# Data Mining

# Introduction

- Data Mining is the process of finding not obvious (but useful) information from the data
- This information can be found by means of
  - data visualization
  - finding relations between variables
  - building (prediction, classification) models
  - clustering
  - anomaly detection (outliers)
  - pattern recognition

# How to find useful information?

by finding

- relationships between variables
- trends
- outliers

or by

- splitting the data by categories –stratify-
- creating new variables from existing ones –feature engineering–
- excluding not useful existing variables –feature selection–

# Relationship between variables

- Correlation between y and $x_1$ is r = 0.7

  On average, y increases when x increases

# Relationship between variables

- Correlation between y and $x_1$ is r = 0.7

  On average, y increases when x increases

- Regression line is $\hat{y} = 0.934 + 2.114\,x_1$

  On avg, *y* increases by 2.114 when *x* increases by 1

# Relationship between variables

- Correlation between y and $x_1$ is r = 0.7

  On average, y increases when x increases

- Regression line is $\hat{y} = 0.934 + 2.114\,x_1$

  On avg, *y* increases by 2.114 when *x* increases by 1

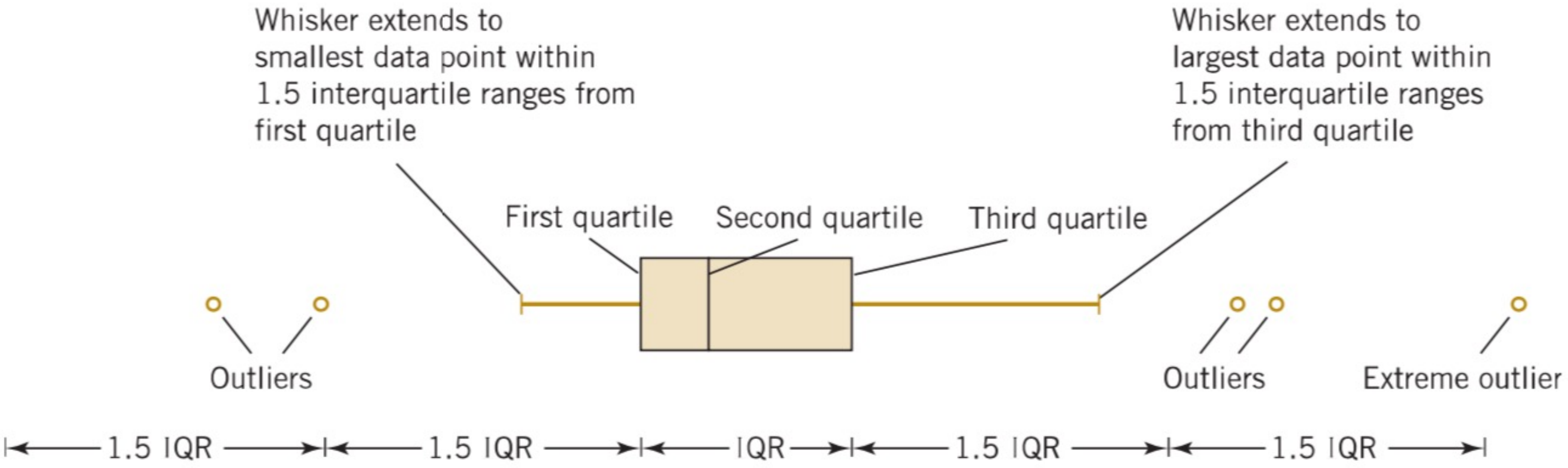
- Add variable $x_2$ and now the regression line is

$$\hat{y} = 0.934 - 0.25\,x_1 + 1.76\,x_2$$

What is the relation between y and x1?

# Box-and-whisker plot (boxplot)

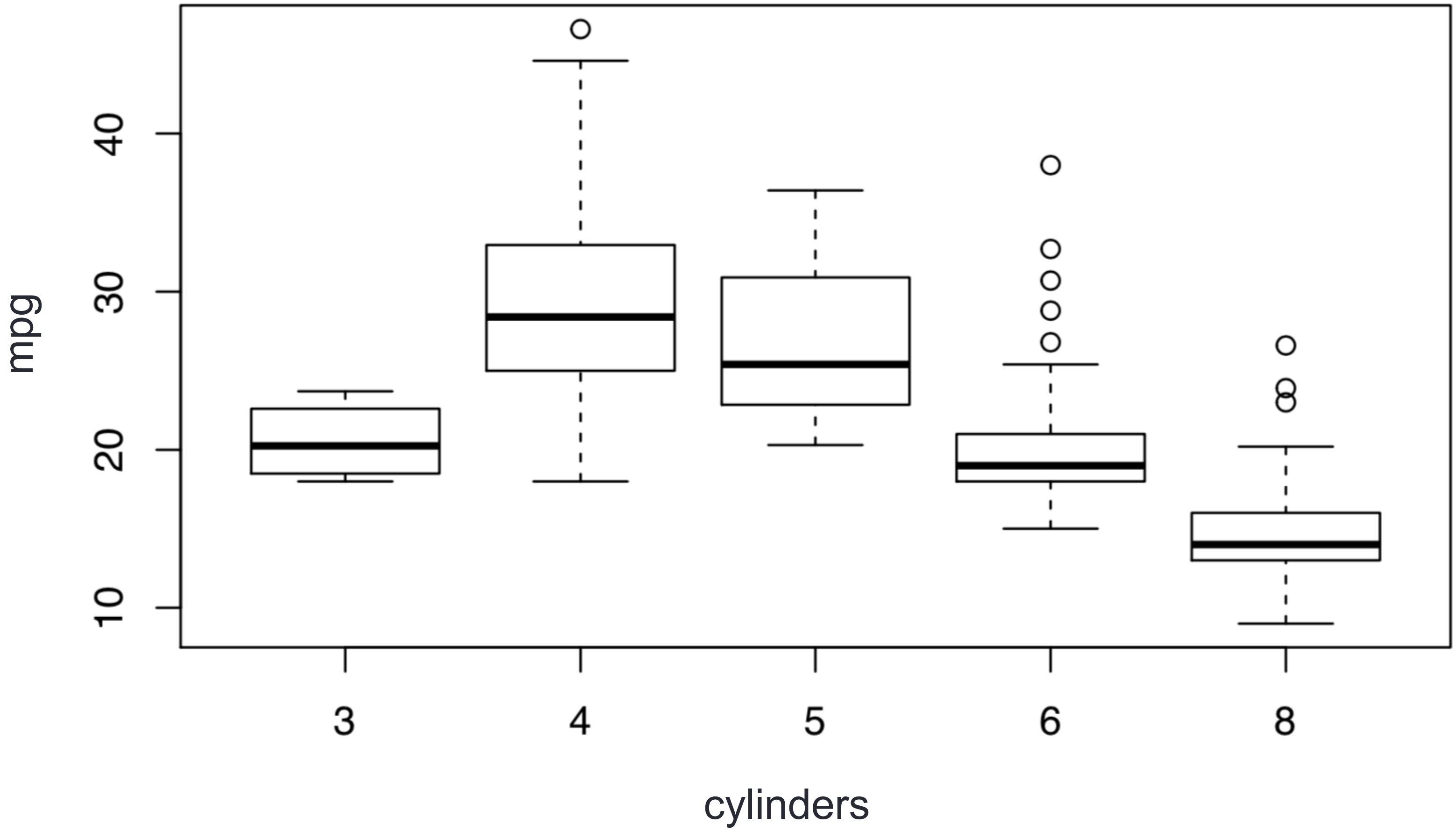
- Graphical display showing key values of a variable distribution
- Key values: Center, spread, symmetry, and outliers
- Useful for comparing same variable on different categories



Whisker extends to smallest data point within 1.5 interquartile ranges from first quartile
Whisker extends to largest data point within 1.5 interquartile ranges from third quartile
First quartile
Second quartile
Third quartile
Outliers
Outliers
Extreme outlier
1.5 IQR
1.5 IQR
IQR
1.5 IQR
1.5 IQR

# Box-and-whisker plot (boxplot)

- Useful for comparing the distribution of a variable on different categories



# Creating new variables

- Combination of variables may be more useful than each individual variable
- For example, in some cases, the difference (of their values) between two predictors may prove more useful than using each one in a model