Inspire…Educate…Transform.

Essentials Engineering Skills for Big Data Analytics

Structured Data Processing and Visualization

June 13, 2015

Data analysis

Data gathering & integration

Per the problem description, what data is available, form of data, attribute types, looking at data summaries, and conduct preliminary analysis. etc. Get an idea of the model. Ensure data availability to meet the objective.

Data cleansing

Identify missing information, outliers or noise etc. Ignoring special characters, etc.

Data transformation

Identify inconsistencies within the data types, conversion of types as required, standardize, discretize, etc.

Data modeling

Identify model(s) and benchmark with simple models. Model output and interpretations.

Patterns to support decision making

Predictions/ Rules/Patterns Data insights/ Visualizations

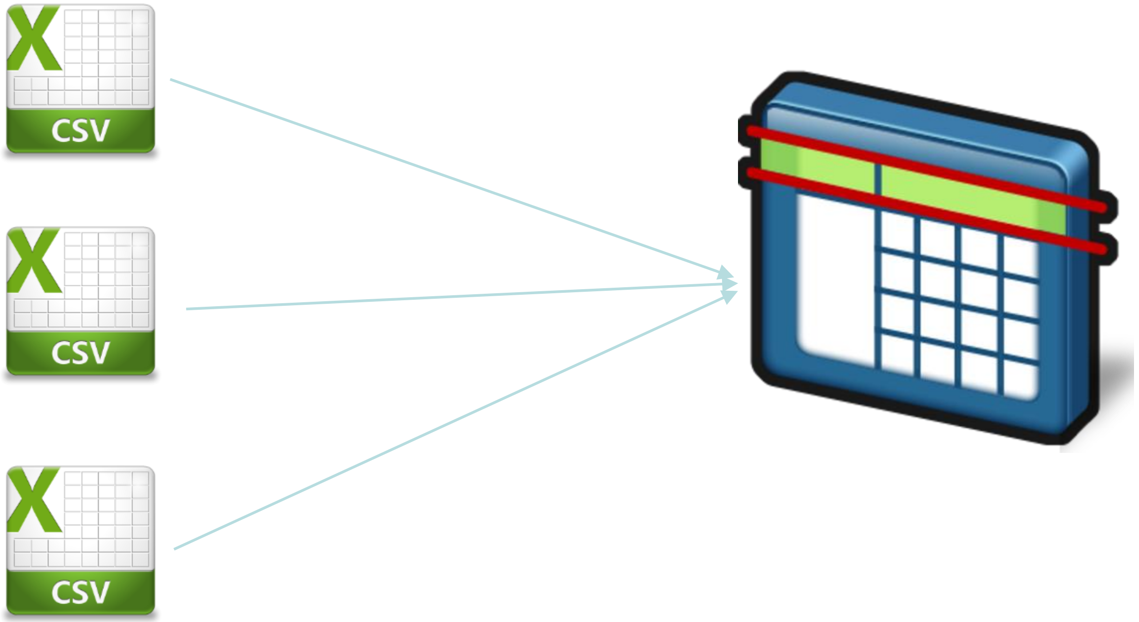
Why pre-process

- Poor model on good data is likely to be better than great model on poor data

- What are the steps in data preprocessing?
  - Data inspection & integration
  - Data cleansing
    - Missing values
    - Outlier detection
  - Data transformation
    - Type conversion
    - Binning
    - Standardization and normalization

Data Integration

- Integrating multiple files, databases, etc. into a single entity. For instance, if data may exist in multiple excel files then we need to merge them in a single file in order to carry out analysis on the entire data.



Attributes types

- Type
  - Numeric, Categorical and Ordinal
  - Date, strings
- Actionable
  - Focus
    - Gender, region, education, etc.
  - Changeable
    - % discount, price, product, etc.

Type conversion

- Some tools can deal with nominal values internally
- While other methods (neural nets, regression, nearest neighbor) require only numeric inputs
- To use nominal fields in such methods need to convert them to a numeric value

Binary to Numeric

- Binary fields

E.g. Gender=M, F

Convert to Field_0_1 with 0, 1 values

e.g. Gender = M → Gender_0_1 = 0

Gender = F → Gender_0_1 = 1

Ordered to numeric

- Conversion: Ordered to Numeric
- Ordered attributes (e.g. Grade) can be converted to numbers preserving Natural order, e.g.

A → 4.0

A- → 3.7

B+ → 3.3

B → 3.0

Why is it important to preserve natural order?

To allow meaningful comparisons, e.g. Grade > 3.5

Categorical to numeric

- How do we set up categorical variables in distance metrics
  - Create as many dummy variables as there are options
  - Code as 100, 010,…

Data cleansing

Data is not always fully available

E.g., many tuples have no recorded value for several attributes, such as customer income in sales data

- Missing data may be due to equipment malfunction
  - E.g., system failure
- data not entered due to misunderstanding
  - E.g., cities/zipcodes
- certain data may not be considered important at the time of entry
  - E.g., age of customers
- no registered history or changes of the data
  - E.g., employee promotion or change in designation

Hence, missing data may need to be inferred.

Data Cleansing: action

- Fill in missing values:
  - Ignore the tuple
  - Fill in the missing values manually: tedious + infeasible?
  - Use a global constant to fill in the missing value: e.g., “unknown”, a new class?!
  - Central imputation
    - Use the attribute mean (or majority nominal value) to fill in the missing value.
  - kNN imputation
    - Imputation using k-nearest neighbors. For each record, identify missing features. For each missing feature find the k nearest neighbors which have that feature. Impute the missing value using the imputation function on the k-length vector of values found from the neighbors.

Central imputation

Stock	Price
Day1	22.4
Day2	20
Day3	19
Day4	
Day5	22.7
Day6	18.5

Suppose the price value for day 4 is missing.

1. Compute the mean/median price for the given data
2. Substitute the value

Average price = (22.4 + 20 + 19 + 22.7 + 18.5)/5
= 20.52

Median price = 20

Hence, the Price for Day 4 = 20.5 or 20

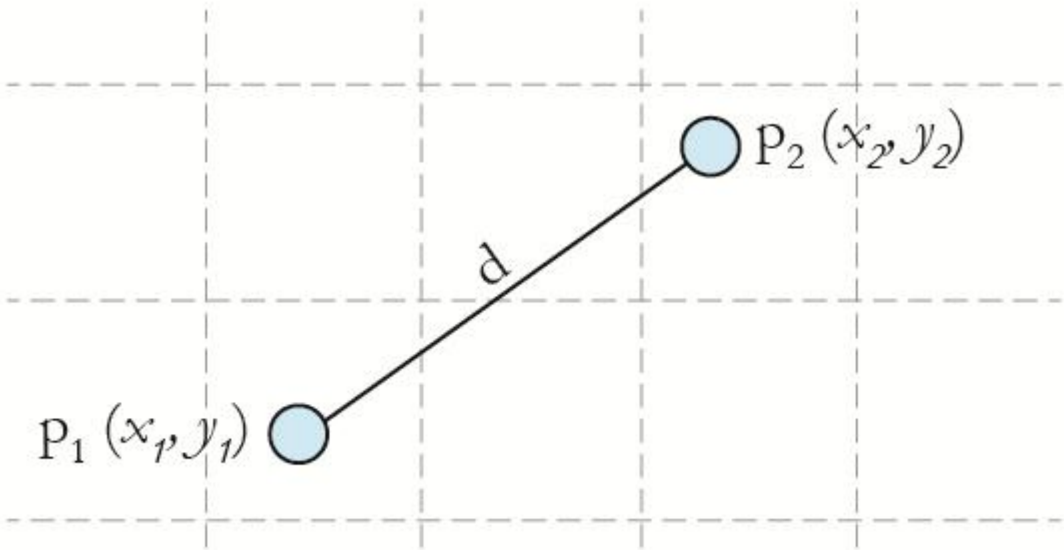
Central imputation

ToyCategory	Price
Infant	46
Infant	20
1 year	19
	45
2-3 years	22.7
4+	18.5

In the case, when the attribute is categorical then substitute with mode

Mode = Infant

Euclidean distance



$$\text{Euclidean distance (d)} = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$$

Compute Euclidean distance

age	exp	inc	family	edu
29	1	49	4	4
45	19	34	3	3
39	15	11	1	1
35	9	100	1	1
35	8	45	4	4
37	13	29	4	4
53	27	72	2	2
50	24	22	1	1
35	10	81	3	3
34	9	180	1	1
65	39	105	4	4
29	5	45	3	3
48	23	114	2	2

The distance between two customers will
Tell us how similar they are behaving.

Let us see how to compute distances

P1= (45, 19, 34, 3, 3)
P2 = (35, 9, 100, 1, 1)

Square root ((45-35)^2 + (19-9)^2+(34-100)^2+(3-1)^2+(3-1)^2

= 67.55

KNN imputation

	age	exp	inc	family	edu
1	29	1	49	4	4
2	45	19	34	3	3
3	39	15	11	1	1
4	35		100	1	1
5	35	8	45	4	4
6	37	13	29	4	4
7	53	27	72	2	2
8	50	24	60	1	1
9	35	10	81	3	3
10	34	9	180	1	1
11	65	39	105	4	4
12	29	5	45	3	3
13	48	23	114	2	2

What is k?

K is the number of neighbors
we want to compare with.

1. Let us assume that K = 1
2. To fill the missing ‘exp’, consider all the other attributes except ‘exp’ .
3. Compute the distances between them.
4. Since K =1 we pick the records that has The least distance and substitute the value corresponding to it.

If K> 1 then we take that many neighbors and compute the average if attribute is numeric or find mode if attribute is categorical.

KNN imputation

	age	exp	inc	family	edu
1	29	1	49	4	4
2	45	19	34	3	3
3	39	15	11	1	1
4	35		100	1	1
5	35	8	45	4	4
6	37	13	29	4	4
7	53	27	72	2	2
8	50	24	90	1	1
9	35	10	81	3	3
10	34	9	180	1	1
11	65	39	105	4	4
12	29	5	45	3	3
13	48	23	114	2	2

For instance, to fill the ‘exp’ value, we will compute the distances between all records after dropping ‘exp’

Sample distance matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	0.675985	0.548	0.724568	0.106191	0.973432	0.412267	0.742238	0.150028	0.672912	0.842089	0.472599	0.810401
2		0	0.708901	0.812123	0.069179	0.170197	0.976874	0.982411	0.686228	0.61787	0.77201	0.903131	0.525195
3			0	0.676776	0.663941	0.650133	0.131188	0.983661	0.902748	0.45684	0.134171	0.278518	0.471798
4				0	0.116646	0.709008	0.294508	0.1718	0.284747	0.12446	0.907106	0.173462	0.254
5					0	0.22803	0.242047	0.518885	0.480488	0.677187	0.16992	0.477329	0.904156
6						0	0.743279	0.148761	0.068481	0.153494	0.94478	0.863299	0.298472
7							0	0.065255	0.614358	0.942889	0.488612	0.142829	0.977342
8								0	0.523811	0.494418	0.087557	0.231459	0.210142
9									0	0.617692	0.29537	0.790905	0.582729
10										0	0.434051	0.282059	0.464114
11											0	0.926031	0.631139
12												0	0.669799
13													0

Suppose, the distance between 4th and 5th record is the least. Then we substitute the ‘exp’ value of 5th record in 2nd In this case the missing value is filled with ‘8’

If K>1

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	0.675985	0.548	0.724568	0.106191	0.973432	0.412267	0.742238	0.150028	0.672912	0.842089	0.472599	0.810401
2		0	0.708901	0.812123	0.069179	0.170197	0.976874	0.982411	0.686228	0.61787	0.77201	0.903131	0.525195
3			0	0.676776	0.663941	0.650133	0.131188	0.983661	0.902748	0.45684	0.134171	0.278518	0.471798
4				0	0.116646	0.709008	0.294508	0.1718	0.128747	0.12446	0.907106	0.173462	0.254
5					0	0.22803	0.242047	0.518885	0.480488	0.677187	0.16992	0.477329	0.904156
6						0	0.743279	0.148761	0.068481	0.153494	0.94478	0.863299	0.298472
7							0	0.065255	0.614358	0.942889	0.488612	0.142829	0.977342
8								0	0.523811	0.494418	0.087557	0.231459	0.210142
9									0	0.617692	0.29537	0.790905	0.582729
10										0	0.434051	0.282059	0.464114
11											0	0.926031	0.631139
12												0	0.669799
13													0

If K = 2 then we consider the 5th and 10th as nearest neighbors and compute the average Of ‘exp’ values corresponding to it. The average value is (15+8)/2 = 11.5
Hence 11.5 will be substituted.

kNN imputation

- Well, the attributes have to be all numeric type.
- For categorical attributes, it is best to create dummies.
  - If there are say 5 levels in the categorical attribute, then 5 dummy variables will be created and converted into binary levels.

Outliers

- Definition of Hawkins [Hawkins 1980]: "An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism"

- The outlier may be due to the differences in the measurement, change in the system behavior, human error, transcription error, etc.

Outlier identification

- General ways of identifying are using scatter plots and box plots.
- Using Box plot: an observation that is outside the interquartile range

Let us recollect how to find the quartiles.

A box plot is constructed by drawing a box between the upper and lower quartiles with a solid line drawn across the box to locate the median. The following quantities (called fences) are needed for identifying extreme values in the tails of the distribution:

lower inner fence: Q1 - 1.5*IQ
upper inner fence: Q3 + 1.5*IQ
lower outer fence: Q1 - 3*IQ
upper outer fence: Q3 + 3*IQ

Outlier detection criteria: A point beyond an inner fence on either side is considered a mild outlier. A point beyond an outer fence is considered an extreme outlier.

Reference: http://www.r-tutor.com/r-introduction

Outlier identification

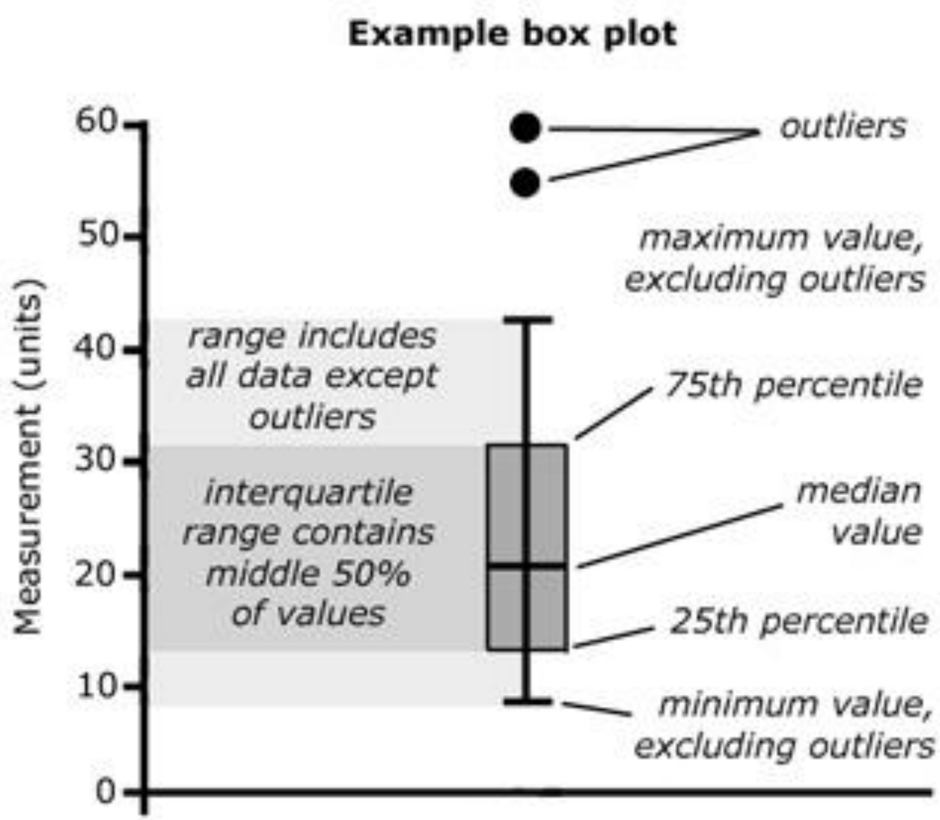


Image source: http://web.anglia.ac.uk/numbers/graphsCharts.html

Noise

Noise: random error or variance in a measured variable

- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention

Handle Noise

Bin data

- first sort data and partition into equal bins
- then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

Example

- Let us say we have the data for price as 4, 8,9,15,21,21,24,25,26, 28, 29, 34

- If we partition in to equal sizes

Bin1: 4, 8, 9, 15

Bin2: 21, 21, 24, 25

Bin3: 26,28,29,34

Smoothing noise

- Use mean/median to smooth

Bin1: 9, 9, 9, 9

Bin2: 23, 23,23, 23

Bin3: 29,29,29,29

- Use bin boundaries to smooth

Bin1: 4, 4, 4, 15

Bin2: 21, 21, 21, 25

Bin3: 26, 26, 26, 34

Outliers and Noise

Distinguish between noise and outliers.

(a) Is noise ever interesting or desirable? Outliers?

No, by definition. Yes.

e.g.: Fraud identification

(b) Are noise objects always outliers?

No. Random distortion can result in an object or value much like a normal one.

(c) Are outliers always noise objects?

No. Often outliers merely represent a class of objects that are different from normal objects.

Quick recap

- Understand why is preprocessing important.
- Understand what steps are carried out in in order to prepare the data
- Be able to read and merge data from multiple files in R
- Create subsets of data, convert data types, create dummies in R
- Understand data cleansing and apply central imputation and kNN imputation methods to fill missing values
- Understand outliers and noise in data and correct them

Standardization

- Standardization is a statistical technique that gives different units of measurement a common base for purposes of comparison

Why standardize?

- When approaching data for modeling, some of the standard procedures should be used to prepare the data.
- We learnt that the data needs to be cleansed, by removing special characters, removing outliers and noise, fill the missing value, etc.
- The next step is to identify and equalize the range and/or data variability
  - Variables measured at different scales do not contribute equally to the analysis.
  - Transforming the data to comparable scales can prevent this problem.

Example

- For e.g., the variable Age that ranges between 0 – 100 outweighs the Income that ranges between 10,000 – 50,000.

Age	Income (£)
24	15000
30	12000
28	30000

Income dominates completely!

Standardization techniques

- 0-1 scaling
- Z-score scaling
- Range scaling
- Stdev scaling

Standardization techniques

- **0-1 scaling**: each variable in the data set is recalculated as

$$(V - \min V)/(\max V - \min V)$$

where V represents the value of the variable in the original data set.

This method allows variables to have differing means and standard deviations but equal ranges.

In this case, there is at least one observed value at the 0 and 1 endpoints.

Example

Age	Income (£)	New value
24	15000	(15000 – 12000)/18000 = 0.16667
30	12000	(12000 – 12000)/18000 =0
28	30000	(30000 – 12000)/18000 =1

Income_Minimum = 12000
Income_Maximum = 30000
(Max – min) = (30000 – 12000) = 18000

Please note, the new values have
Minimum = 0
Maximum = 1

Hence, we have converted the income values between 0 and 1.

Standardization techniques

- Dividing each value by the **range**:

recalculates each variable as V /(max V - min V).

In this case, the means, variances, and ranges of the variables are still different, but at least the ranges are likely to be more similar.

Example

Age	Income (£)	New value
24	15000	(15000)/18000 = 0.8333
30	12000	(12000)/18000 =0.6666
28	30000	(30000)/18000 =1.6666

Income_Minimum = 12000
Income_Maximum = 30000
Income_Range = 30000 – 12000 = 18000

Hence, we have converted the income to lower values using the Range method.

Standardization techniques

- **Z-score scaling**:

variables recalculated as (*V* - mean of *V*)/s,

where "s" is the standard deviation. As a result, all variables in the data set have equal means (0) and standard deviations (1) but different ranges.

Example

Age	Income (£)	New value
24	15000	(15000 - 19000)/9643.65 = -0.4147
30	12000	(12000 - 19000)/9643.65 = -0.7258
28	30000	(30000-19000)/9643.65 = 1.1406

Average = (15000 + 12000 + 30000)/3 = 19000
Standard deviation = 9643.65

Hence, we have converted the income values to lower values using the z-score method.

x = c(-0.4147, -0.7258, 1.1406)
mean(x) = -0.000003 ~ 0
var(x) = 0.999 ~1

Standardization techniques

- **Sd scaling**:

variables recalculated as *V* /s,

where "s" is the standard deviation. Dividing each value by the **standard deviation**. This method produces a set of transformed variables with variances of 1, but different means and ranges

Example

Age	Income (£)	New value
24	15000	15000 /9643.65 = 1.5554
30	12000	12000/9643.65 = 1.2443
28	30000	30000/9643.65 = 3.1108

x = c(1.5554, 1.2443, 3.1108)
mean(x) = 1.970167

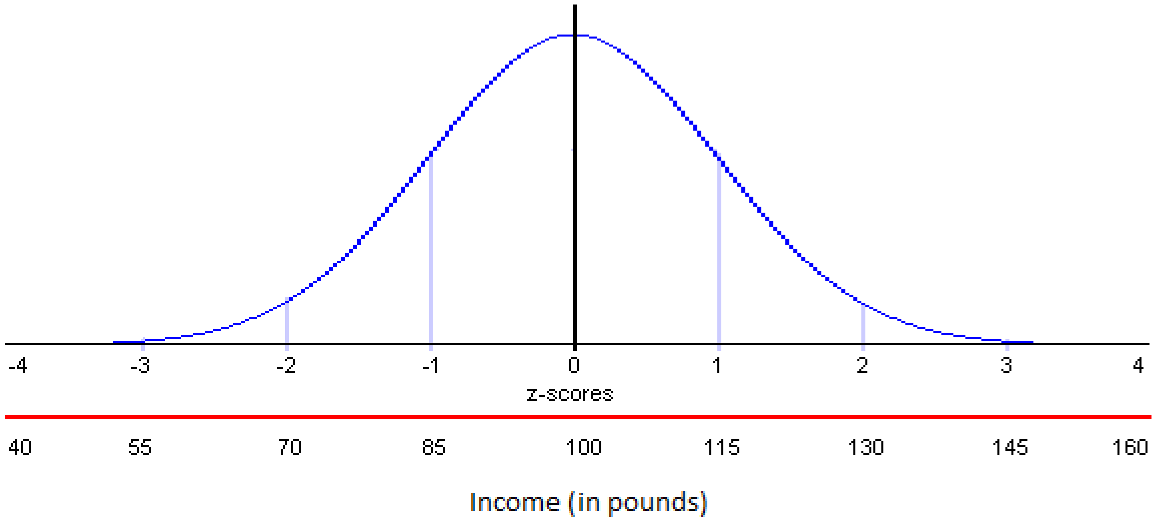
variance (x) = 0.999 ~ 1

Standardization

- Min-max is extremely sensitive to outliers

  Min-max (1,2, 1001) is (0, 0.001, 1)

- The z-score is what will be calculated to standardize the data and it reflects, how many standard deviations is it from the average that the data point falls.



Discretization/Binning

- Converting numeric to categorical
  - Manual
  - Equal width
    - Interval is same
  - Equal frequency
    - Number of sample in each bin is same

Binning

Manual bin

If age is less than 40 then 1,

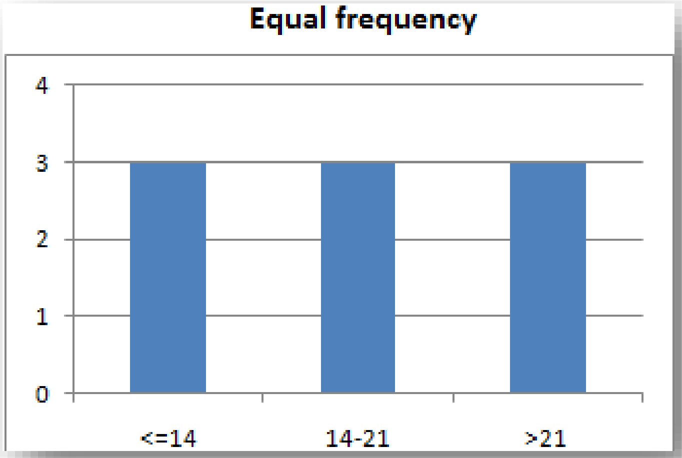
If between 40 to 60 then 2,

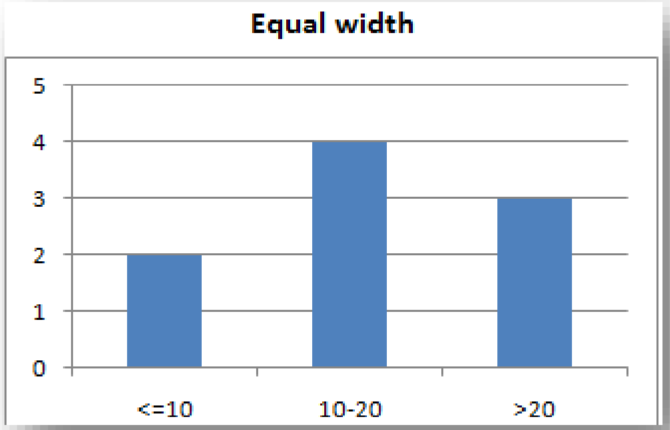
If more than 60 then 3

Binning

- **equal frequency**: when the data is divided into *k* groups where each group has approximately the same value. For both equal frequency and equal width, the best way to determine *k* is to look at the histogram and experiment with different intervals or groups

- **equal width**: when the data is divided into *k* intervals of equal size. The width of intervals is $w = (max–min) / k$. The interval boundaries are $min + w, min + 2w, \ldots , min + (k – 1)w$.





Recap!

- What are the steps in data preprocessing?
  - Data inspection & integration
  - Data cleansing
    - Missing values
    - Outlier detection
  - Data transformation
    - Type conversion
    - Binning
    - Standardization and normalization

Data visualizations

- The primary goal of data visualization is to communicate the information clearly and effectively into one collective, illustrative graphic.

Data visualizations

Few other definitions:

- According to Friedman (2008) the "main goal of data visualization is to communicate information clearly and effectively through graphical means. It doesn't mean that data visualization needs to look boring to be functional or extremely sophisticated to look beautiful. To convey ideas effectively, both aesthetic form and functionality need to go hand in hand, providing insights into a rather sparse and complex data set by communicating its key-aspects in a more intuitive way. Yet designers often fail to achieve a balance between form and function, creating gorgeous data visualizations which fail to serve their main purpose — to communicate information".

- Fernanda Viegas and Martin M. Wattenberg have suggested that an ideal visualization should not only communicate clearly, but stimulate viewer engagement and attention.

Emerging importance of visualization

- show the data
- induce the viewer to think about the important elements in data rather than about methodology
- present many numbers in a small space
- make large data sets coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from a broad overview to the fine structure
- display patterns or relationships in data
- be closely integrated with the statistical and verbal descriptions of a data set.

Graphics *reveal* data. Indeed graphics can be more precise and revealing than conventional statistical computations.

Tufte, Edward (1983). *The Visual Display of Quantitative Information.* Cheshire, Connecticut: Graphics Press. ISBN 0961392142.

Designing useful graphs

- Tell one very important story per graph and identify elements that tell the story

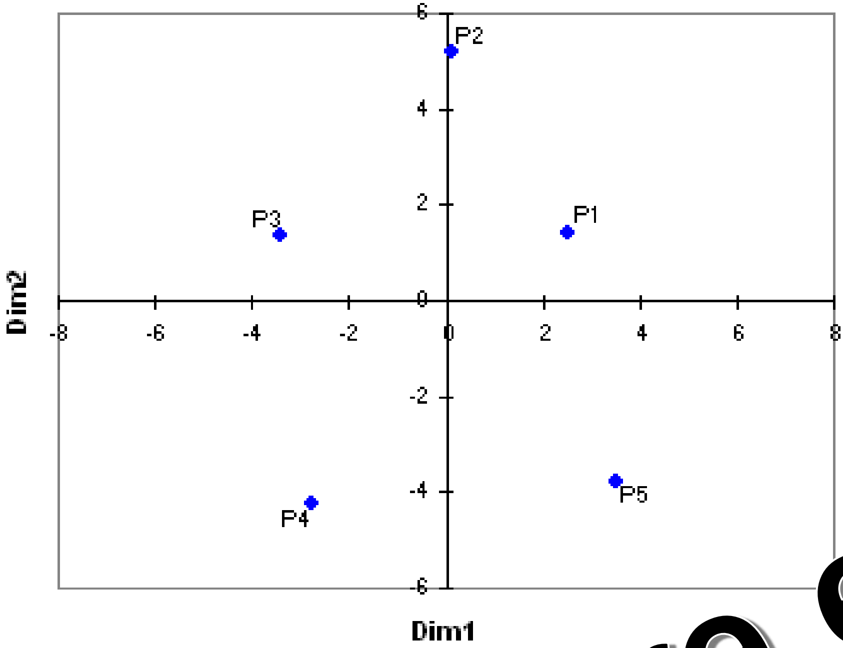
- The most important elements that tell the story should take the X and Y axis

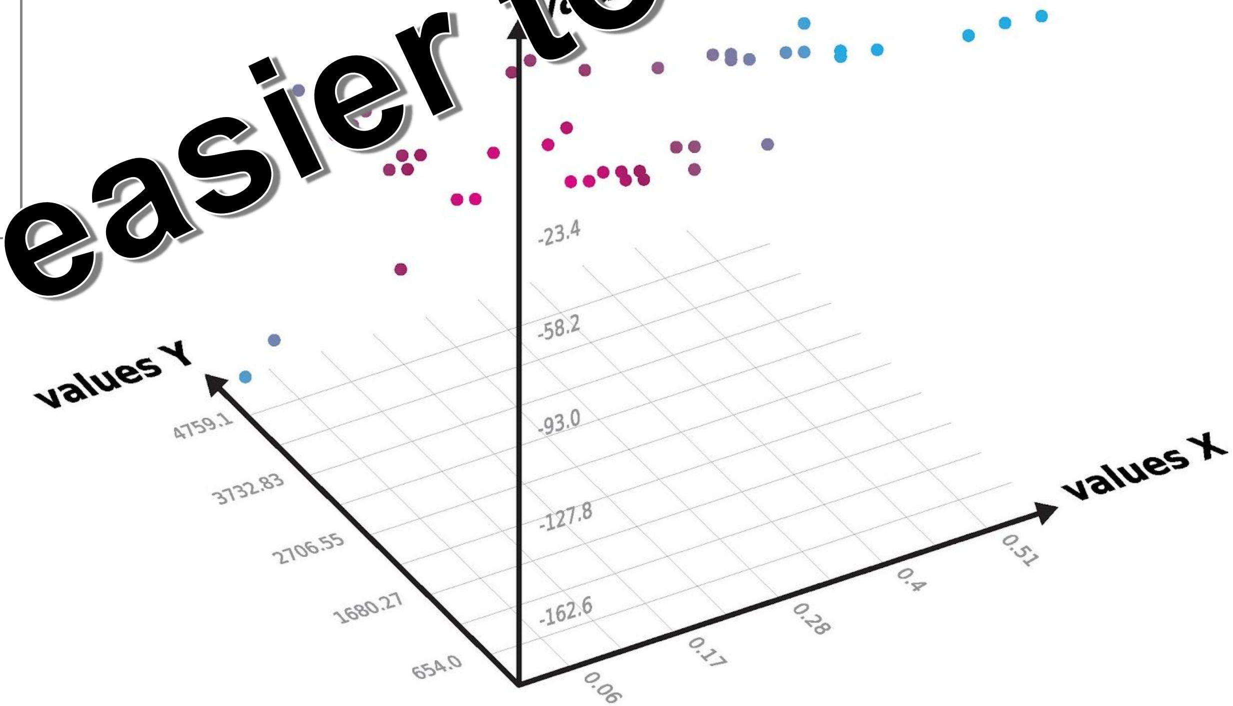
  – Help me choose a flight

Airline

- Important attributes
  - Price, time of travel, start and end time, number of stops, name of airlines, etc.
  - Hipmunk

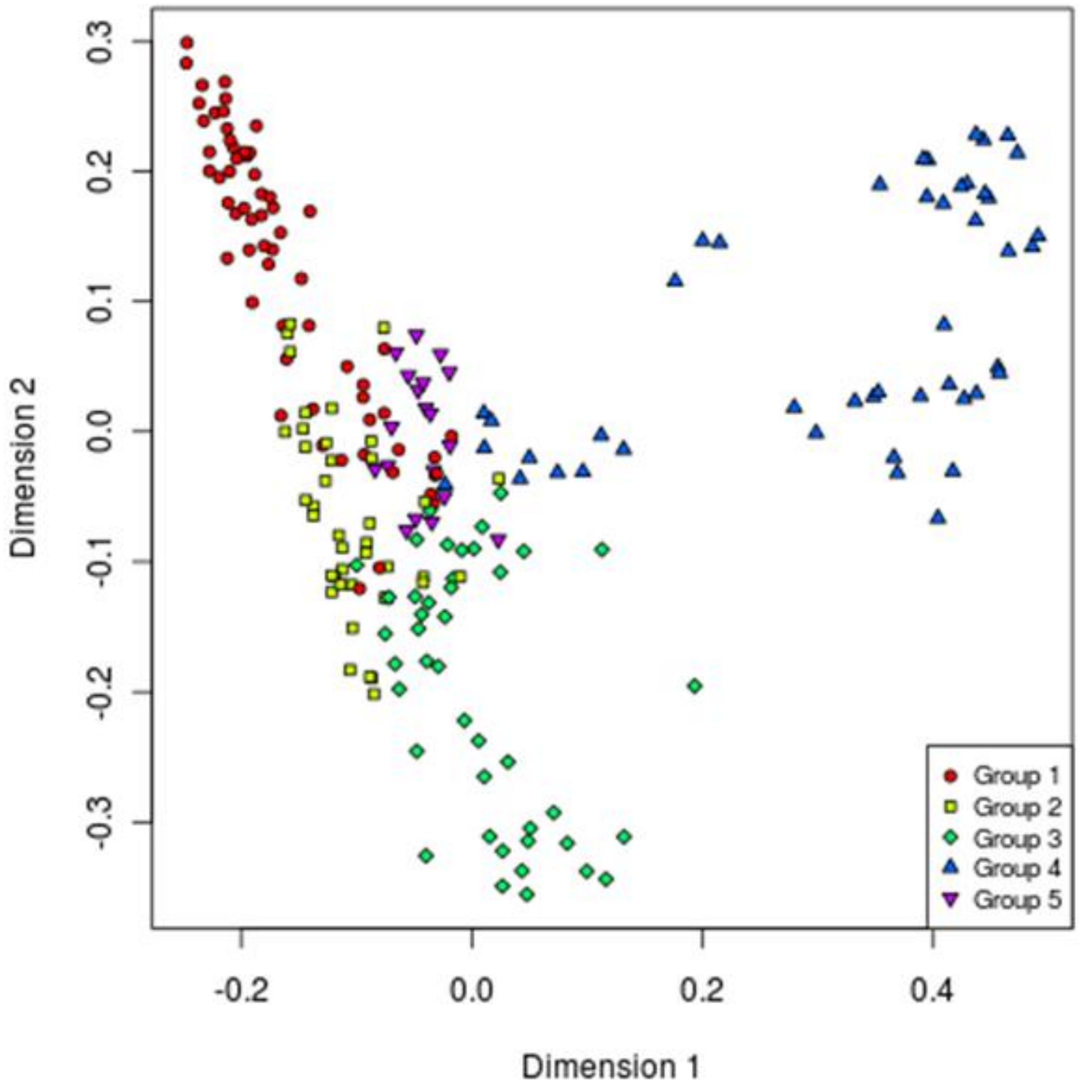
What if there are more elements?



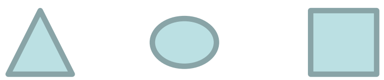


2 D are easier to read

Use color, shape for additional features



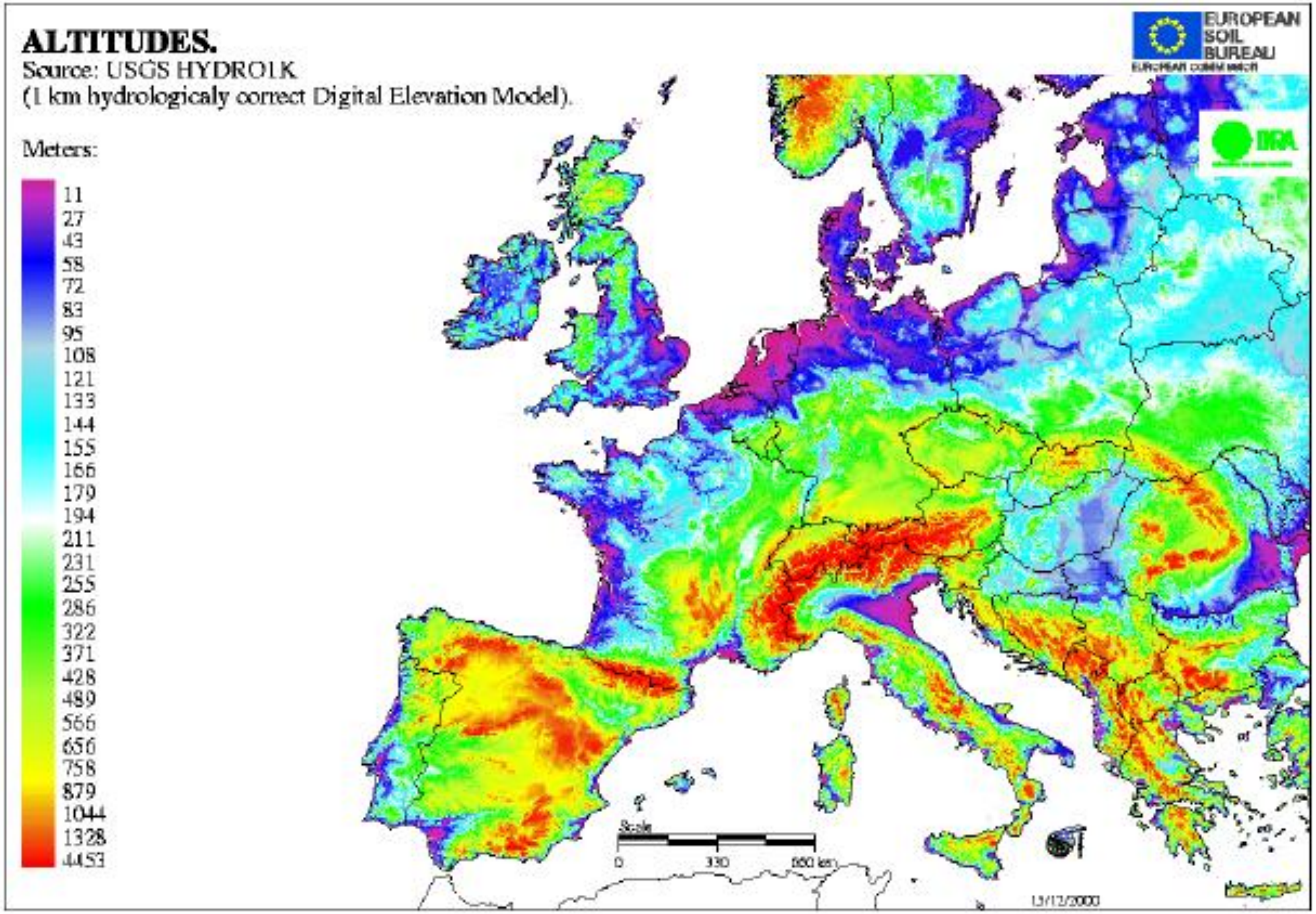
Super powerful maps

- Size ()
- Color ()
- Shape ()

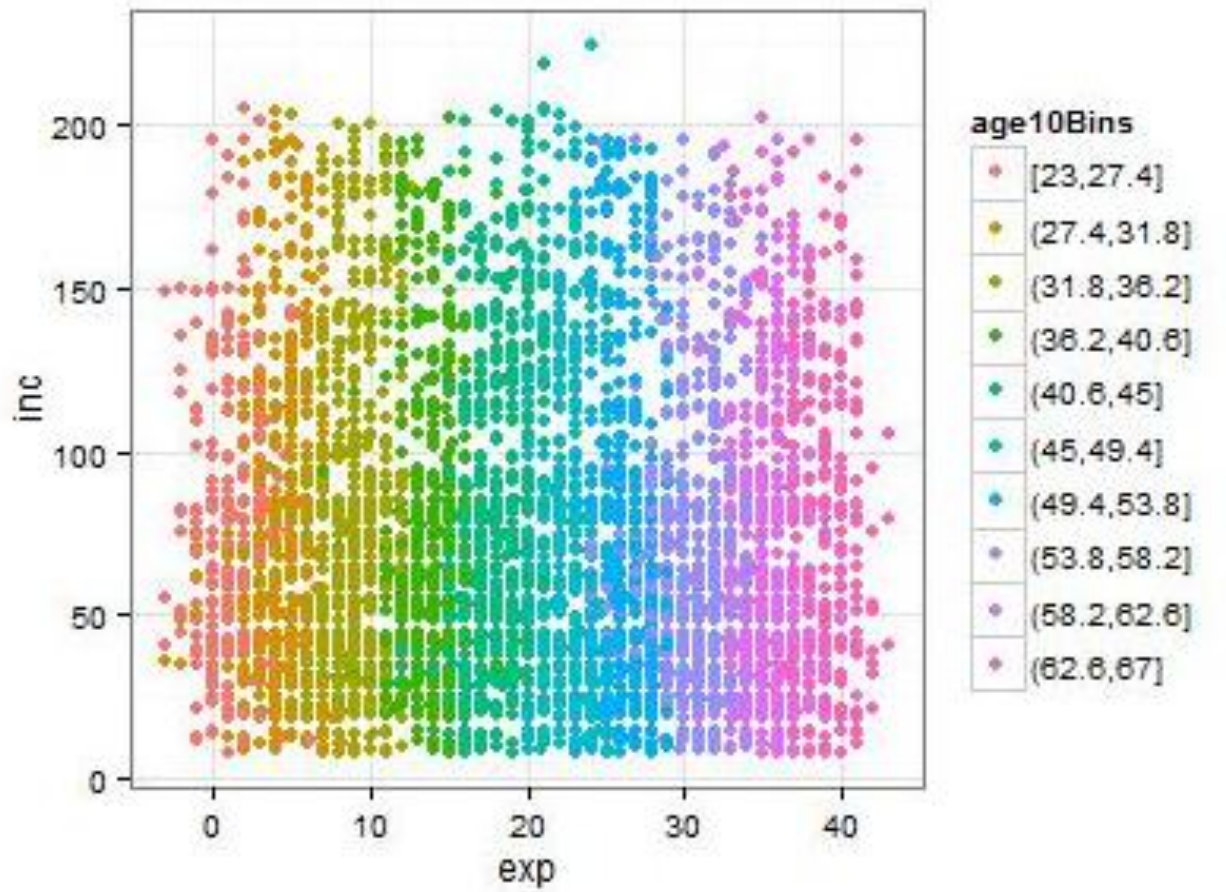
Attributes

- Numeric: Numbers (age, height, income etc.)
- Categorical: Gender, country and all the above when expressed as bins

Shape and color are bad for numeric



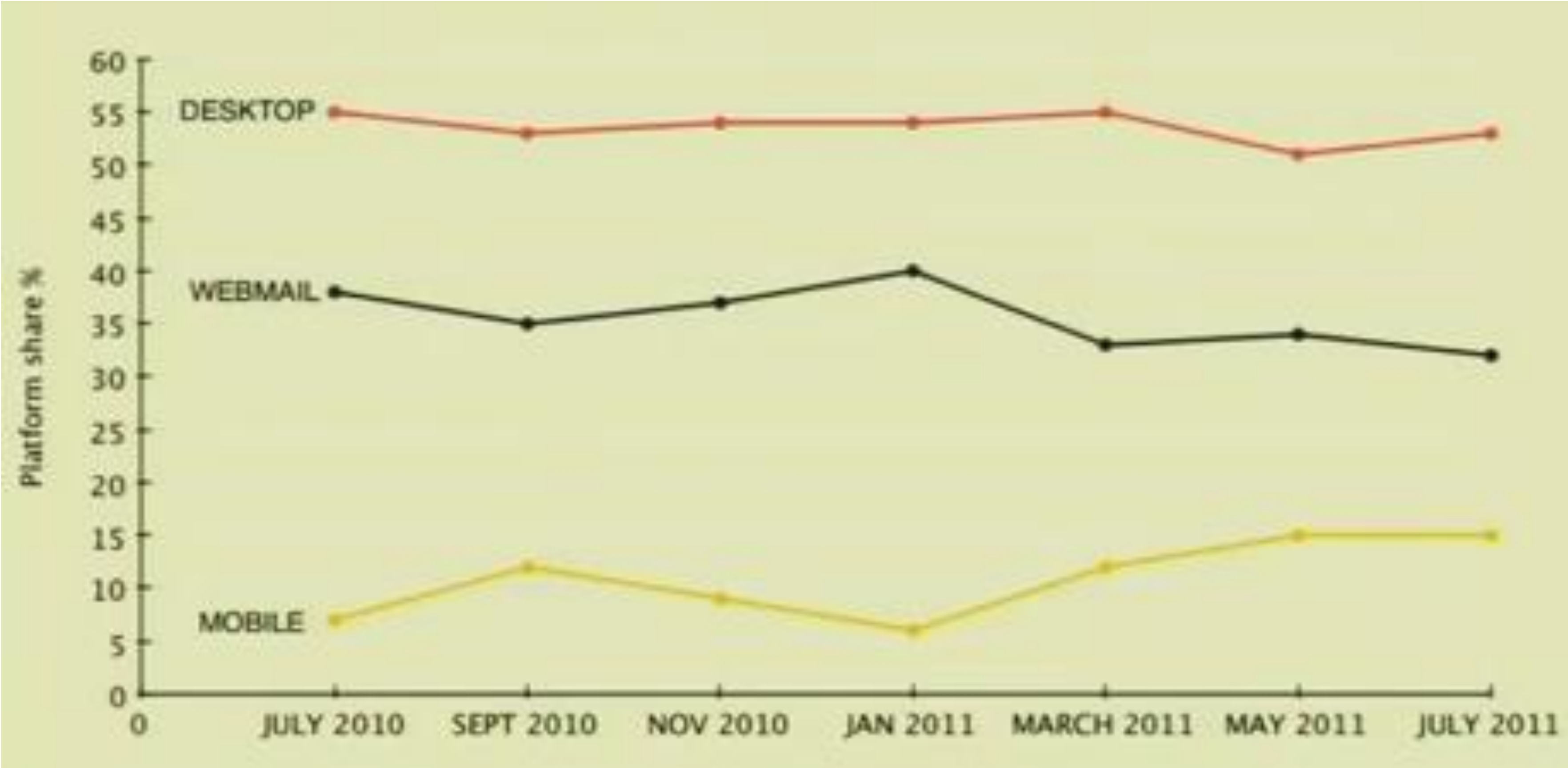
Shape and color are bad for categorical variables with too many levels



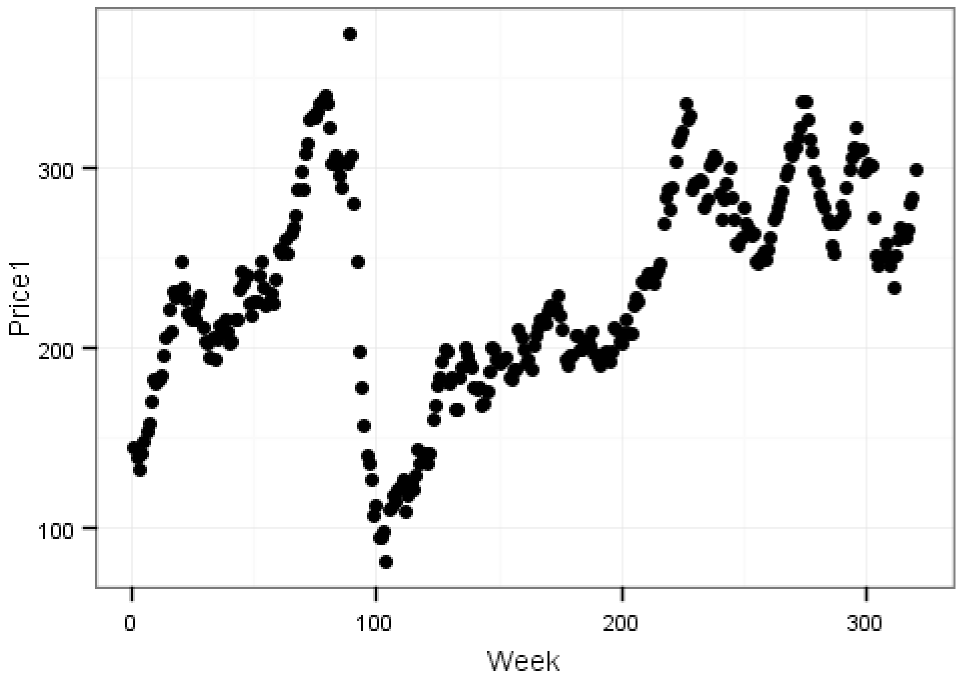
Best graphs

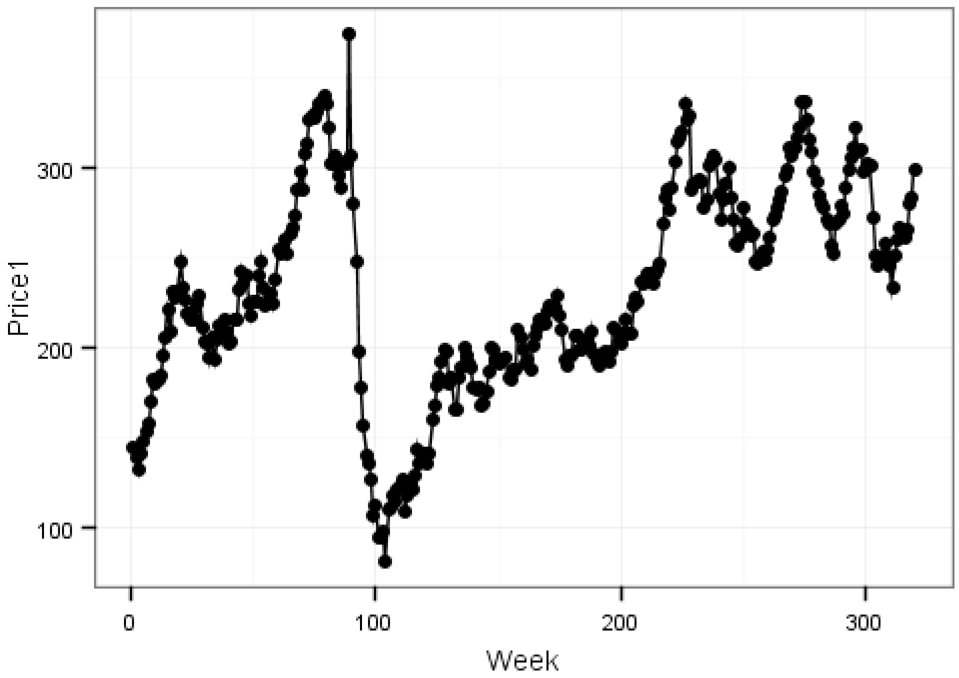
- Most important attributes get X and Y axis
- Next most important must be mapped to
  - Size if numeric
  - Color or shape if categorical

Lengths are better



When data has a time dependence



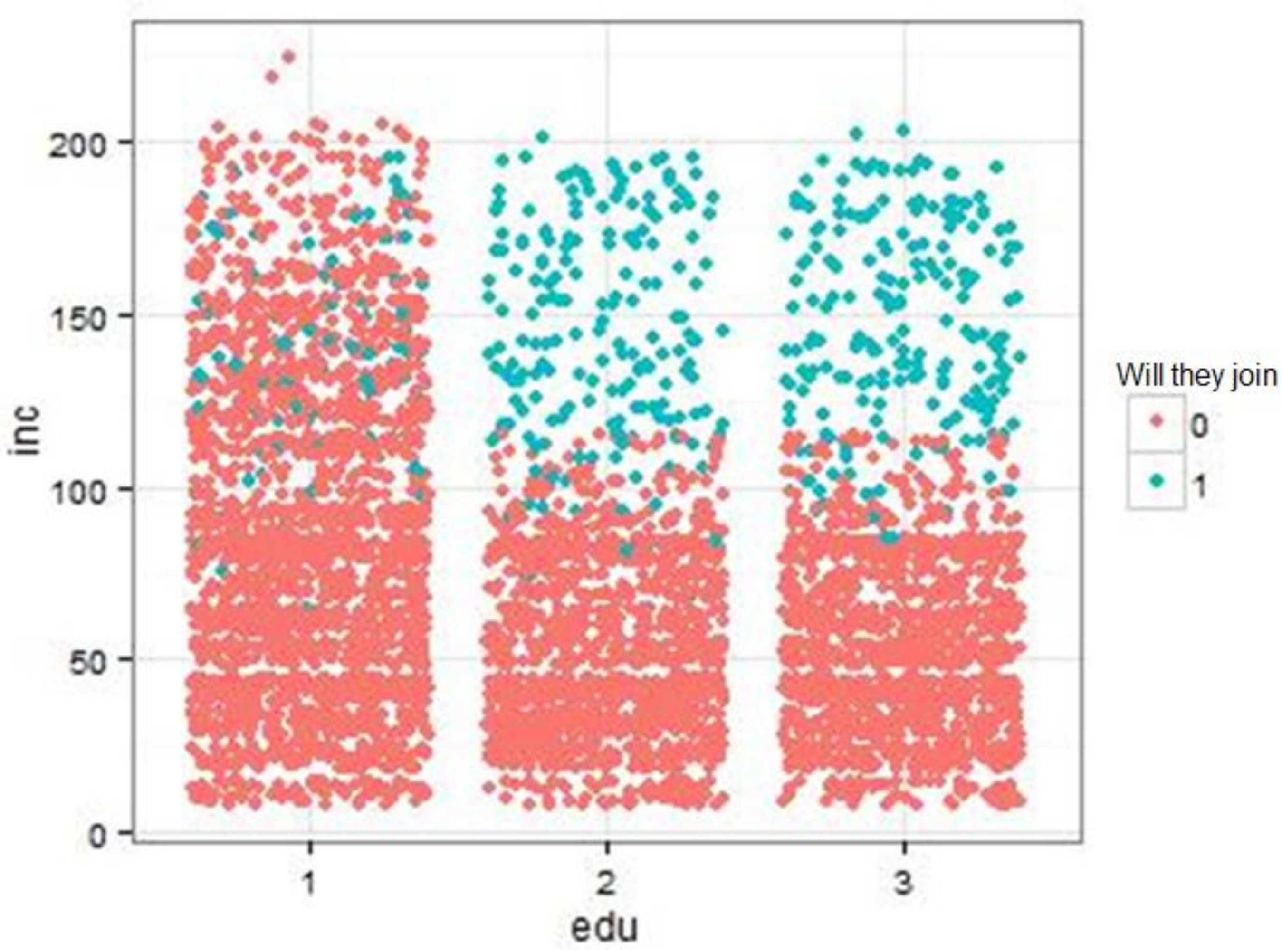


Multiple attributes

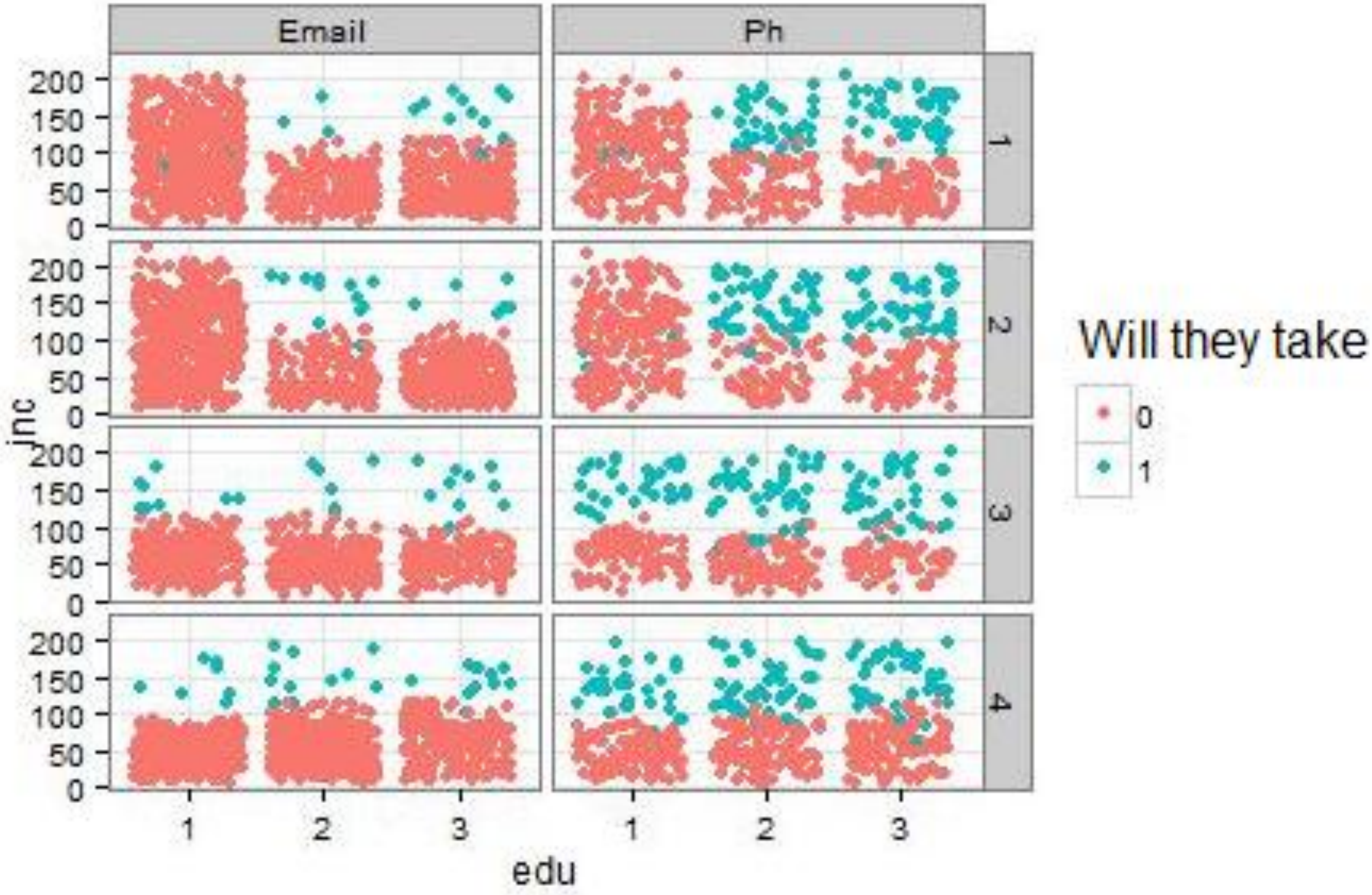
- Do education and income level have a bearing on whether they join the course?

- Let us also analyze the impact of family size and contact mode

Multiple attributes: Scatter



Multiple dimensions



Numeric/Categorical	0	1 categorical	2 categorical	3 categorical
0		Bar	Grouped Bar	
1 Numeric	Histogram/Box or Line	Box plot for all levels/Histogram with color mapping	Point graph with color mapping	Faceting
2 Numeric	Scatter	Scatter with color/shape mapping	Point graph with color and size	
3 Numeric	Scatter and map size to third one	Scatter with size, color mapping	Point graph with color, size and shape	

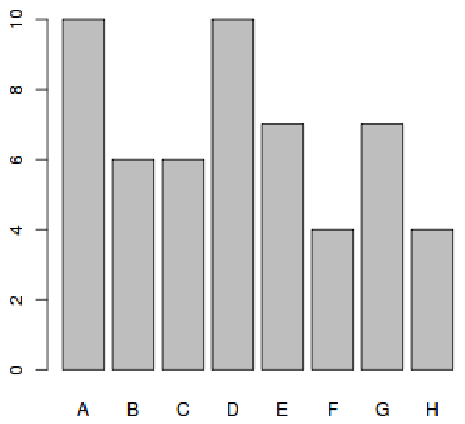
The key takeaway is that, no matter what they look like, data visualizations need to be factually accurate and help viewers make conclusions based on the data.

Let us now look at the Bar charts

Bar Graph

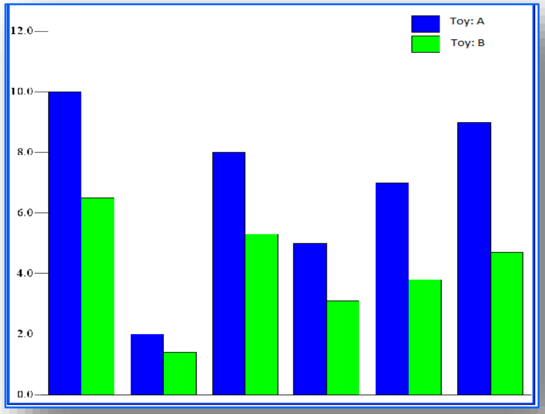
- A bar graph of a qualitative data sample consists of vertical parallel bars that shows the frequency distribution graphically

Understand the sales of toys



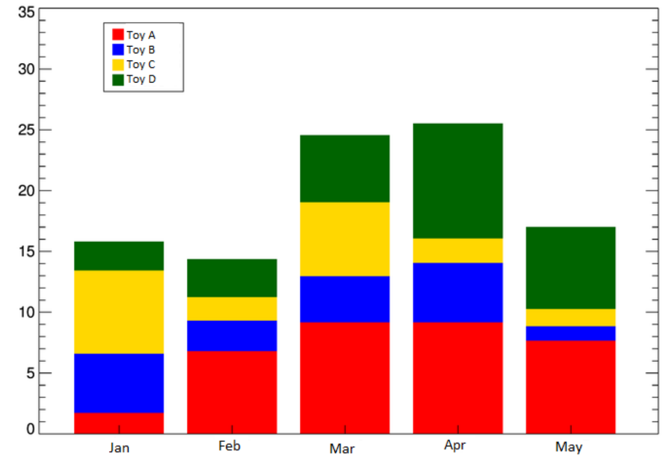
X axis: type of toys,
{A, B, C, D, E, F, G, H}

Y axis: count of items sold



X axis: Month 1 through 6

Y axis: count of items sold



X axis: Jan through May

Y axis: count of items sold

Histogram

A histogram consists of parallel vertical bars that graphically shows the frequency distribution of a quantitative variable. The area of each bar is equal to the frequency of items found in each class.

Box plots

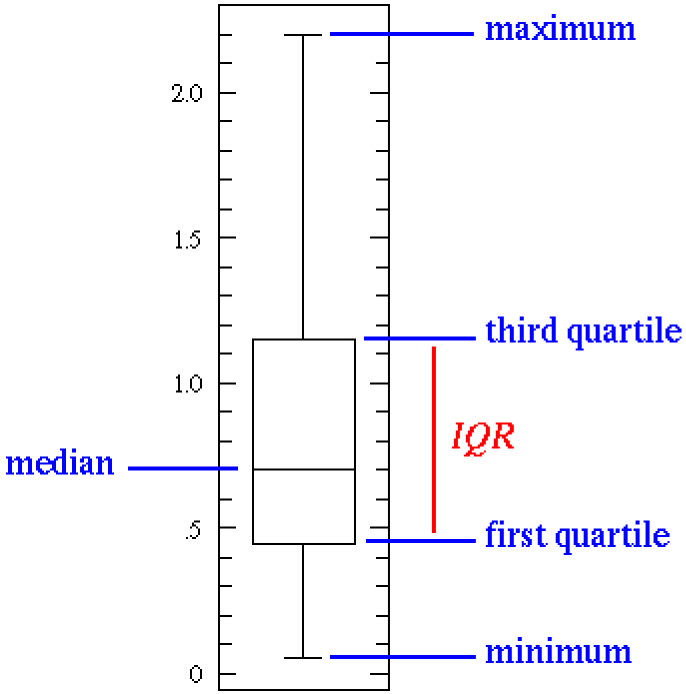
- Also called as box-and-whiskers plot
- In descriptive statistics, boxplot is a graphical way of describing the quantiles in the data.
- If you recall, we have seen how box plots can be used to identify outliers.

Box plots

- Outlier detection aims to find patterns in data that do not conform to expected behavior.

- Box plot is the standardized way of displaying the distribution of data based on the five number summary: minimum, first quartile, median, third quartile and maximum.

Box plot

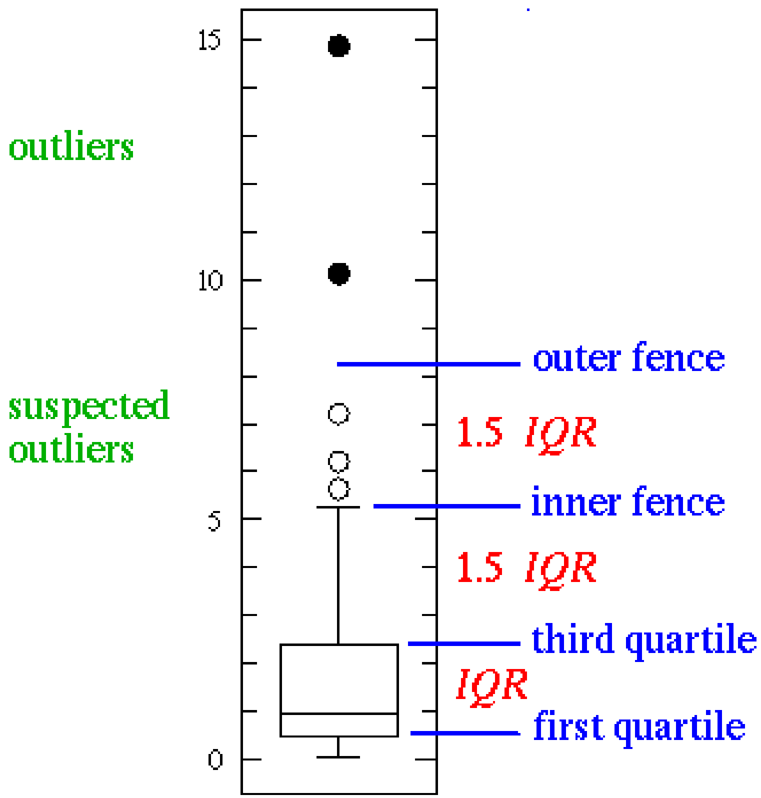


The central rectangle spans the first and third quartile which is also called as Inter quartile range (IQR)

The segment inside the rectangle shows the median

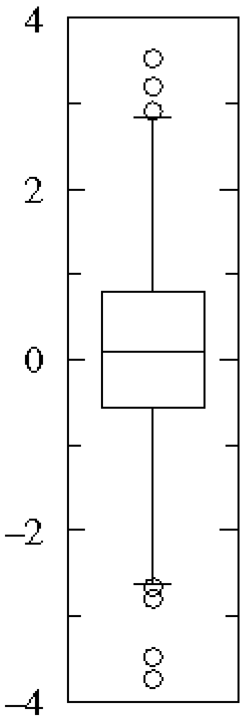
The whiskers above and below the box shows the locations of minimum and maximum

Outlier analysis



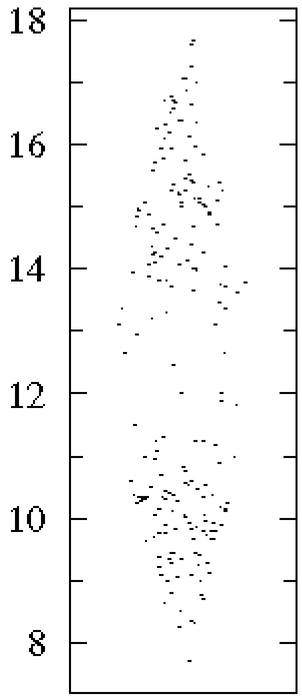
- John Tukey has provided a precise definition for two types of outliers:
  - **Outliers** are either 3×*IQR* or more above the third quartile or 3×*IQR* or more below the first quartile.
  - **Suspected outliers** are slightly more central versions of outliers: either 1.5×*IQR* or more above the third quartile or 1.5×*IQR* or more below the first quartile.

Outlier analysis



- Outliers are not necessarily 'bad'
- Outliers may deserve special consideration as they may be key to some phenomenon under study or result of human errors
- Some applications areas:
  - intrusion detection in cyber security
  - military surveillance for enemy activities
  - fraud detection for credit cards
  - fault detection in safety critical systems

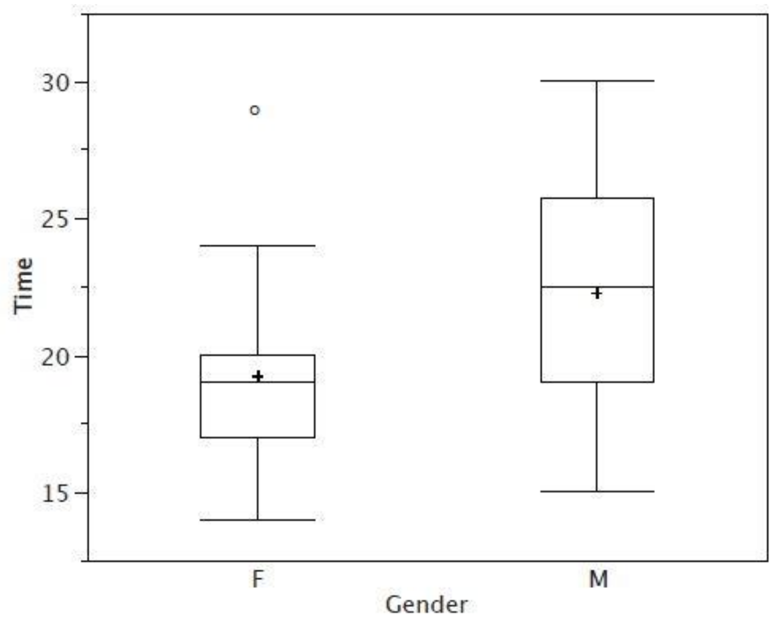
Outlier analysis



- In some cases, a box plots can be deceptive, when the data has more than one pattern.
  - Income levels are recorded for rich and poor

Parallel box plots

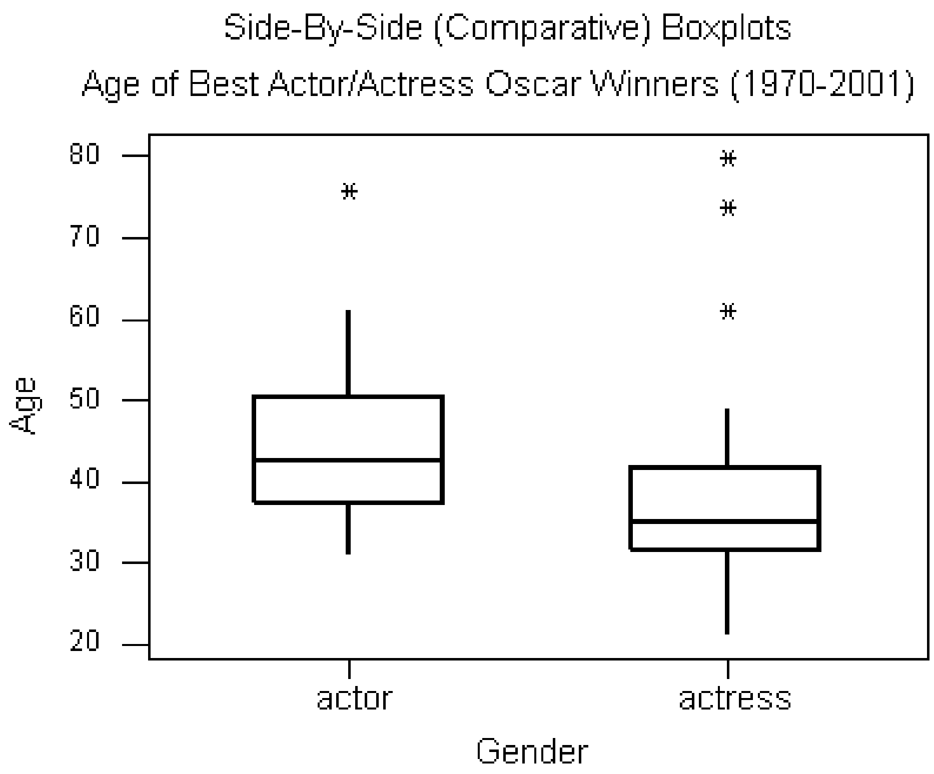
- Students in an introductory statistics class were each given three tasks.
  - In the "words" task, students read the names of 60 color words written in black ink;
  - in the "color" task, students named the colors of 60 rectangles;
  - in the "interference" task, students named the ink color of 60 conflicting color words.
  - The times to read the stimuli were recorded. There were 31 female and 16 male students



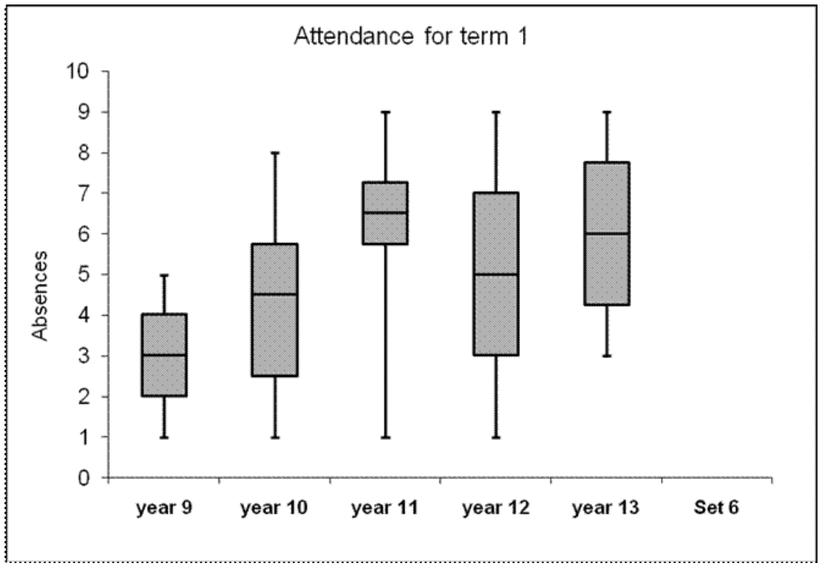
We see that half the women's times are between 17 and 20 seconds, whereas half the men's times are between 19 and 25.5. We also see that women generally named the colors faster than the men did, although one woman was slower than almost all of the men

Read data stroop.xlsx and plot the above for your practice

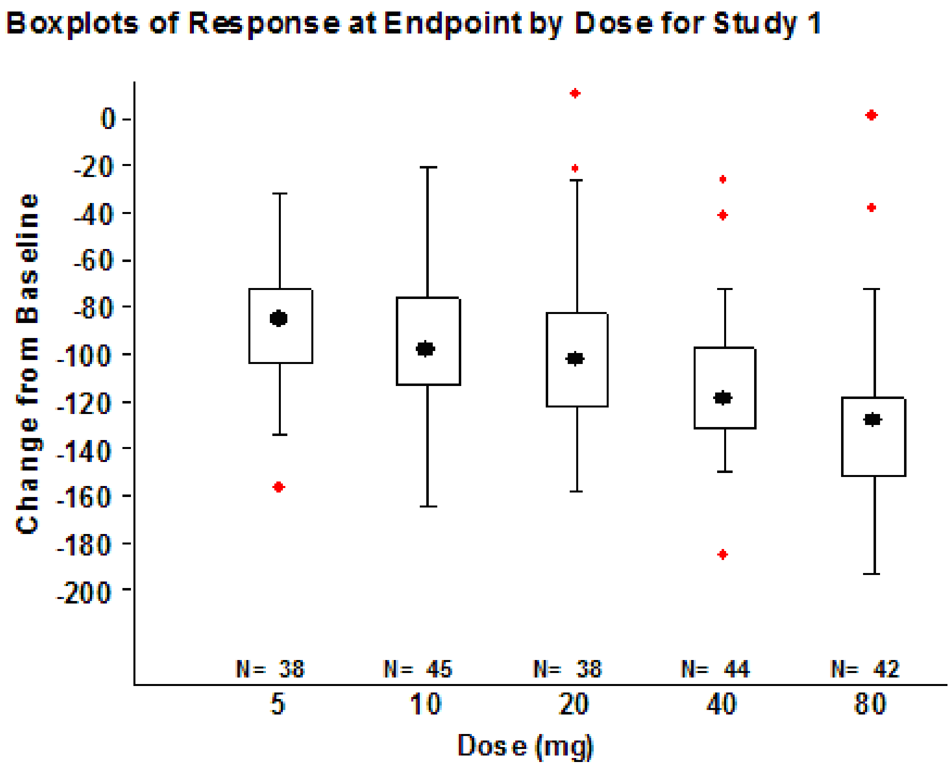
Parallel box plots for comparison



The distribution of absentees for term I across 5 years



Parallel box plots

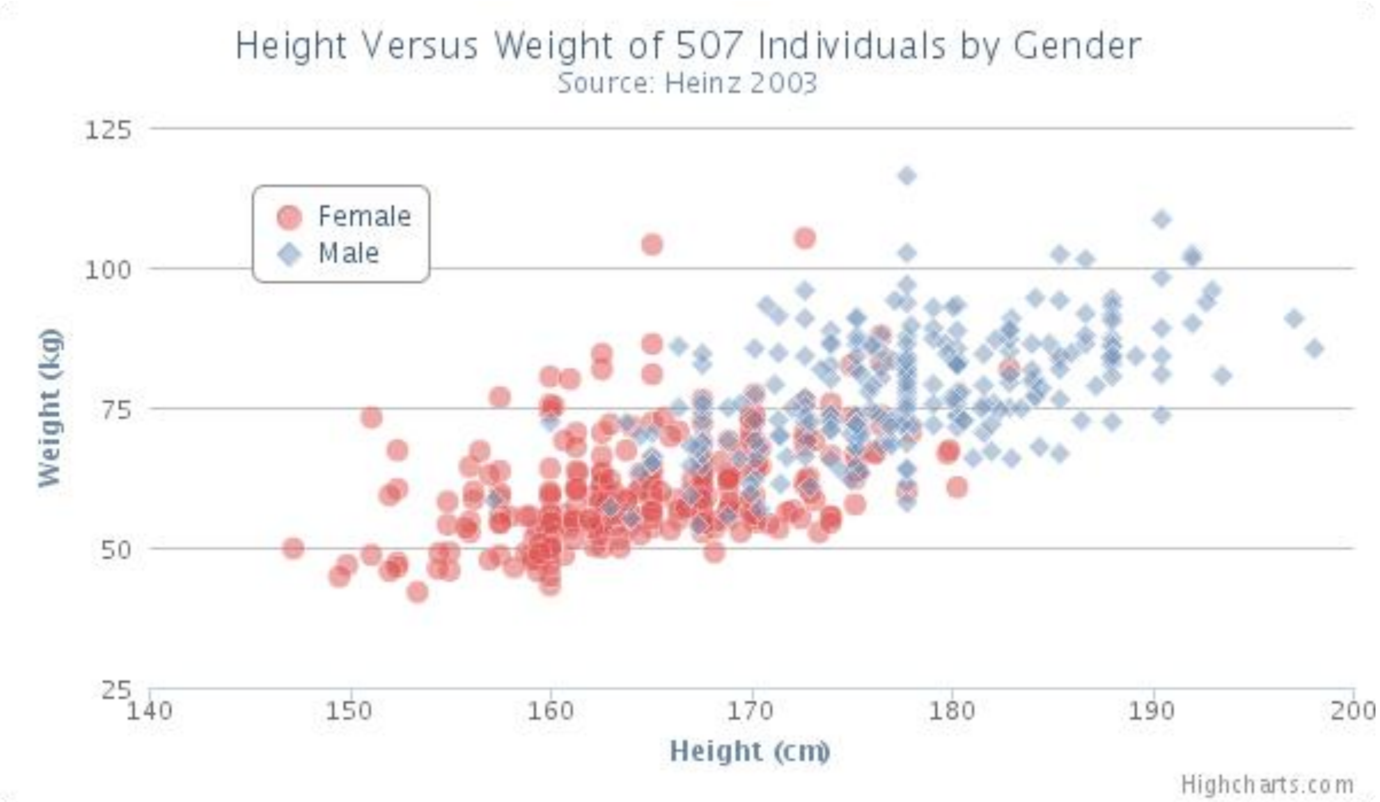


For more exercises: http://onlinestatbook.com/2/graphing_distributions/boxplot_demo.html

Scatter plots

- A scatter plot (X,Y) has points that show the relationship between two sets of numeric data.
  - Specifically, these plots show how much one variable is affected by the another.
  - Correlation between two variables

Height and weight of individuals by Gender

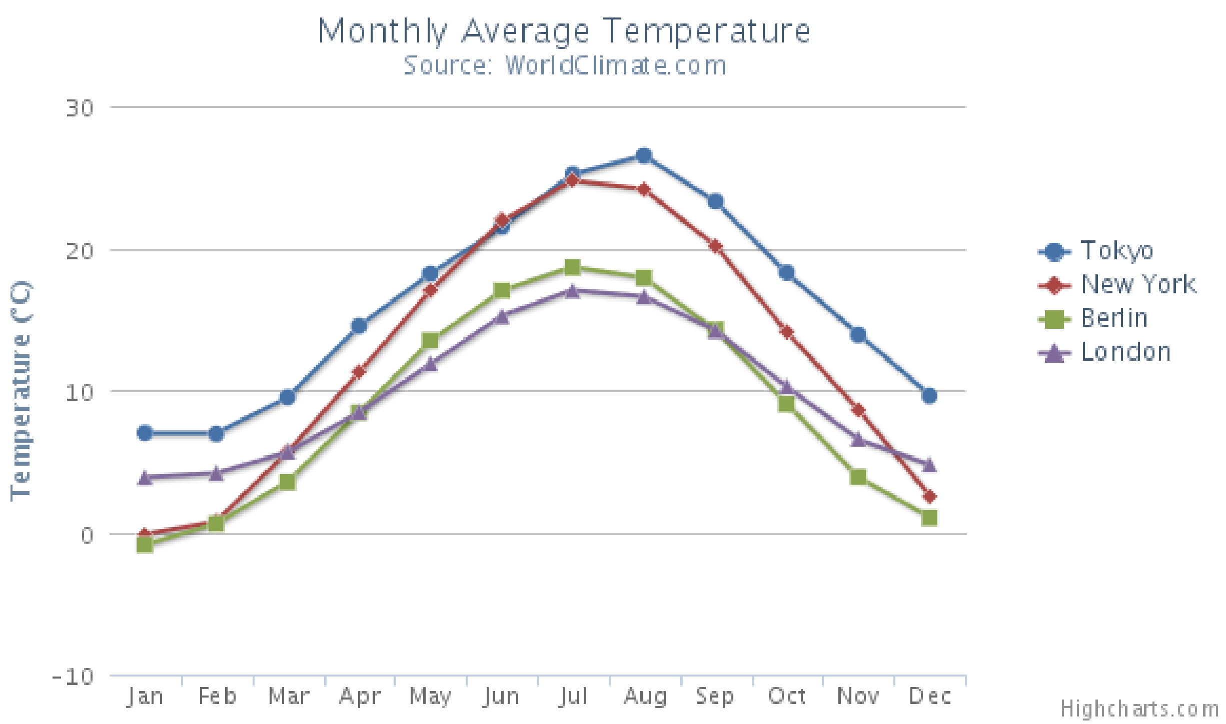


We note that the height and weight move together and show a positive correlation

Line charts

- Line chart has series of data points connected using a line.
- It is similar to scatter plot except that the measurements are ordered and joined using a line
- It is used very often to visualize the trend in the data over intervals of time

Change in monthly average temperature in four locations across months in a year



We note that the temperature goes up starting June and fall in September in all four locations. We note that there is positive correlation as the temperatures across Locations fall and rise together.

N/C	0	1	2	3
0		Bar	Grouped Bar	
1	Histogram/ Box or Line	Box plot for all levels/Histogram with color mapping	Point graph with colormapping	
2	Scatter	Scatter with color/shape mapping	Point graph with color and size	Faceting
3	Scatter and map size to third one	Scatter with size, color mapping	Point graph with color, size and shape	

Inspire...Educate...Transform.

International School of Engineering

Plot 63/A, 1st Floor, Road # 13, Film Nagar, Jubilee Hills, Hyderabad - 500 033

For Individuals: +91-9502334561/63 or 040-65743991

For Corporates: +91-9618483483

Web: http://www.insofe.edu.in

Facebook: https://www.facebook.com/insofe

Twitter: https://twitter.com/Insofeedu

YouTube: http://www.youtube.com/InsofeVideos

SlideShare: http://www.slideshare.net/INSOFE

LinkedIn: http://www.linkedin.com/company/international-school-of-engineering