Data Exploration Part 1

Lesson 1

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Data Exploration (Descriptive Statistics)

- Purpose: To gain a clear understanding of your data.
  - How large is it?
  - What columns are of interest?
  - Missing data?
  - Outliers?
  - Assumptions inherent in the data

Assumptions and why they matter:

W T F

Assumptions and why they matter:

S M T W T F S

Because the first idea might not be the right one

Assumptions and why they matter:

01-12

02-12

03-12

04-12

05-12

06-12

07-12

08-12

09-12

10-12

11-12

12-12

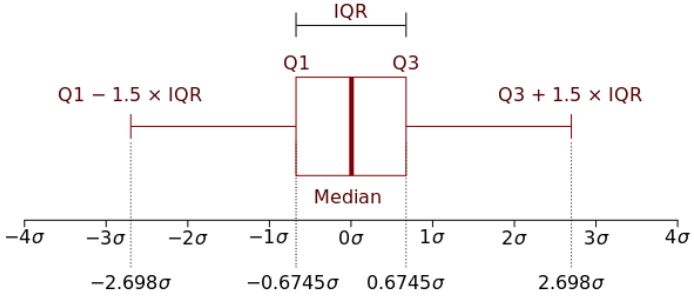
Assumptions and why they matter:

01-December	January-12
02-December	February-12
03-December	March-12
04-December	April-12
05-December	May-12
06-December	June-12
07-December	July-12
08-December	August-12
09-December	September-12
10-December	October-12
11-December	November-12
12-December	December-12

Because the data is different based on location

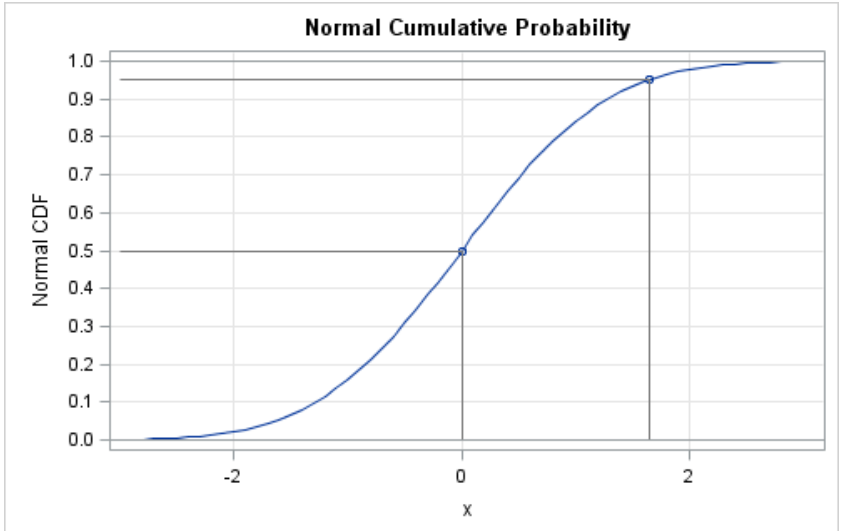
Numerical Exploration

> inner quartile range (Q3 – Q1)



Numerical Exploration

- Quantiles of numerical vectors
  - Quantiles are inverse values of the CDF (cumulative distribution function).
  - Standard Normal: (shown in figure)
    - Quantile(0.5) = 0, means at x=0, 50% of the distribution lies to the left. (This is also the median)
    - Quantile(0.95) = 1.65



Numerical Exploration

- Relationships:
  - covariances

$$Cov(X,Y) = \frac{\sum (X_i - \bar{X}) * (Y_i - \bar{Y})}{n}$$

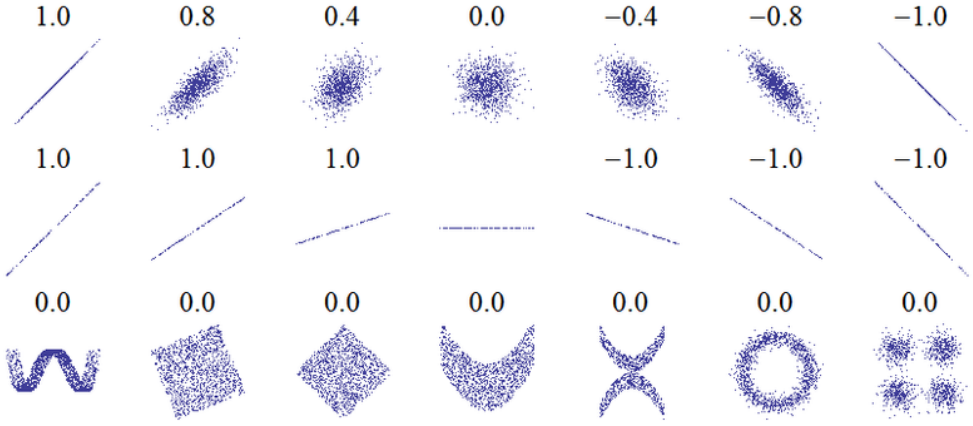
  - Interpretation: Expected value of the differences between x and y and their corresponding mean.
  - E.g. if x is above it's mean when y is also above it's mean, then they will have a high covariance.
  - Highly interpretable, but not bounded.

Numerical Exploration

- Relationships:
  - Correlations (pearsons) = scaled covariance

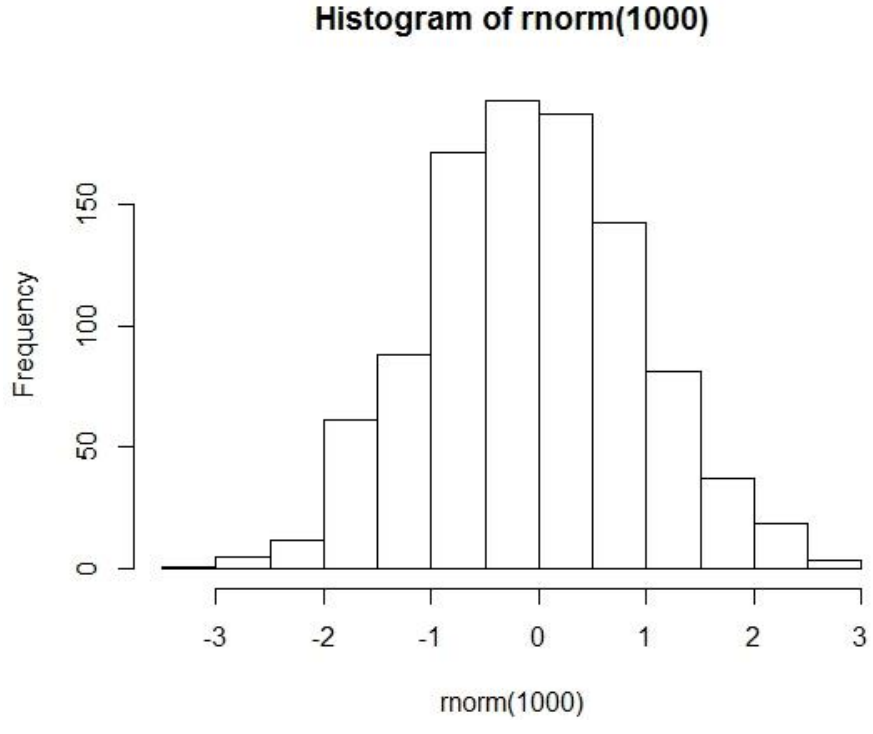
$$r = r_{xy} = \frac{\mathrm{Cov}(x, y)}{S_x \times S_y}$$

  - Bounded between 0 and 1.
  - Not as interpretable.



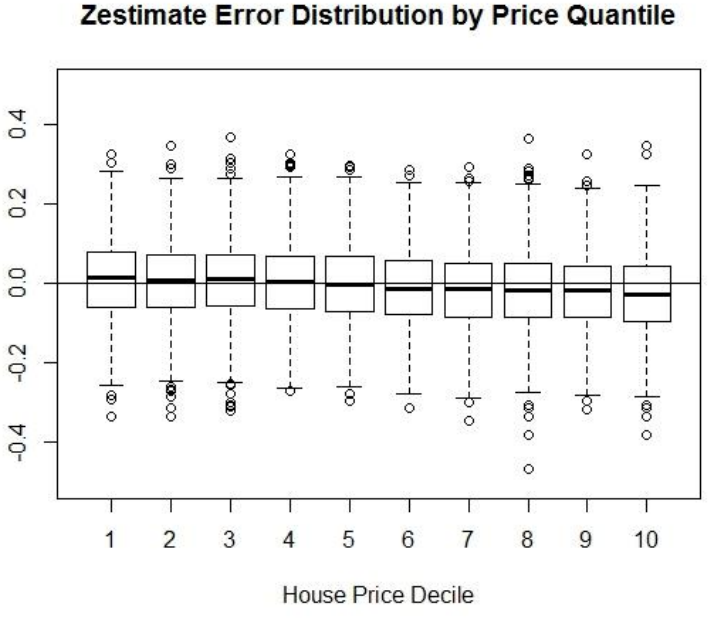
Visual Exploration

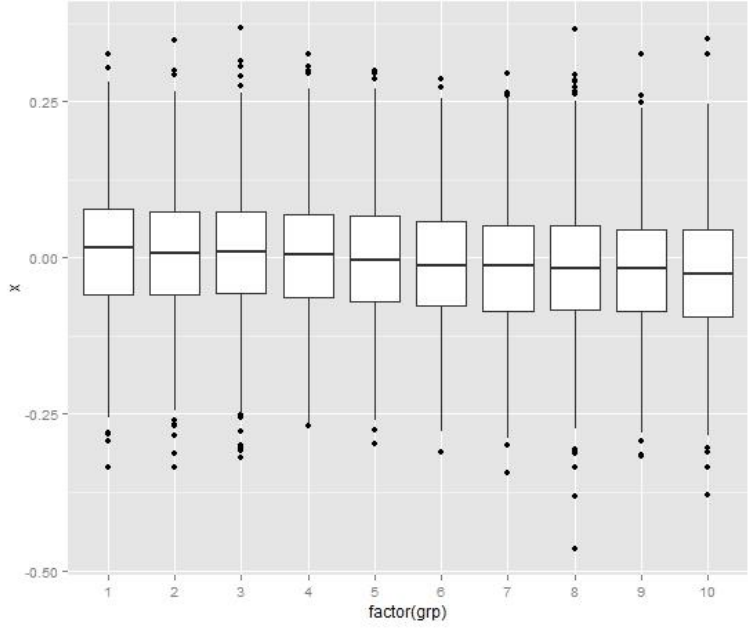
> Histograms:



Visual Exploration

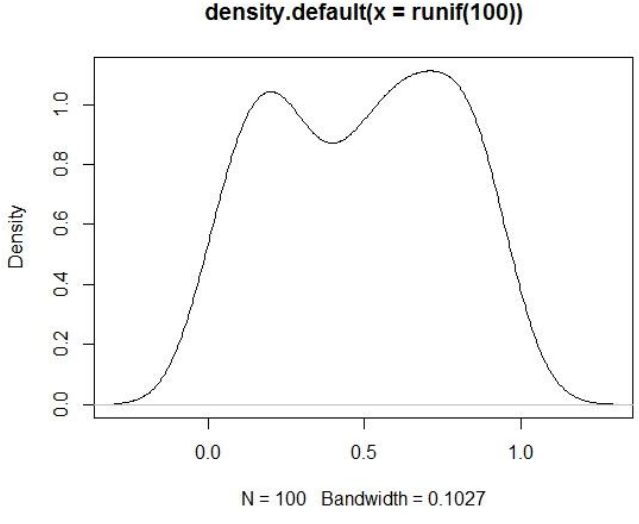
> Boxplots:

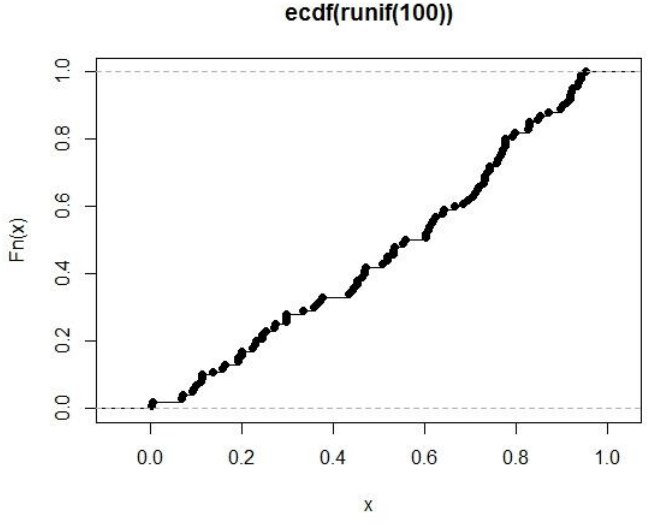




Visual Exploration

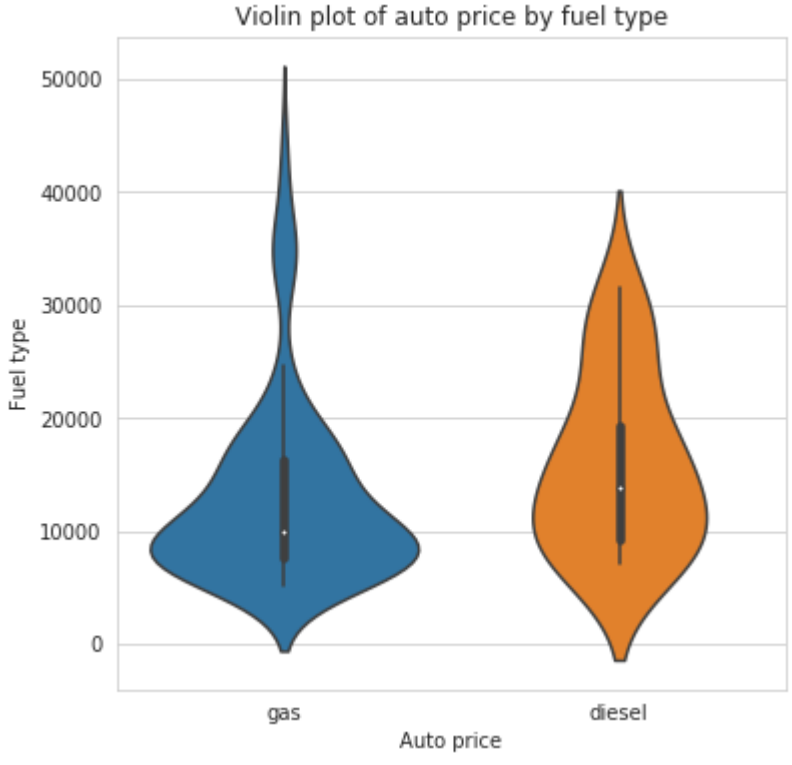
> Densities/CDFs:





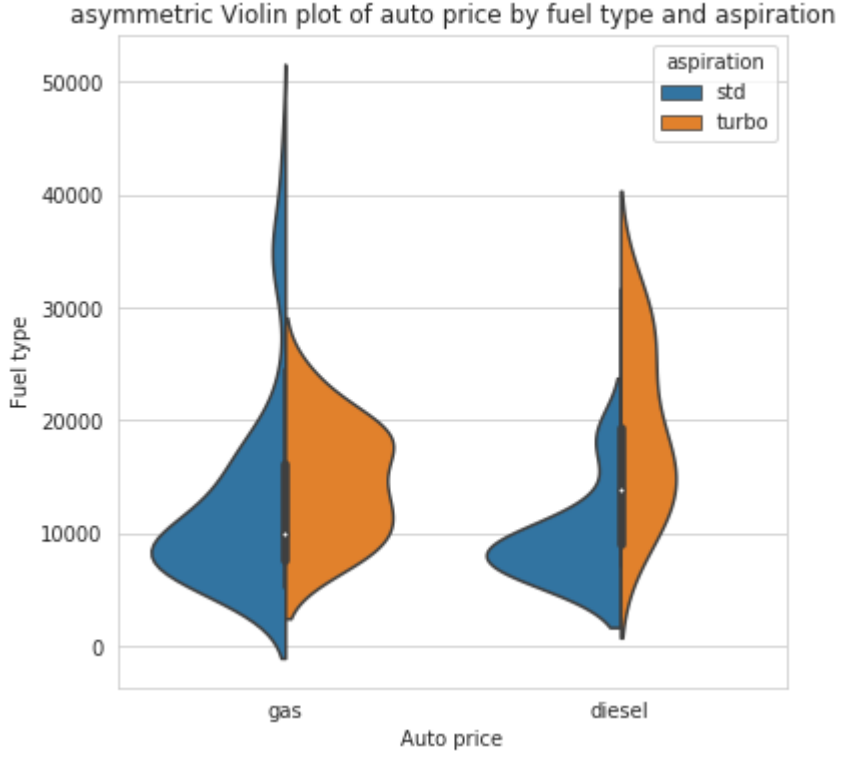
Visual Exploration

> Violin Plots:



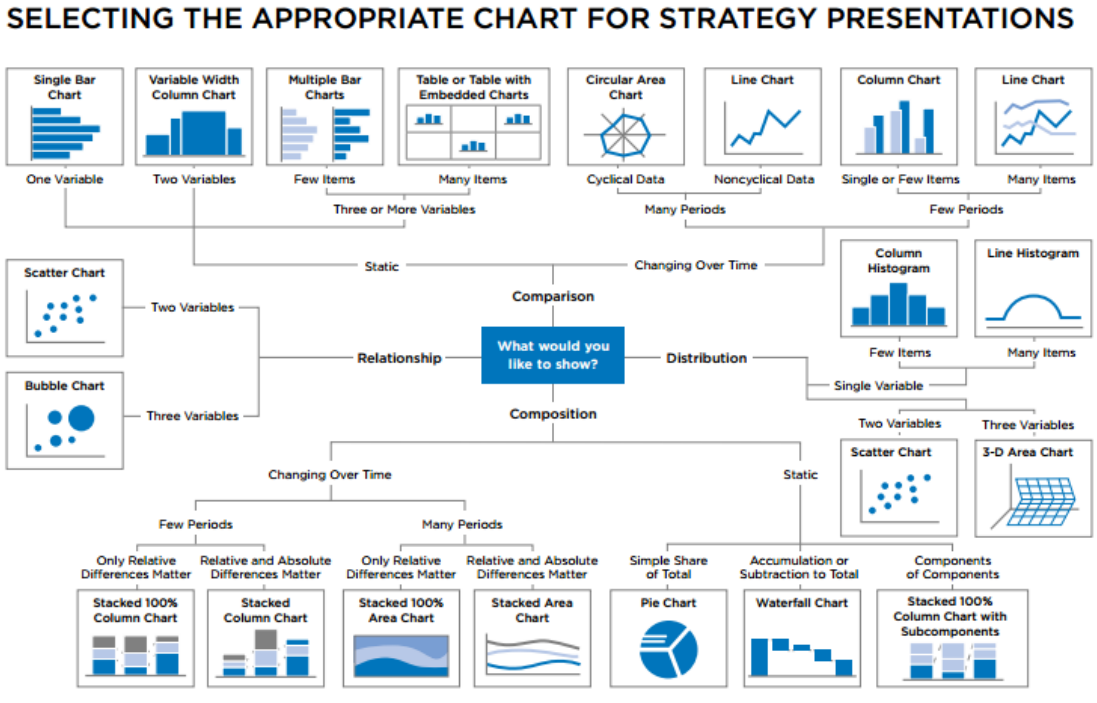
Visual Exploration

> Asymmetric Violin Plots:



Visual Exploration

> Suggested Chart Selection (not to be understood as iron rule)
> The chart conveying the message in the clearest way is the right chart.



(source: multiple, can’t trace the original author)

Data Exploration Part 1

Lesson 1

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON