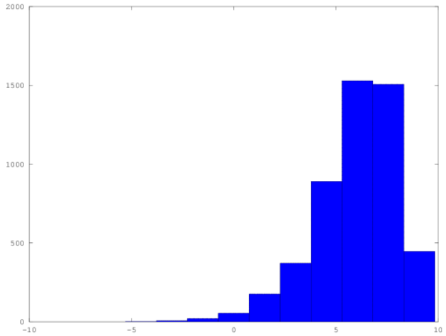
1.808



1.366



0.038

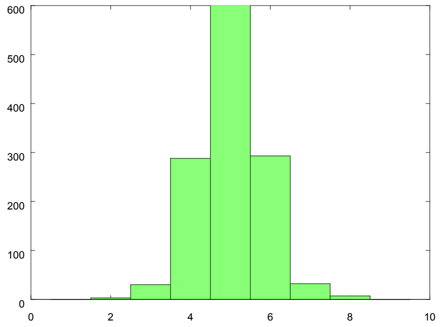


-0.985

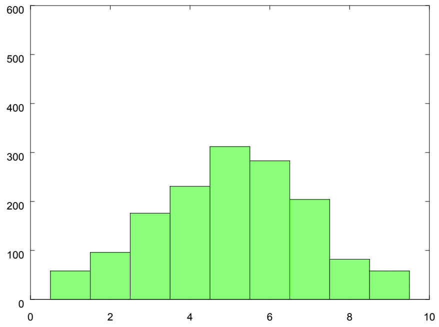
  - Skew measures deviation from a symmetric distribution
  - Skew positive: Left-skewed, negative: Right-skewed

Measures of shape

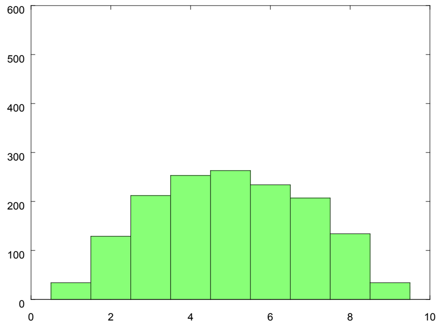
- Kurtosis
  - Measures weight of tails
  - Leptokurtic vs. Platykurtic: Heavy vs. Light-tailed distribution



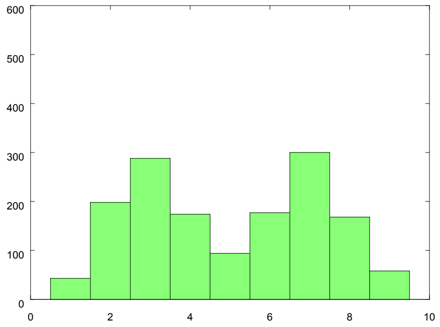
1.648



-0.036



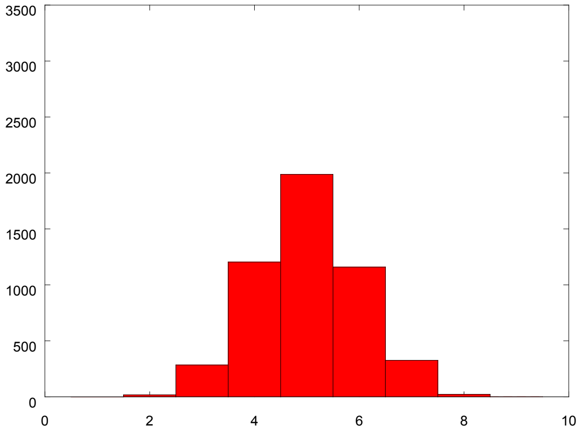
-0.816



-1.282

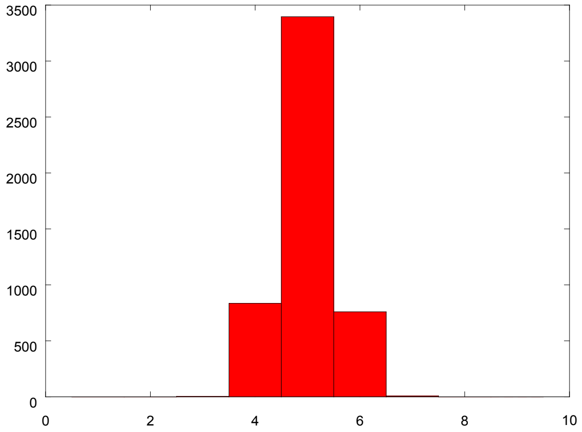
  - Weight of tails compared to normal distribution (kurtosis=0)

Measures of shape



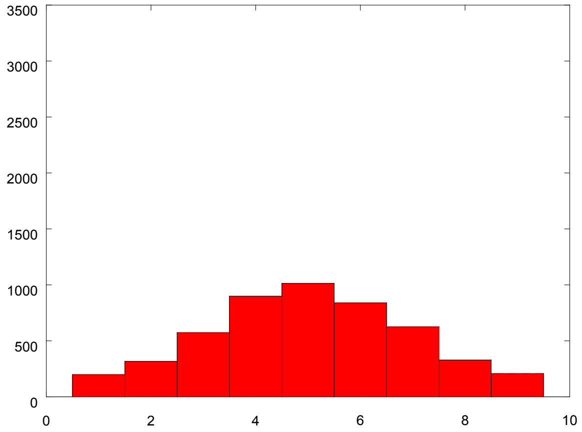
-0.056

s=0.982



-0.042

s=0.500



0.002

s=2.000

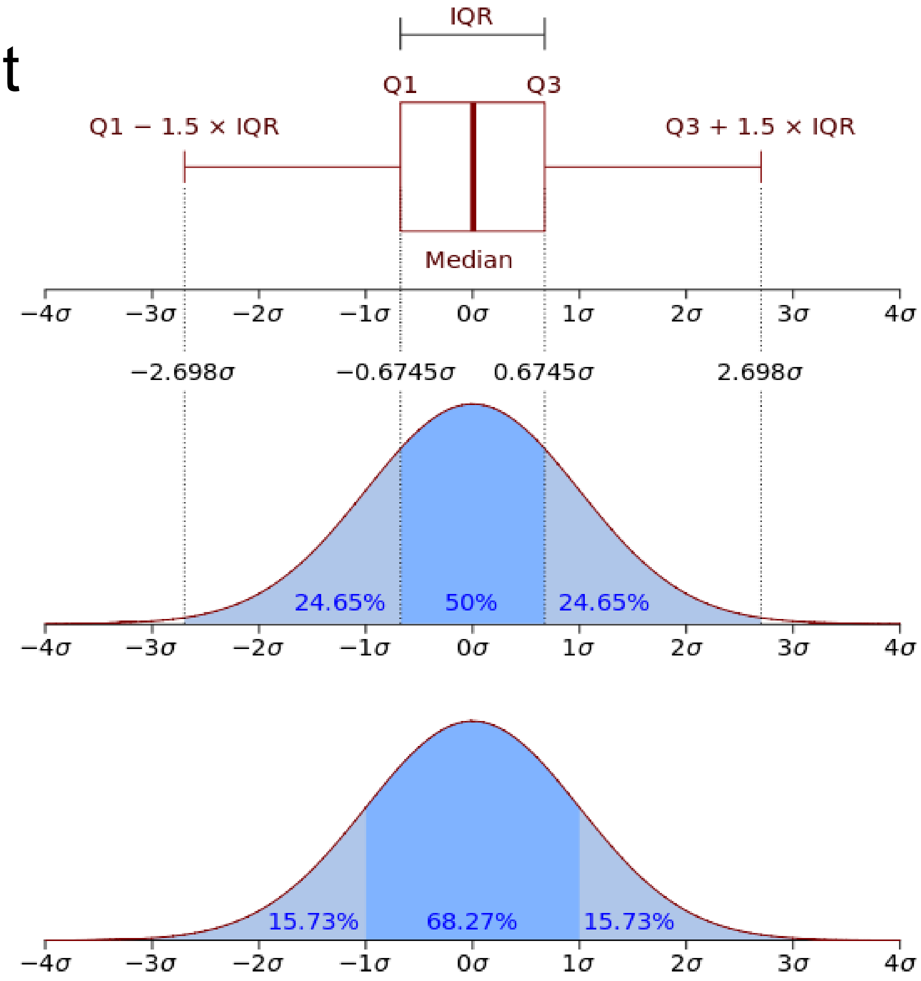
- Kurtosis ≠ Standard Deviation
- Skew and Kurtosis can be used to measure divergence from a normal distribution

Measures of dispersion

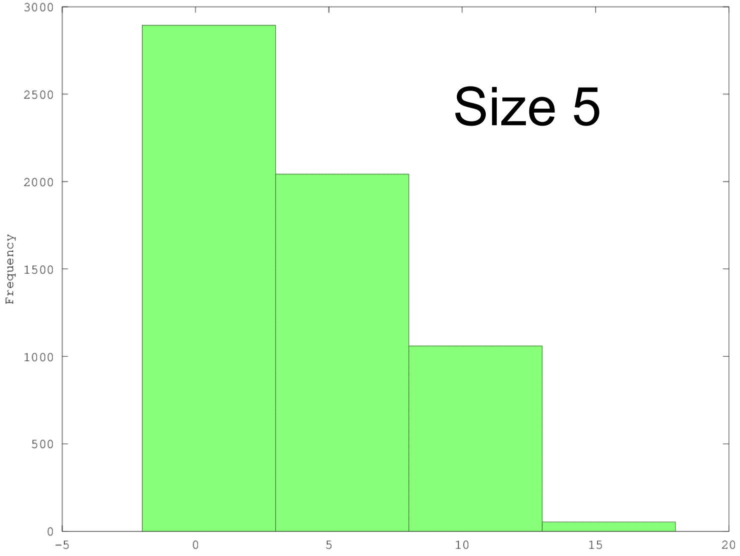
- Standard deviation & variance
- Maximum, minimum, Range
- Quartiles
  - Divide ordered distribution into four equal parts, quartiles are the values that split those parts
  - Quartiles are numbered in ascending order
  - Q2 = median(x), Q1=median(x| x<Q2), Q3=median(x| x>Q2)
  - e.g. (1,2,2,4,4,5,6,8,8,8,9,100) = (1,2,2),(4,4,5),(6,8,8),(8,9,100) $\Rightarrow$ Q1=3; Q2=5.5; Q3=8
  - Q1-3 also 25th, 50th and 75th percentile
  - Interquartile Range IQR: Range between Q1 and Q3
    in example: Range = 99, IQR=5

Measures of dispersion

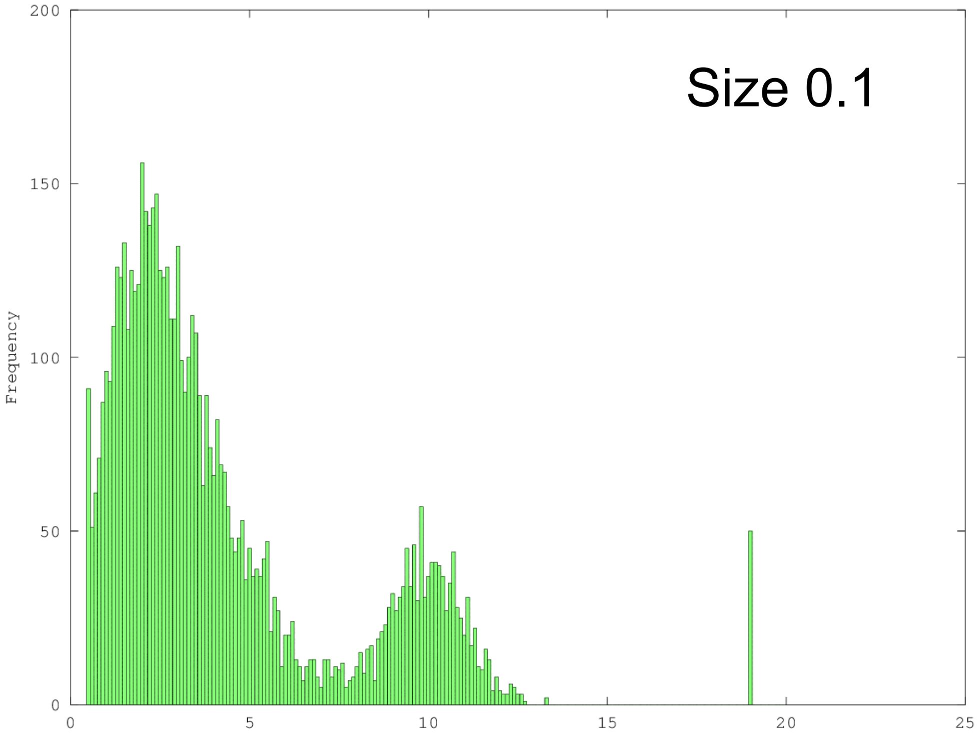
- Quartiles displayed in Boxplot
- Command to calculate statistics for one variable:
  - octave>statistics(x)
  - Displays:
    - minimum,
    - 1st, 2nd, and 3rd quartile,
    - maximum,
    - mean, standard deviation,
    - skewness, and kurtosis



More interesting data



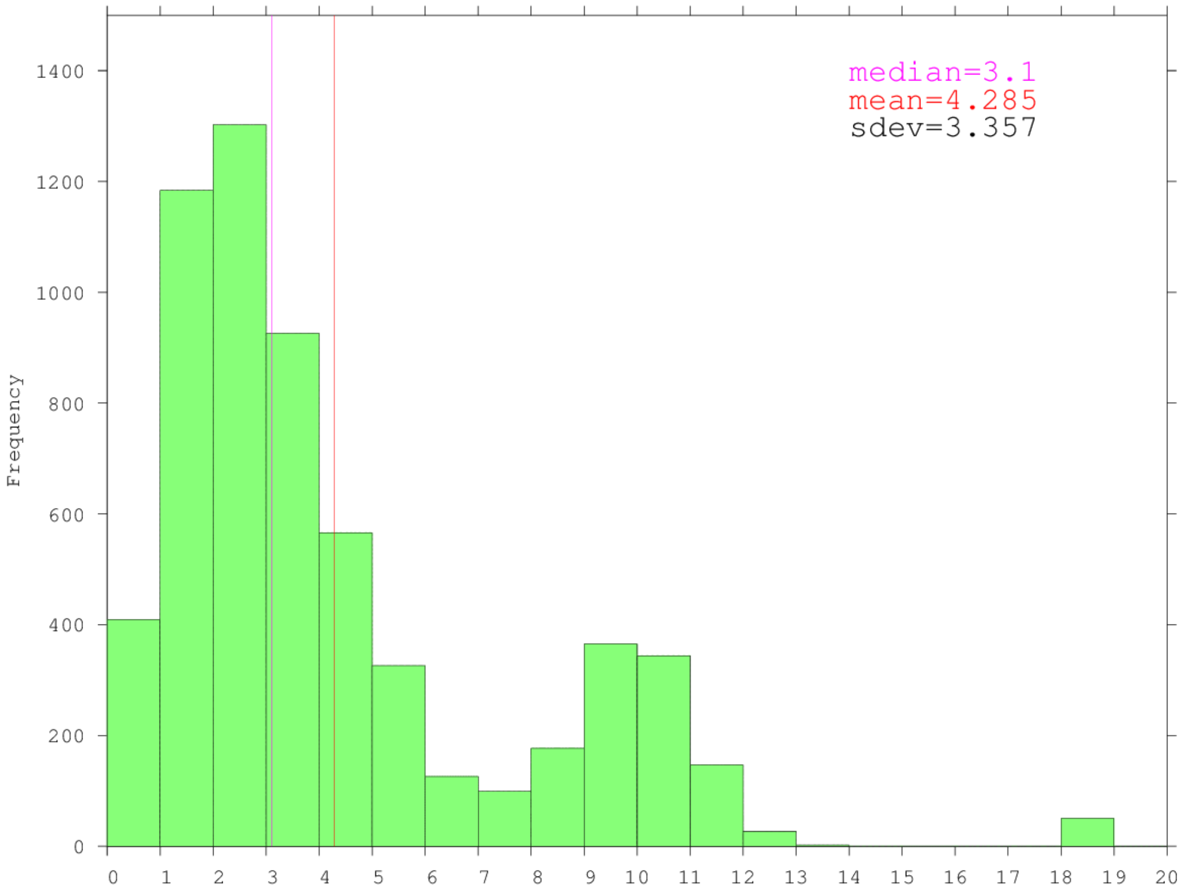
Too little detail



Too much detail

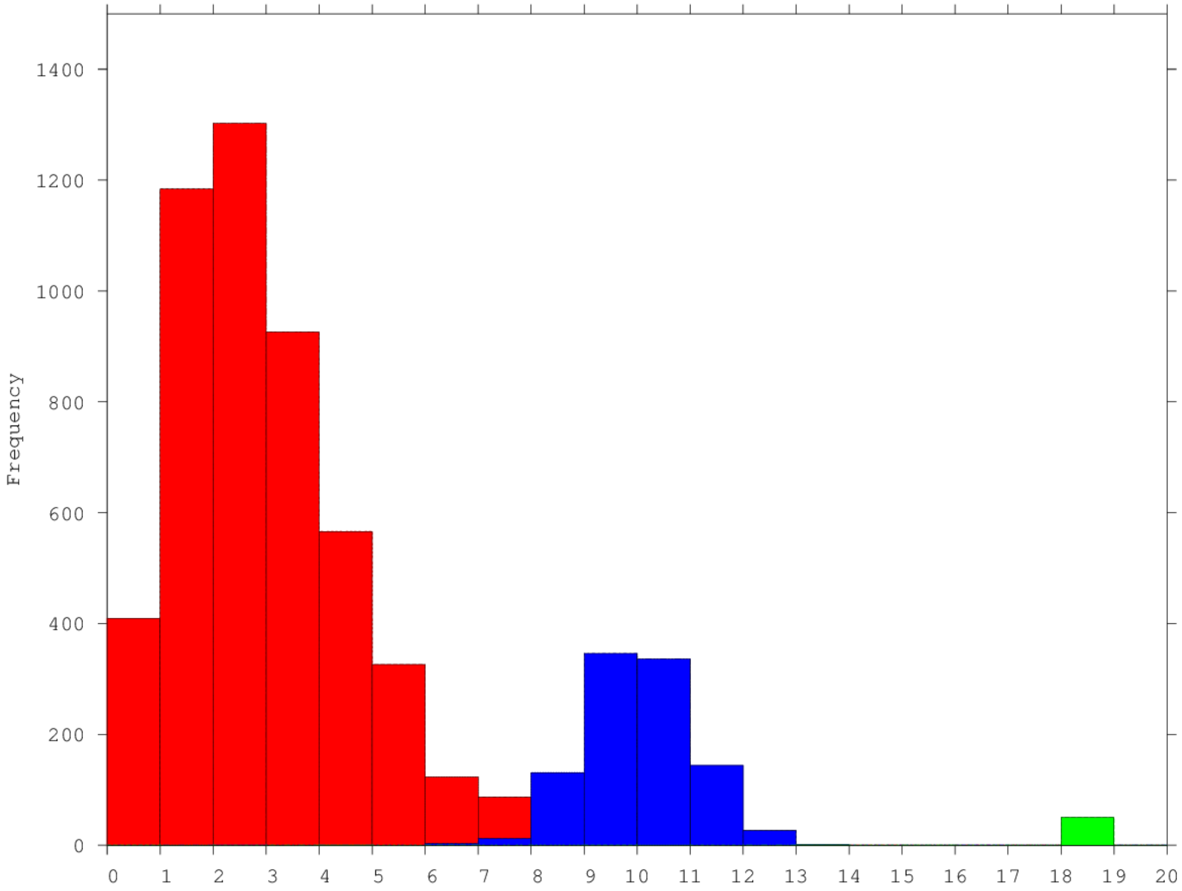
Bin size matters!

What can we see?



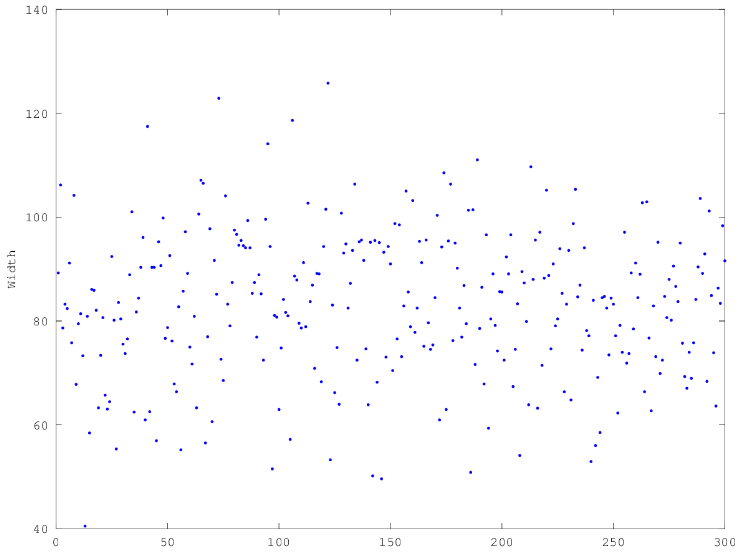
- 2 bumps
- 1 outlier bump
- Evidence of two sub populations?
- Evidence of another factor?
- Interesting areas around 2.5 and 10 or in between
- Don’t forget the outliers!

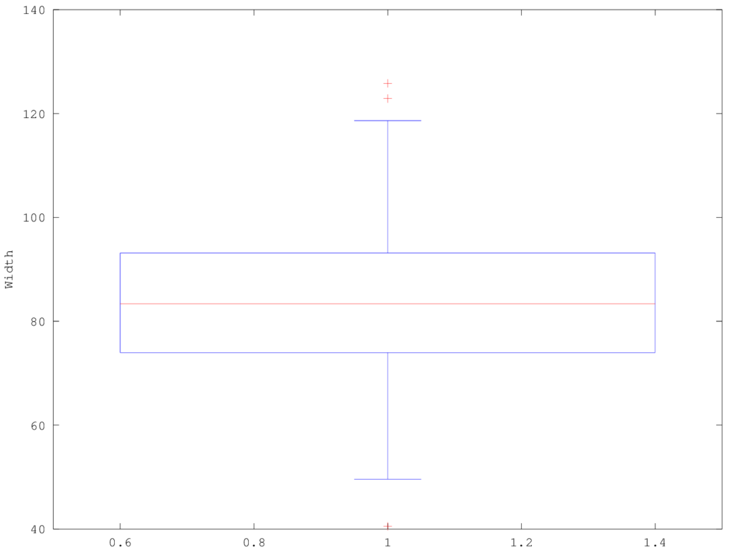
What can we see?

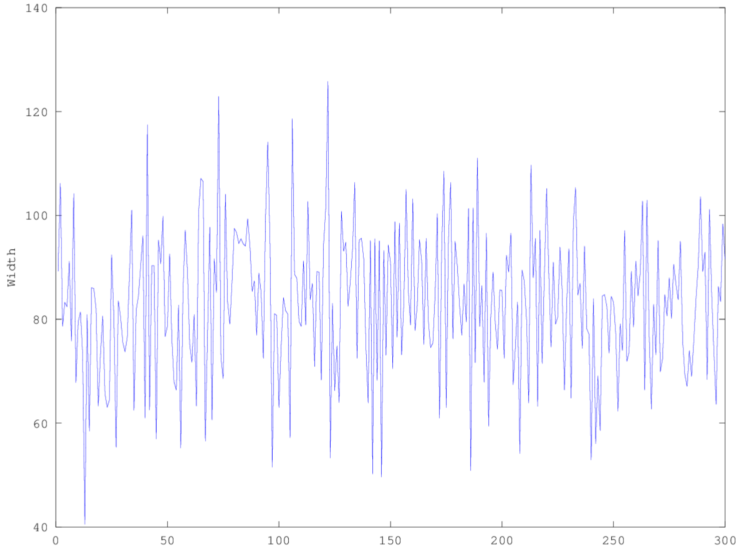


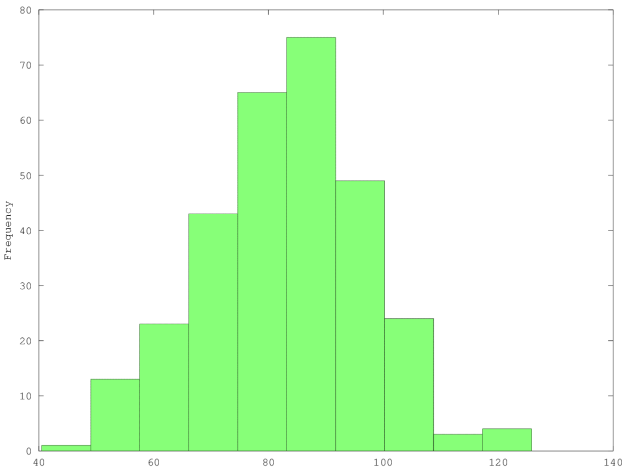
- Robot returned
  - Success: red
  - Failure: blue
- Experiment aborted: green
- Continue search with sub-matrices

Treerings



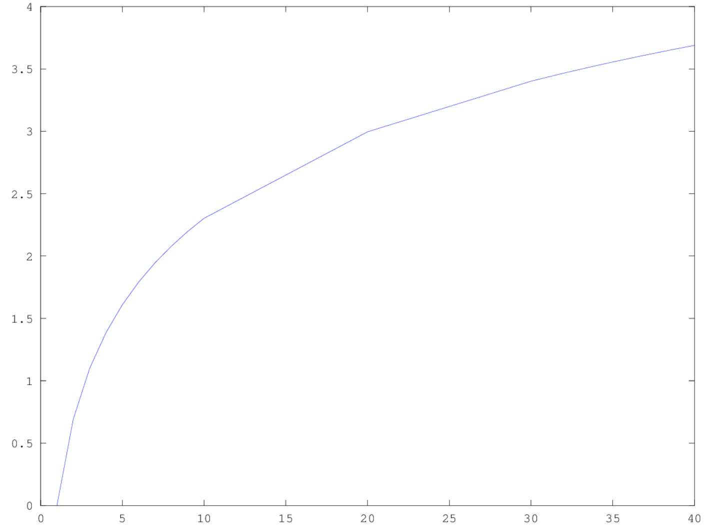


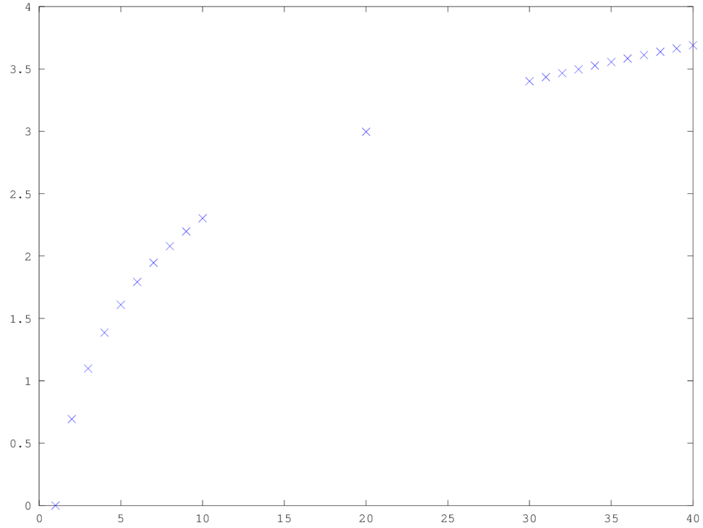


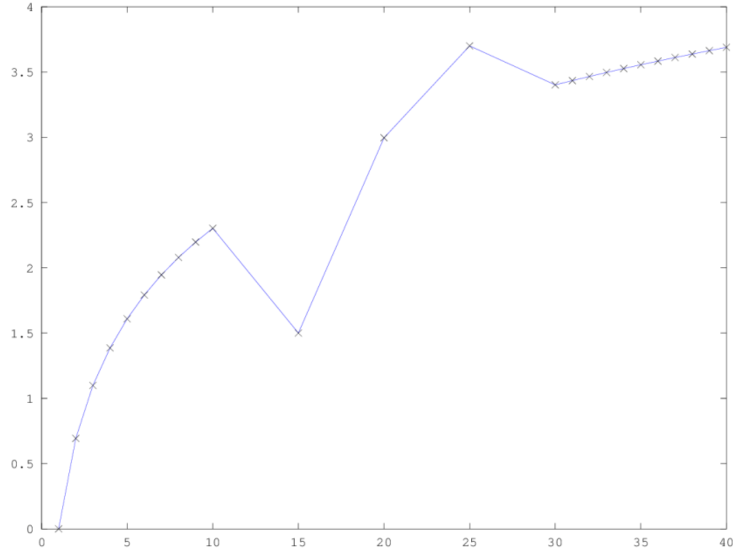


Min:	40.51
Q1:	73.98
Median:	83.36
Q2:	93.13
Max:	125.8
Mean:	82.98
sDev:	14.07
Skew:	-0.078
Kurt.:	0.043

Line or no line?

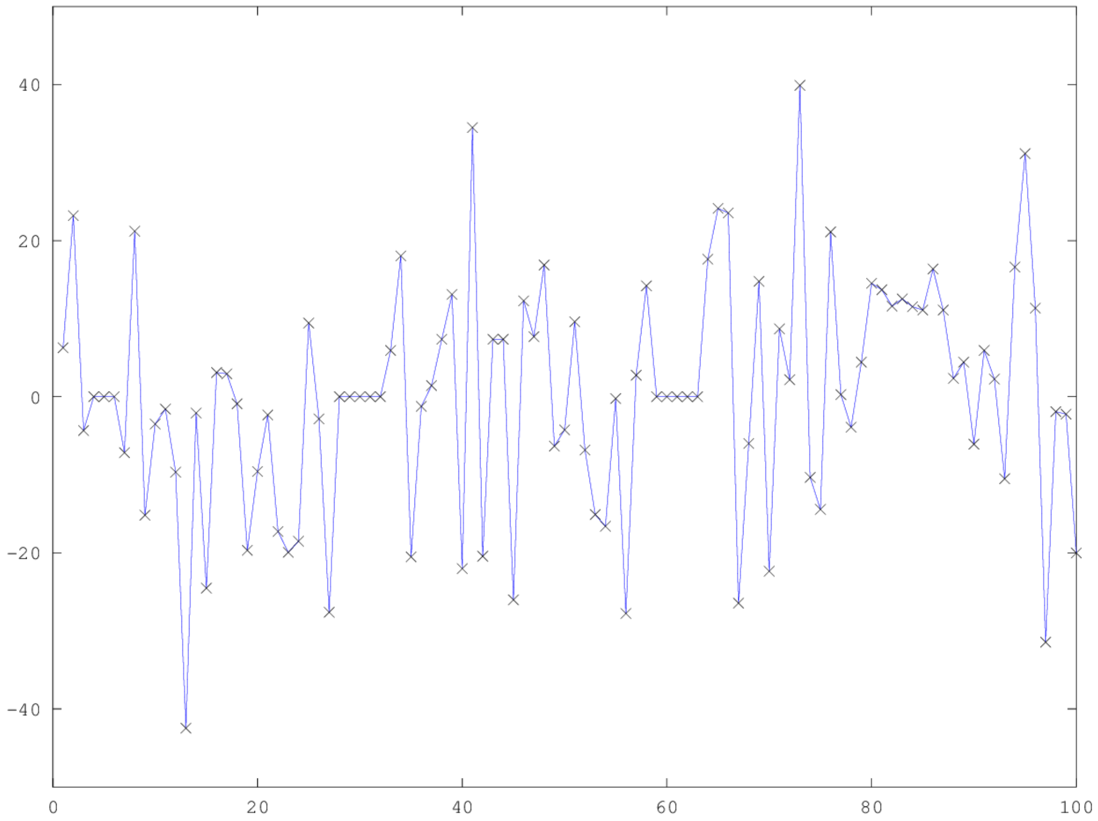






- Always plot data points!
- Watch out for interpolated gaps!

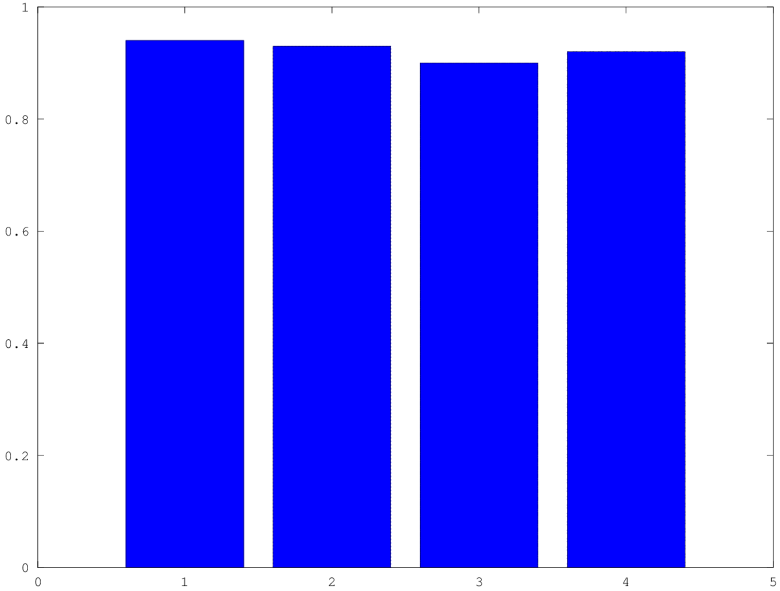
More lines

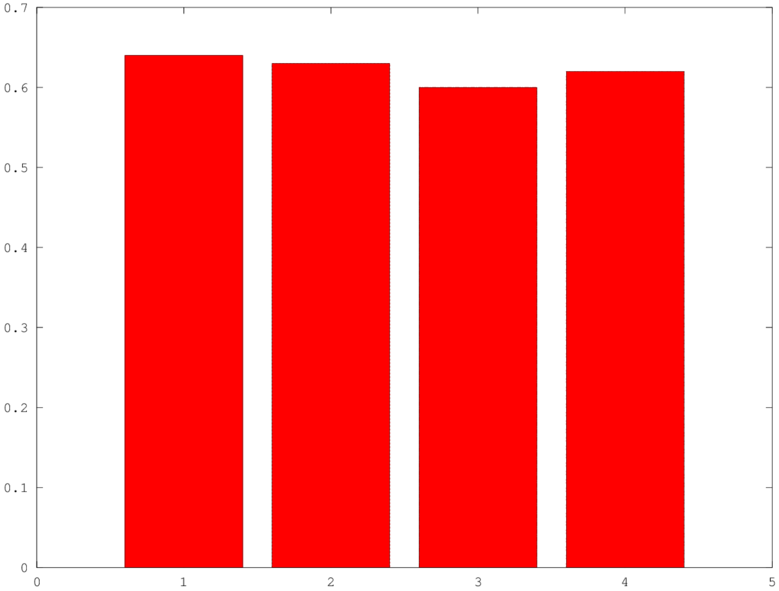


- Some flat areas at 0
- 0 is a special number!

- Evidence of missing data
- Better to use “NA”

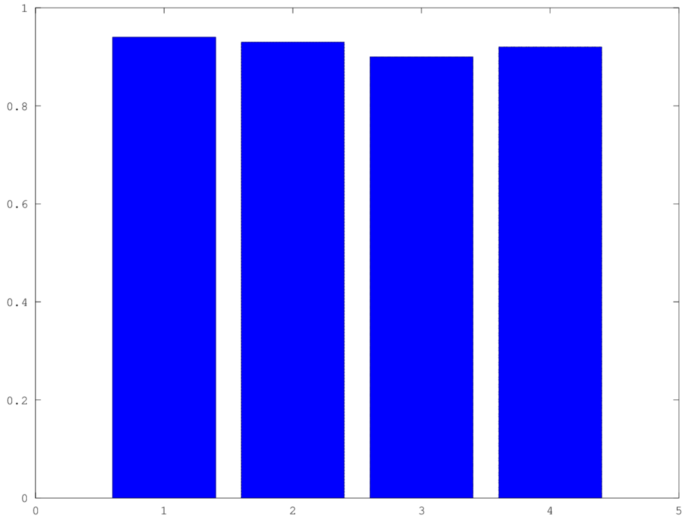
Comparing graphs

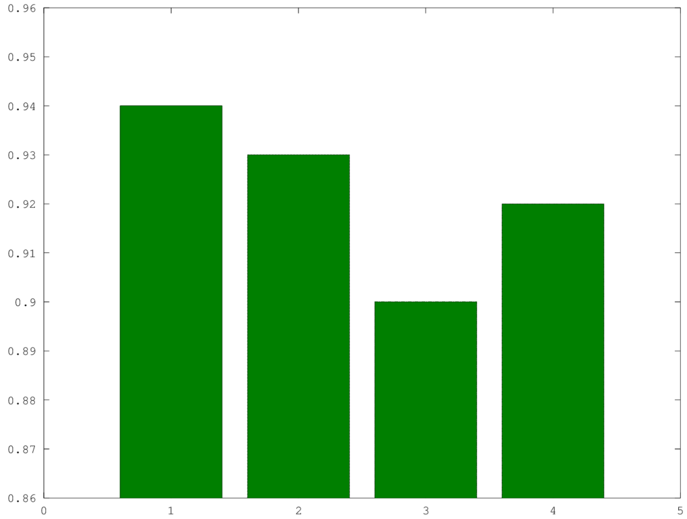


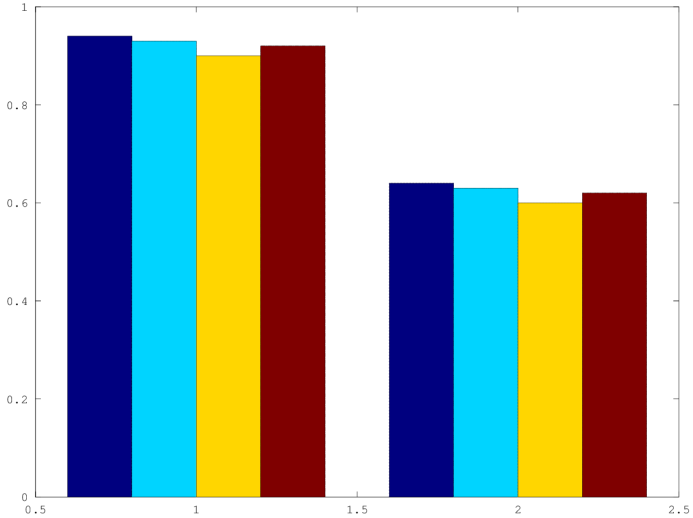


- Graphs look similar due to different axis
- Always make sure the axis match!

Comparing graphs







- If possible, plot into the same figure
- Open questions:
  - are the differences significant?

What have we learned?

1. EDA can reveal structure in data and help tell the causal story behind the data
2. With EDA you can find weak or strong influences of x on y, or identify hidden factors in ε
3. Tools that can be used for univariate distributions:
    1. Frequency histogram
    2. Line plots / Bar charts
    3. Descriptive statistics to quantify/check what you see
4. Look for and try to explain unusual phenomena
5. Use all tools available, like colours, transformations, etc.