Machine Learning & Data Mining

- Exploratory Data Analysis -

Kyung-Ah Sohn

Ajou University

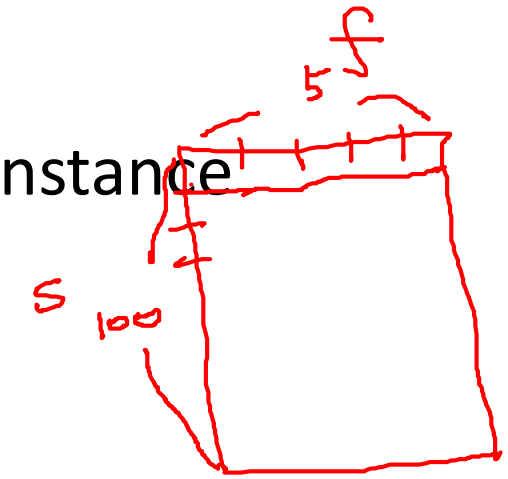
Content

- Data
- Data quality
- Exploratory data analysis
  - Numerical summary
  - Graphical summary

DATA

Terminology

- Components of the input (data)
  - Instances: the individual, independent examples of a concep
    - Sample/point/record/object/case
  - Attributes: measuring aspects of an instance
    - Variable/feature/characteristic/field



Data

- Collection of data instances and their attributes
- An attribute or a feature is a property or characteristic of an instance
- A collection of attributes describe an instance

Attributes (variable/feature/characteristics/field)

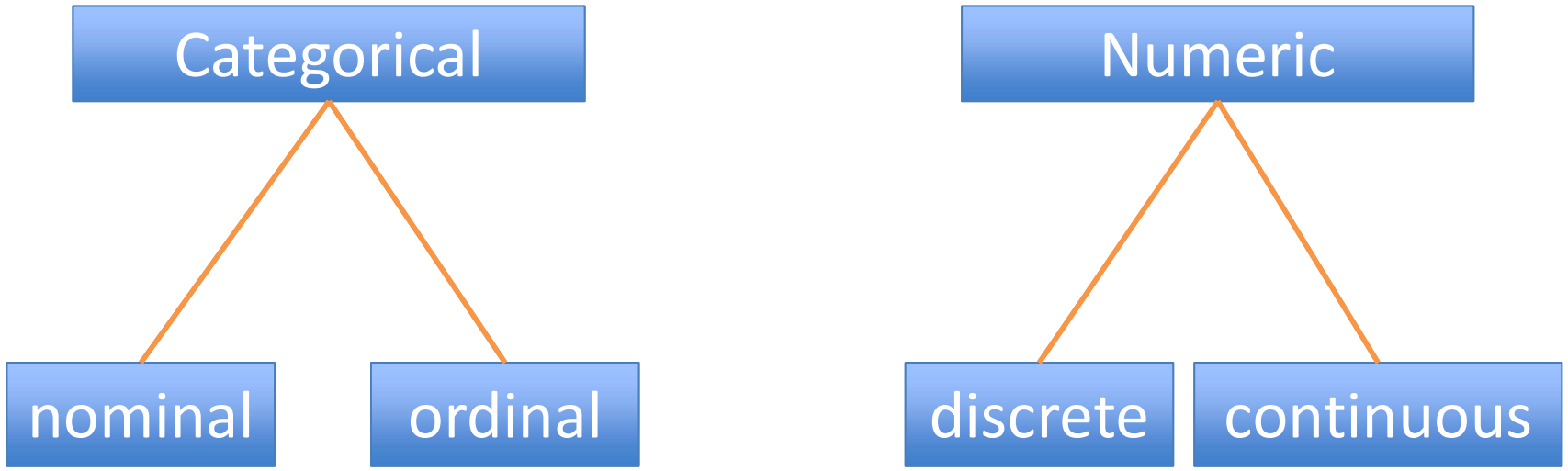
Instances (sample/ record/ point/ case/ object)

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	Normal	False	Yes
...	...	...	...	...

Attribute Values

- Attribute/feature values are numbers or symbols assigned to an attribute
- Distinction between attributes and attribute values
  - Same attribute can be measured in feet or meters
- Possible attribute types ("levels of measurement"):
  - *Nominal, ordinal, interval* and *ratio, …*

Types of Variables



categories
(명목형)

order matters
(순위형)

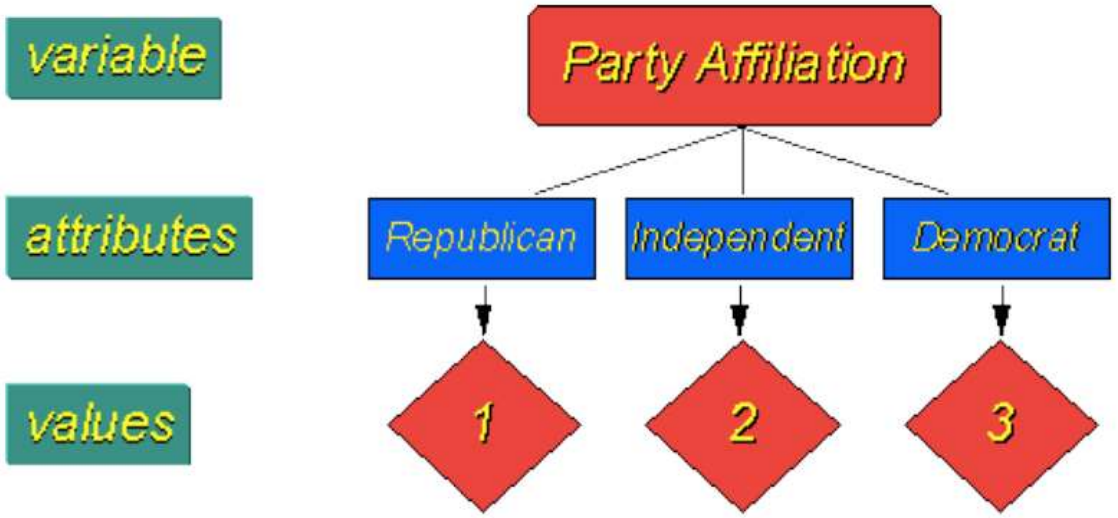
numerical

uninterrupted

Nominal quantities

- Values are distinct symbols
  - Values themselves serve only as labels or names
  - *Nominal* comes from the Latin word for name
- Example: attribute “outlook” from weather data
  - Values: “sunny”,”overcast”, and “rainy”
- No relation is implied among nominal values (no ordering or distance measure)
- Only equality tests can be performed

Nominal Variables



- Numerical values have no semantic meaning, just indices
- No ordering implied

- Example
  - ID numbers, eye color, zip codes
  - Jersey numbers in basketball

Ordinal quantities

- Impose order on values
- But: no distance between values defined

- Example: attribute “temperature” in weather data
  - Values: “hot” > “mild” > “cool”

- Note: addition and subtraction don’t make sense

- Example: Rankings, grades, height in {tall, medium, short}

- Example rule: temperature < hot ⇒ play = yes

- Distinction between nominal and ordinal not always clear (e.g. attribute “outlook”)

Discrete and Continuous

- Discrete attributes
  - Has only a finite or countably infinite set of values
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes
- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight

Why is this important?

- Many models require data to be represented in a speci form
- e.g. real-valued vectors

- What do we do with non-real valued inputs?
  - Nominal with M values
    - Not appropriate to “map” 1 to M (why?)
    - Could use M binary “indicator” variables

Mixed data

- Many real-world data sets have multiple types of variables

- Unfortunately, many data analysis algorithms are suited to only one type of data…

범주형 데이터 표현

- T-shirt size: M < L < XL
- Color: blue, green, red
- 정수로 매핑:
  - blue: 0, green: 1, red: 2
  - 문제점
    - 알고리즘이 이 값을 수치형으로 간주하여 기대하지 않은 (최적
      아닌) 결과를 초래할 수 있음.

	price	size	color
0	10.1	M	Green
1	13.5	L	Red
2	15.3	XL	Blue

One-hot encoding

- Standard approach for categorical data
- Create a binary vector to represent each categorical value
- Allows the representation of categorical data to be more expressive

	A	B	C	D	E	F	G	H	I
1	Original data:			One-hot encoding format:					
2	id	Color		id	White	Red	Black	Purple	Gold
3	1	White		1	1	0	0	0	0
4	2	Red		2	0	1	0	0	0
5	3	Black		3	0	0	1	0	0
6	4	Purple		4	0	0	0	1	0
7	5	Gold		5	0	0	0	0	1
8									

One-hot 인코딩 예시, 출처: stackoverflow

After one-hot encoding

	price	size	color
0	10.1	1	Green
1	13.5	2	Red
2	15.3	3	Blue

	price	size	color_blue	color_green	color_red
0	10.1	1	0	1	0
1	13.5	2	0	0	1
2	15.3	3	1	0	0

데이터 품질

Data Quality

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

- Examples of data quality problems:
  - missing values
  - Noise and outliers
  - duplicate data

결측 데이터

- Reasons for missing values
  - Errors during data collection
  - Information is not collected
    (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)

- 데이터 테이블에서 주로 빈 공간이나 NaN(not a number)로 표시

- 결측 데이터를 무시하는 경우 예측할 수 없는 결과를 낳게 됨
  - 예: 0으로 표시 (x)

결측 데이터 처리

- Handling missing values
  - Eliminate Data Objects
  - Estimate Missing Values (imputation)

결측 데이터 제거 [1]

```
import pandas as pd
from io import StringIO

csv_data = '''A,B,C,D
1.0,2.0,3.0,4.0
5.0,6.0,,8.0
10.0,11.0,12.0,'''

# If you are using Python 2.7, you need
# to convert the string to unicode:
# csv_data = unicode(csv_data)

df = pd.read_csv(StringIO(csv_data))
df
```

```
     A   B   C   D
0    1   2   3   4
1    5   6 NaN   8
2    0  11  12 NaN
```

결측값을 갖는 행 제거

```
>>> df.dropna()
```

```
     A   B   C   D
0    1   2   3   4
```

결측값을 갖는 열 제거

```
>>> df.dropna(axis=1)
```

```
     A   B
0    1   2
1    5   6
2    0  11
```

https://github.com/rasbt/python-machine-learning-book/blob/master/code/ch04/ch04.ipynb

결측 데이터의 제거 [2]

only drop rows where all columns are NaN

```
>>> df.dropna(how='all')
```

	A	B	C	D
0	1	2	3	4
1	5	6	NaN	8
2	0	11	12	NaN

only drop rows that have not at least 4 non-NaN values

```
>>> df.dropna(thresh=4)
```

	A	B	C	D
0	1	2	3	4

only drop rows where NaN appear in specific columns (here, 'C')

```
>>> df.dropna(subset=['C'])
```

	A	B	C	D
0	1	2	3	4
1	0	11	12	NaN

결측 데이터 제거 [3]

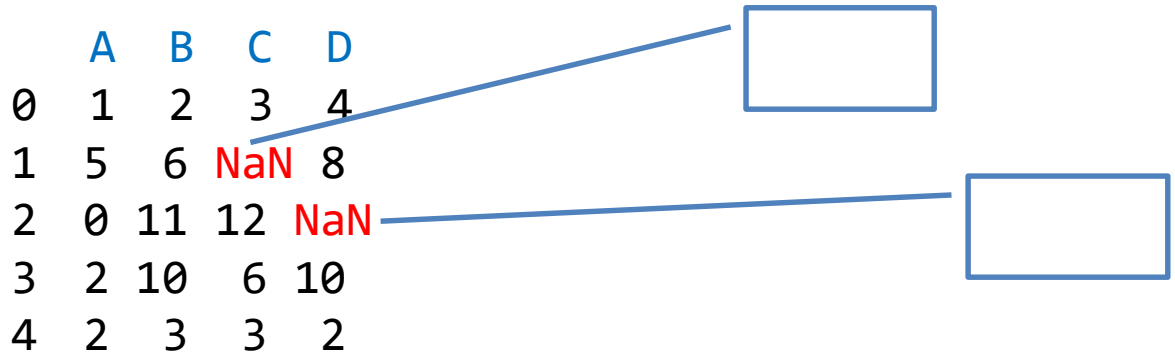
- 장점: 편리함
- 단점: 정보 손실
  - 샘플 수 부족 → 신뢰성 있는 분석 불가
  - 피처 수 부족 → 샘플에 관한 정보 부족

결측값 보정

- Use mean value
- Nearest-neighbor based imputation
- Model-based imputation

평균보정법

- 열(column) 평균 사용



	A	B	C	D
0	1	2	3	4
1	5	6	NaN	8
2	0	11	12	NaN
3	2	10	6	10
4	2	3	3	2

```
>>> df.fillna(df.mean())
```

* 행 평균?

* Median?

* 범주형 변수이면?

K-Nearest Neighbor Imputation

- Find the k closest neighbors and use the non-missing values in the neighbors

	A	B	C	D
0	1	2	3	4
1	5	6	NaN	8
2	2	10	6	10
3	2	3	3	2

$d(x_1,x_0)^2 = 4^2+4^2+4^2$

$d(x_1,x_2)^2 = 3^2+4^2+2^2$

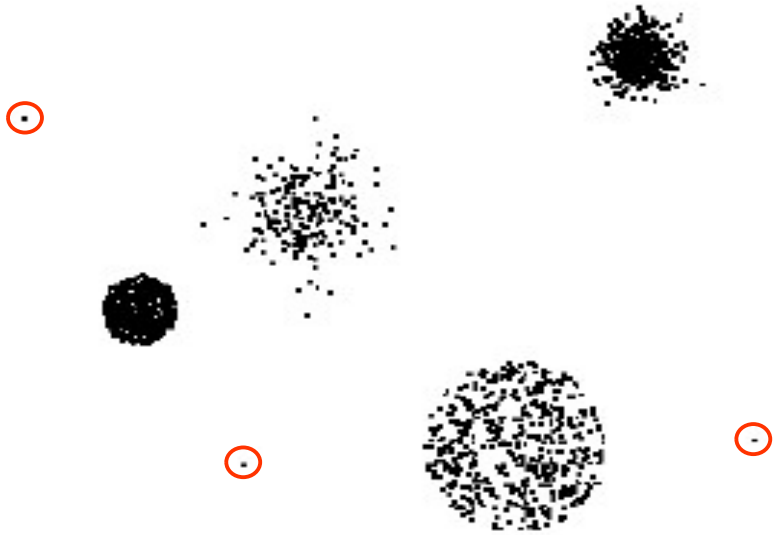
$d(x_1,x_3)^2 = 3^2+3^2+6^2$

K=1:

K=2:

Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



이상값

- 발생 원인
  - Measurement error
  - Human error
  - Sampling error
  - Natural outlier

Outlier detection

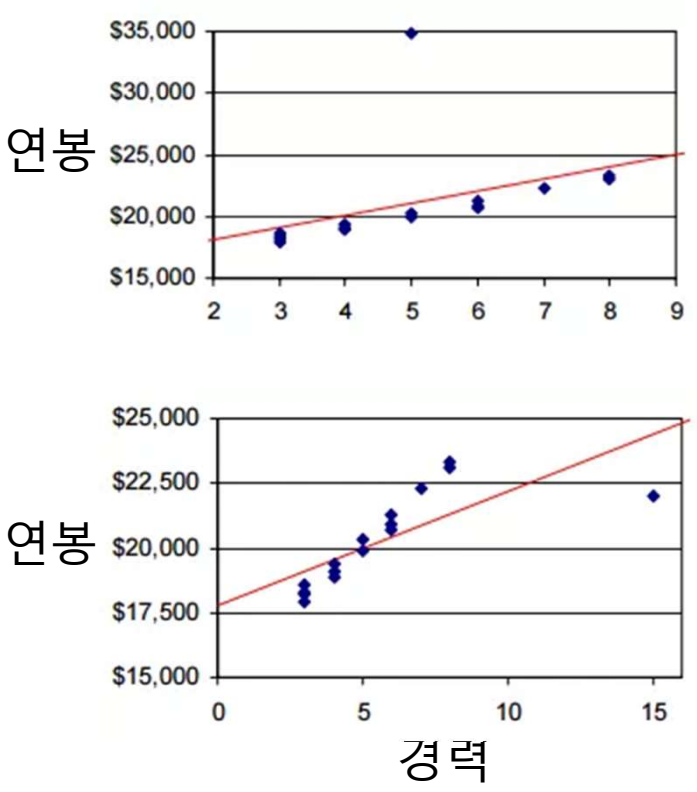
- Visualize the distribution
  - Boxplot, histogram: one variable
  - Scatter plot: two variables or more
- Use statistical methods

이상값 처리 [1]

- 단순 삭제
- 다른 값으로 대체
  - 평균 등
  - 회귀모형 등 예측 모델을 이용

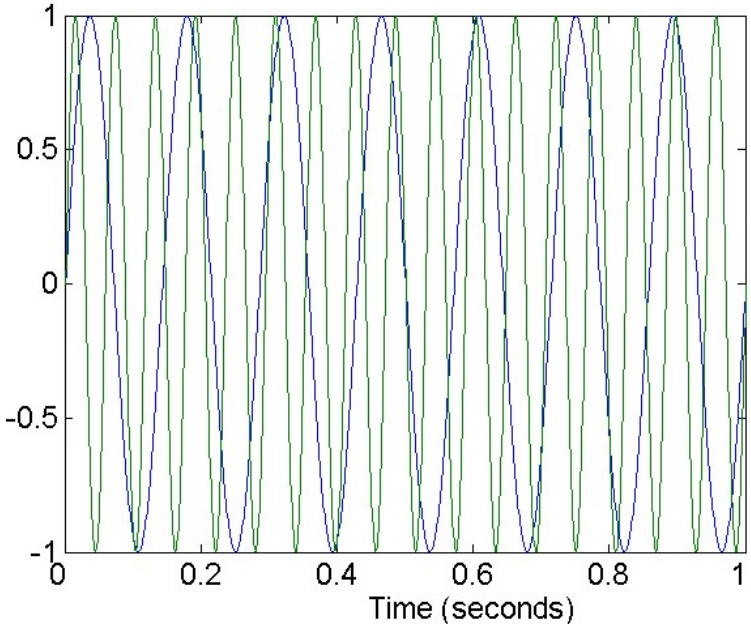
이상값 처리 [2]

- 변수화
  - 예: 의사 등 전문적 종사자
  - 전문직 종사 여부를 0-1로 변수화
- 리샘플링
  - 이상치를 삭제하고 분석 범위를 10년 이내 경력자로 하여 새로 모델 구성

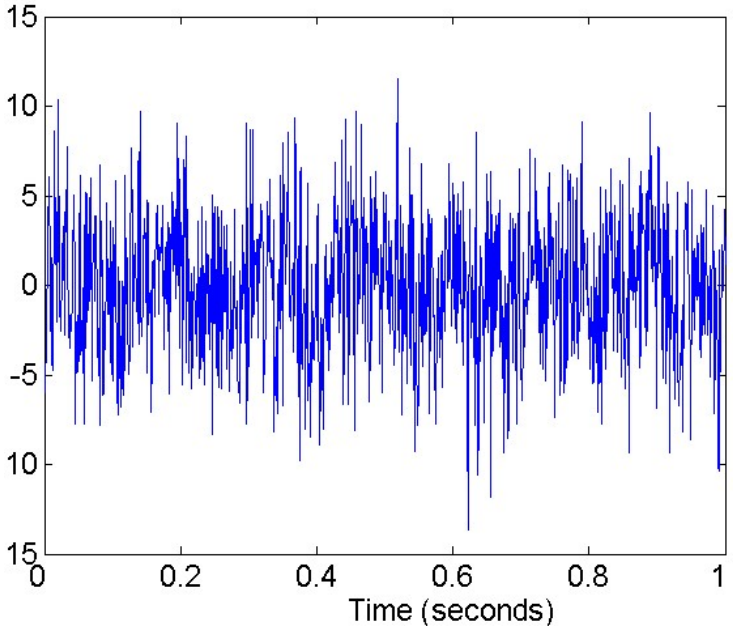


Noise

- Noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone



Two Sine Waves



Two Sine Waves + Noise

노이즈 처리

- Collecting more data
  - The more data, the better you're able to identify the underlying pattern
- Smoothing by binning (discretization)
- Regularization
  - L1/L2 regularization

(will be discussed later)

EXPLORATORY DATA ANALYSIS

EDA

- Graphical summaries of data
  - Visualization
- Numerical summaries of data
  - Descriptive statistics

Exploratory Data Analysis (EDA)

- To get a general sense of the data
- You should always look at every variable - you will learn something!
- Data-driven (model-free)
- Think interactive and visual
  - You can use more than 2 dimensions (space, color, time, …)
- Especially useful in early stages of data analysis
  - detect outliers (e.g. assess data quality)
  - test assumptions (e.g. normal distributions or skewed?)
  - identify useful raw data & transforms (e.g. log(x))
- Bottom line: it is always well worth looking at your data!

Numerical Summaries of Data

- Not visual
- Summary statistics
  - mean, median
  - mode: the most common value
  - variance, standard deviation
  - quartiles
  - Number of distinct values for a categorical variable

- Don't need to report all of theses: Bottom line…do these numbers make sense???

Exploring numeric variables

- Measuring the central tendency

- Measuring spread – quartiles and the five-number summary

```
> summary(usedcars$year)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2000    2008    2009    2009    2010    2012
```

Using the mean in data analysis

- Back in the mid-1980's at the University of North Carolina, the average starting salary of geography students was over $100,000. Knowing that, would you have considered making a career change?
- What if I told you that basketball great Michael Jordan – formerly the world's highest paid athlete – graduated from UNC with a degree in geography?

http://blog.minitab.com/blog/michelle-paret/using-the-mean-its-not-always-a-slam-dunk

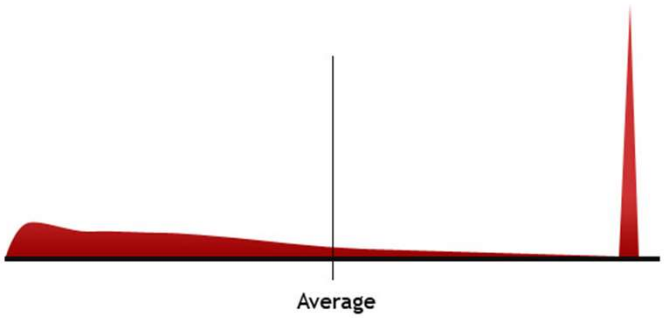
The Mean can Mislead

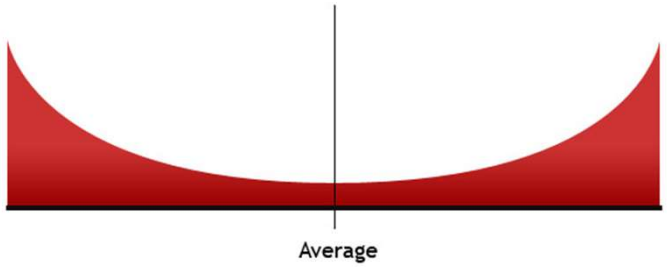
- Jordan's earnings from his athletic career raises the "average" salary for geography graduates in a way that doesn't accurately convey what graduates are likely to earn

- By almost any measure, Jordan's earnings would be an outlier

- How to identify this anomaly?

he average is not a good representation of the true enter of the data





http://www.baekdal.com/notes/twitter-average

Median

Median: the exact middle value

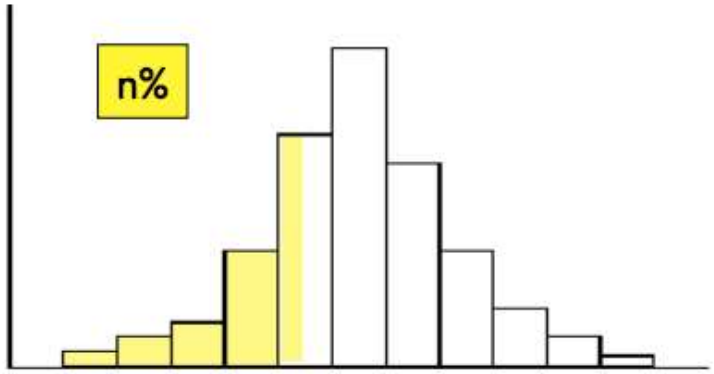
- Useful for skewed distributions or data with outliers
- More robust than mean
- Difficult to handle theoretically (no easy mathematical formula)

Example

Data	1 2 3 4 5	1 2 3 4 100
Mean		
Median		

Percentiles (aka Quantiles)

- The **n^{th} percentile** is a value such that n% of the observations fall at or below of it



- Q1 : 25^{th} percentile
- Median: 50^{th} percentile
- Q3 : 75^{th} percentile
- IQR: Interquartile range (25 to 75%: Q3-Q1)

Visualizing numeric variables

- Boxplot: a common visualization of the five-number summary
- Histogram: another way to graphically depict the spread of a numeric variable

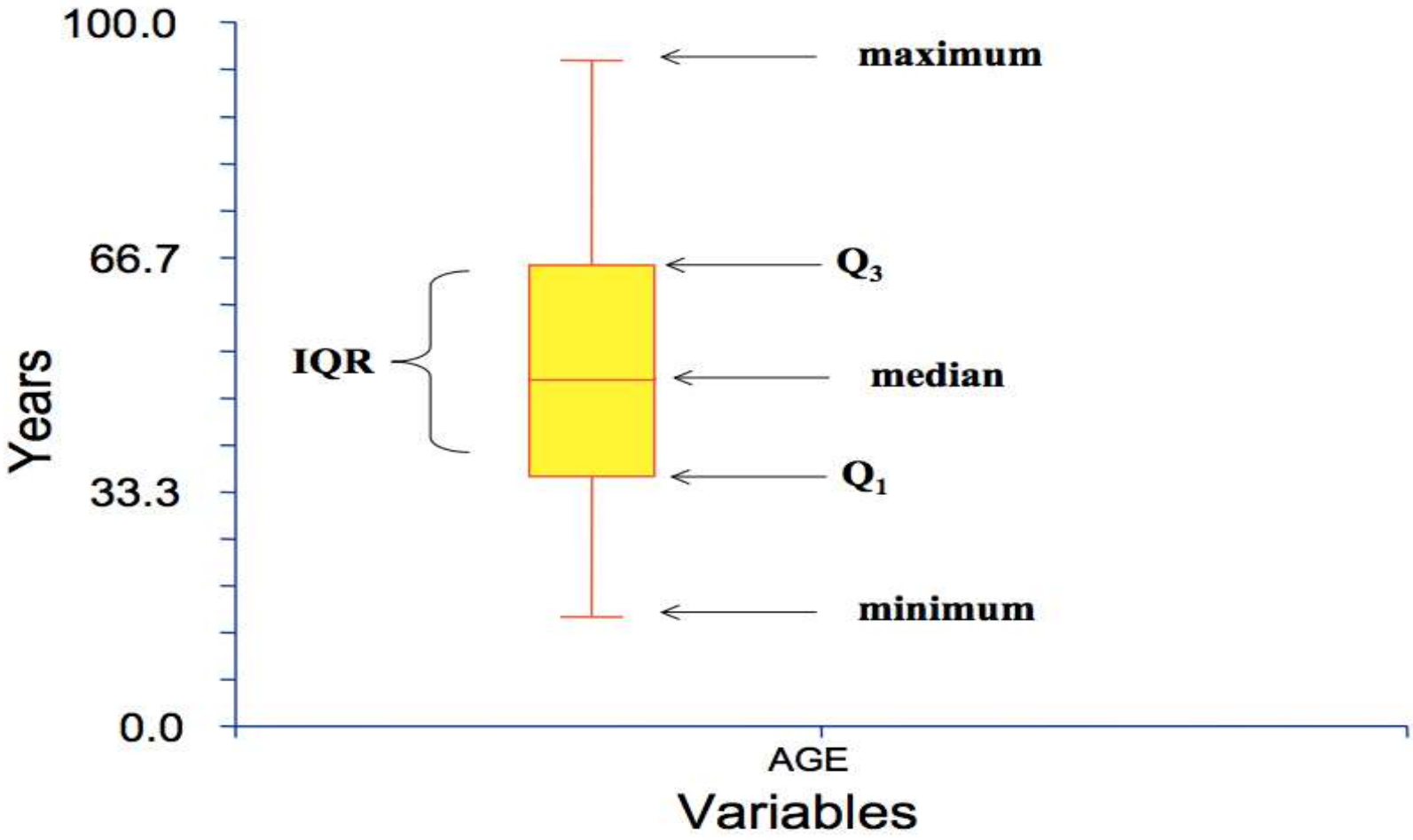
Boxplot

- x-axis: categorical variable
- y-axis: real-valued or integer variable

- For each group, the boxplot shows
  - Median
  - Interquartile range (25 to 75%) (IQR)
  - Whiskers (most extreme points not considered to be outliers)
  - Outliers

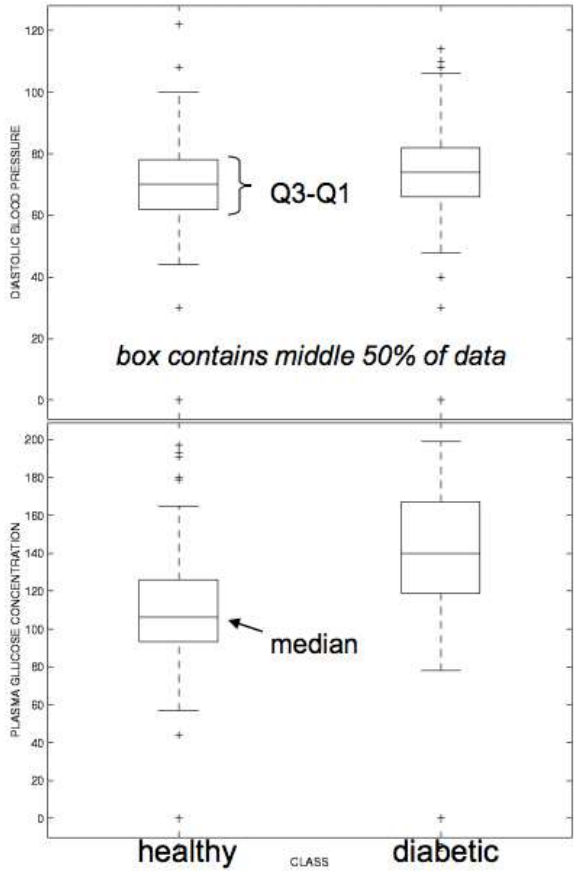
- Negatives
  - Over-plotting
  - Hard to tell distributional shape

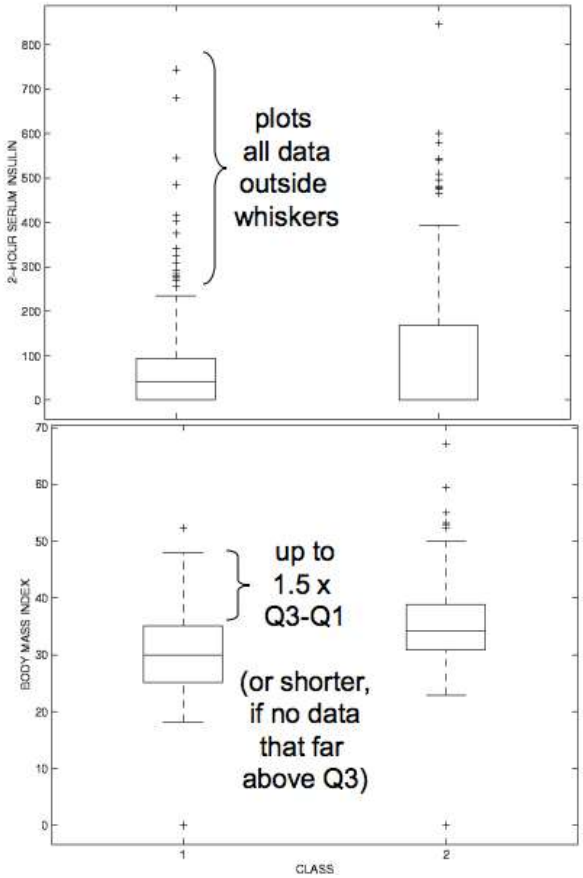
Boxplot


100.0
66.7
33.3
0.0
Years
maximum
Q_3
IQR
median
Q_1
minimum
AGE
Variables

Box (and Whisker) Plots
- Pima Indian data-

```
> library(MASS)
> data(Pima.te)
```

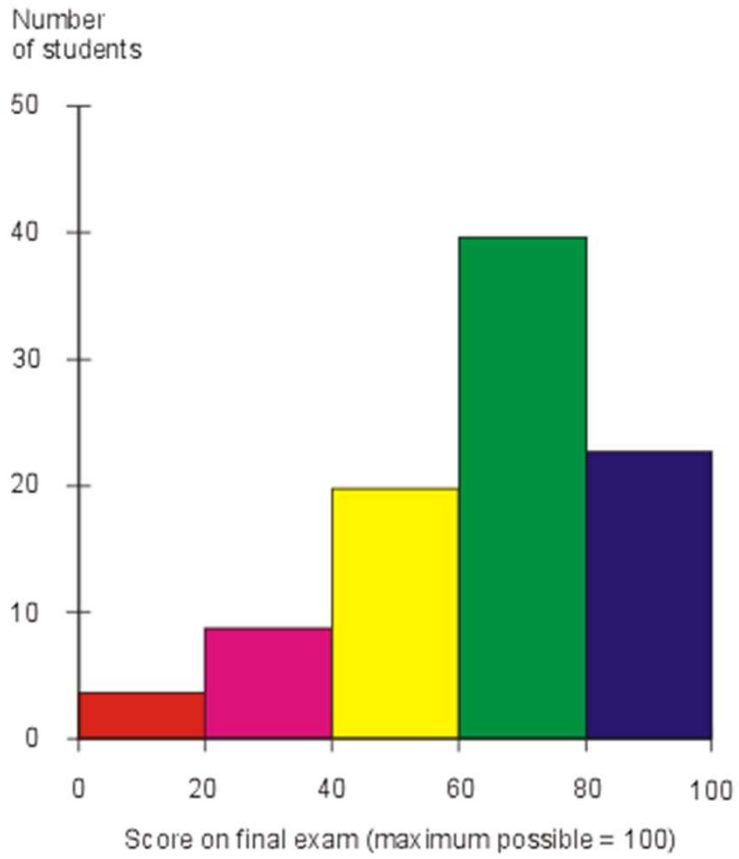





```
## e.g.
> boxplot(Pima.te$bmi ~ Pima.te$type)
```

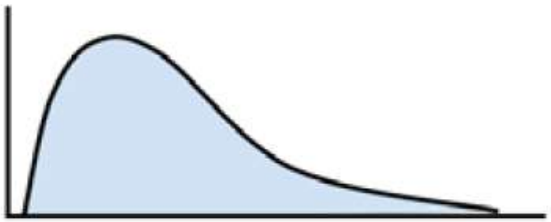
Histograms

- Histogram
  - Split data range into equal-sized bins
  - Count the number of data points falling into each bin
  - x axis: values of the variable
  - y axis: frequency (counts for each bin)

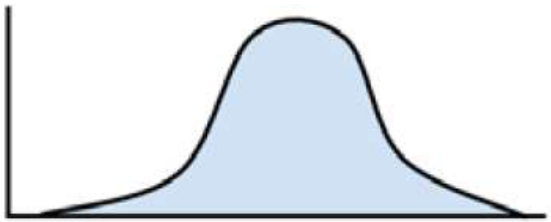


http://www.webquest.hawaii.edu/kahihi/mathdictionary/H/histogram.php

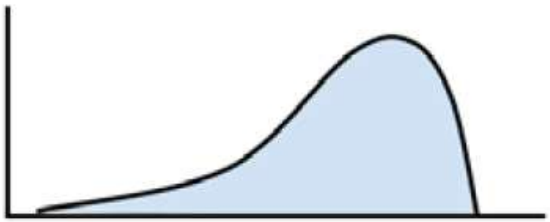
Shape of histograms



Right Skew

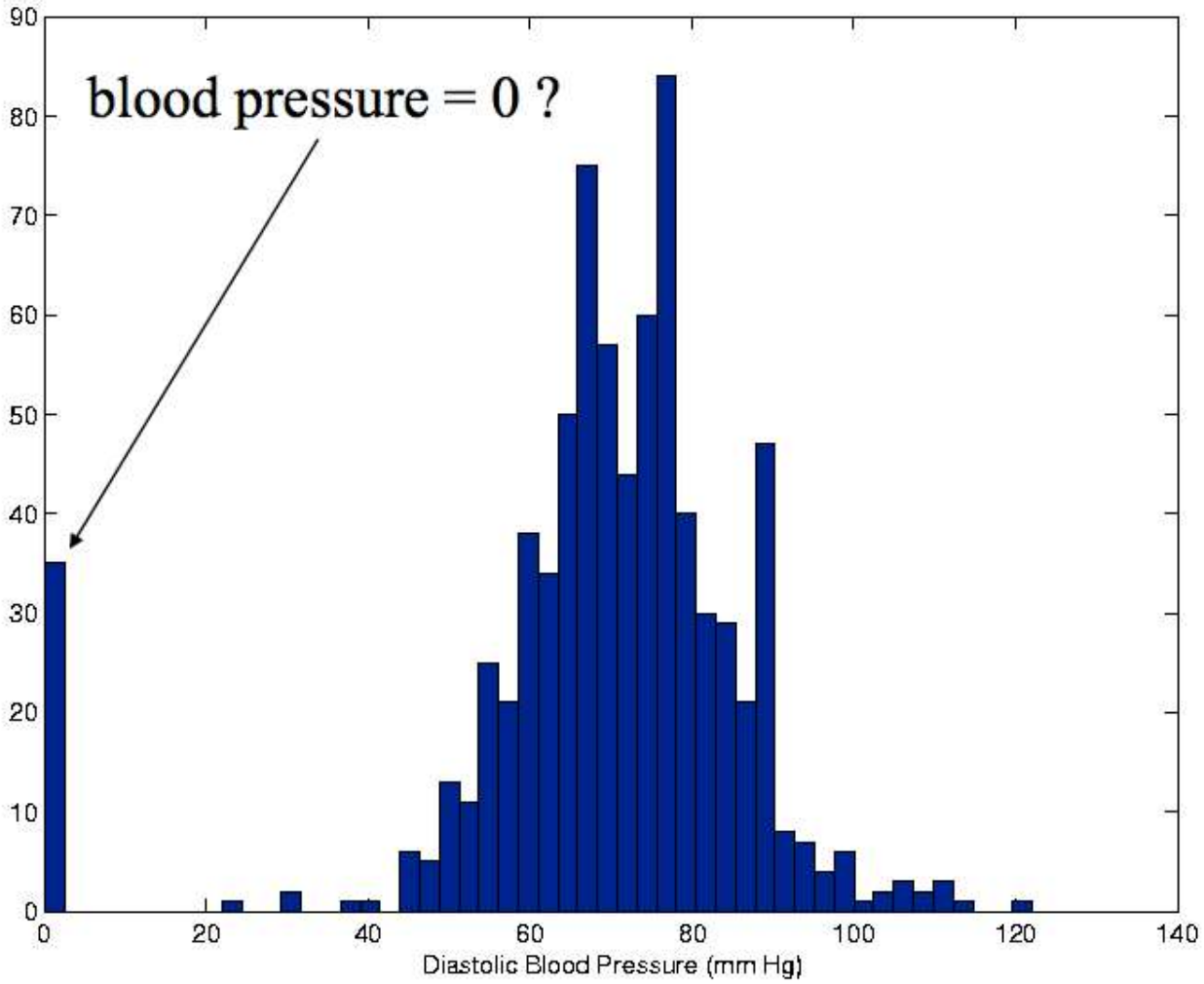


No Skew



Left Skew

Histogram detecting outliers



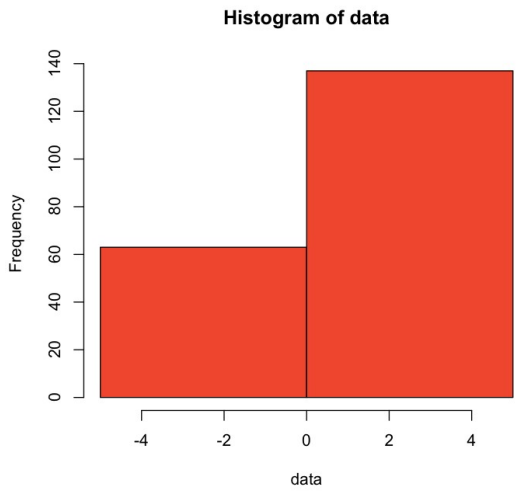
Issues with Histograms

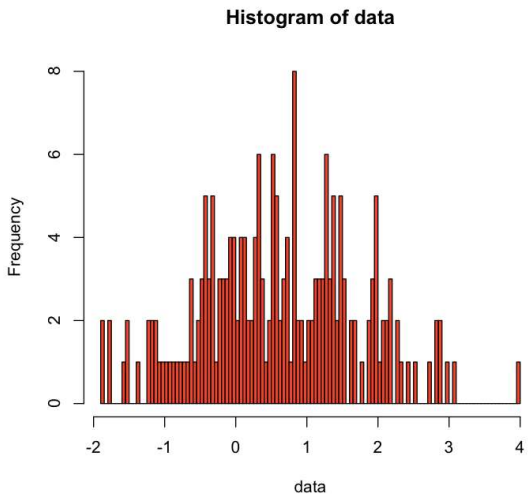
- Histograms can be misleading for small data sets
- For large data sets, histograms can be quite effective at illustrating general properties of the distribution
- Effective only with one variable
- Can smooth histogram using a variety of techniques

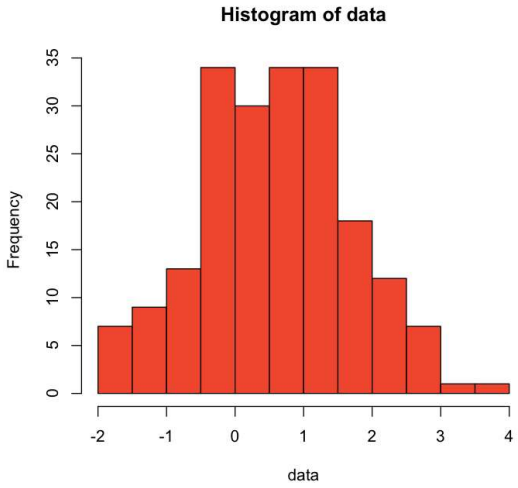
Effect of Bin Size on Histogram

```
> data <- c(rnorm(100), rnorm(100)+1)
```

: Simulated 100 points from N(0,1) and 100 points from N(1,1)








```
> hist(data,breaks=2, col="red")
```

```
> hist(data,breaks=100, col="red")
```

```
> hist(data,breaks=10, col="red")
```

Exploring categorical variables

- Categorical data is examined using tables rather than summary statistics
  - e.g. one-way table

- Measuring the central tendency – the mode
  - The value occurring most often
  - Often used for categorical data

Exploring relationships between variables

- Bivariate, or multivariate relationships

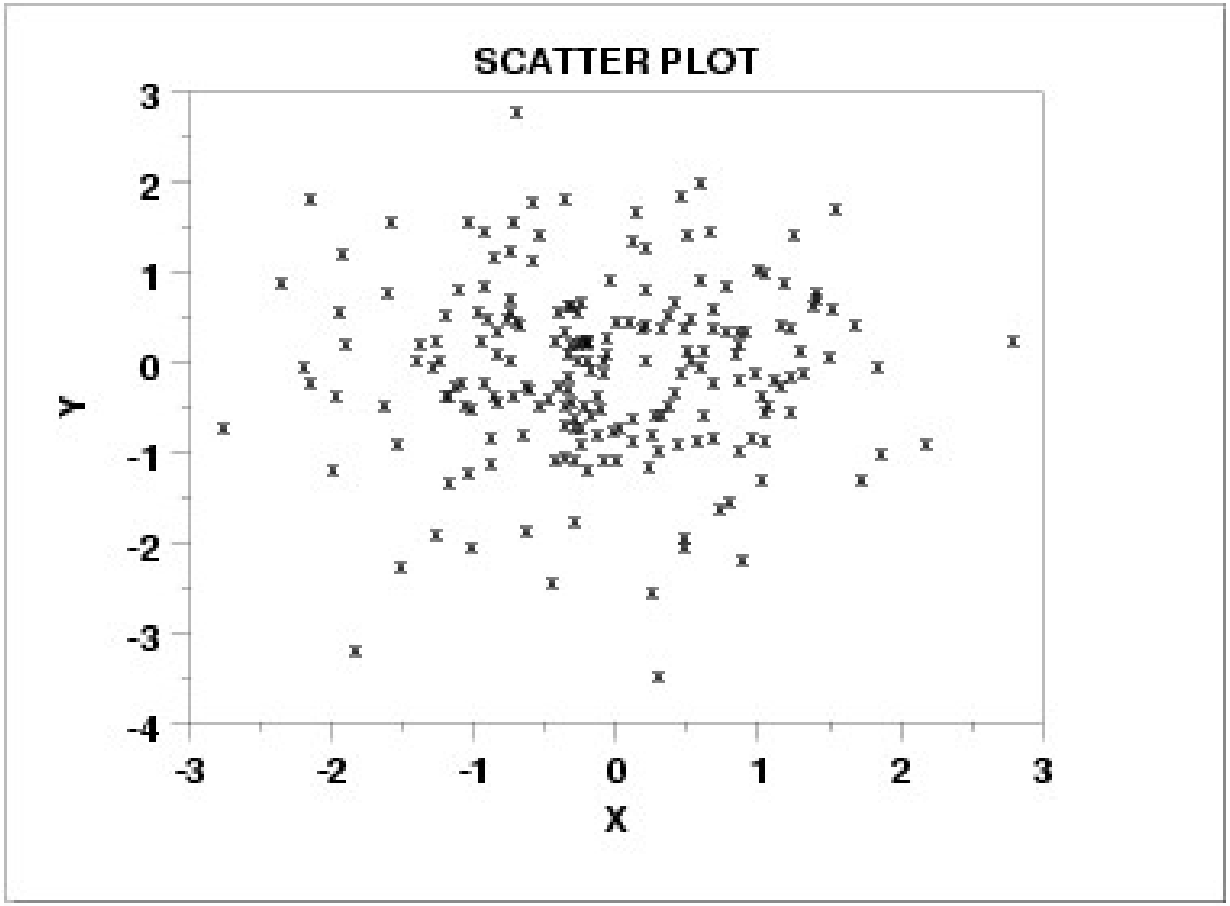
- Scatter plots, two-way cross-tabulation (contingency table)

2D Scatter plots

- standard tool to display relation between 2 variables

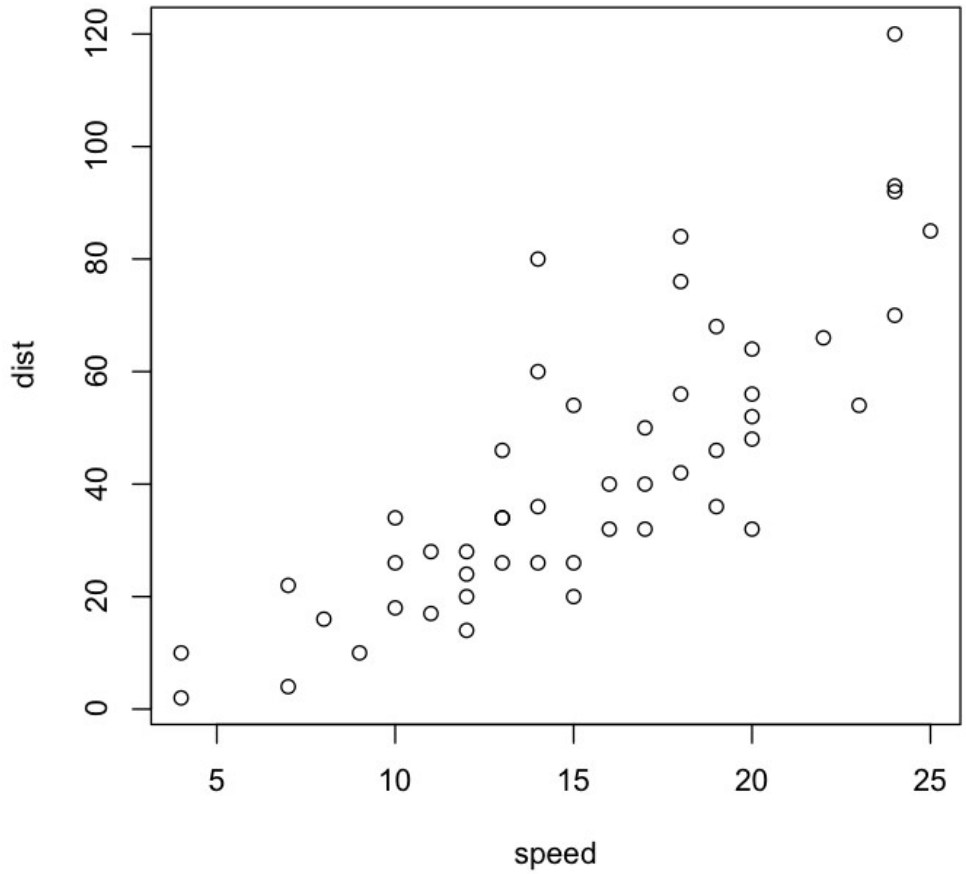
- Useful to answer
  - x and y related?
  - Variance(y) depend on x?
  - Outliers?

Scatter Plot: No apparent relationship



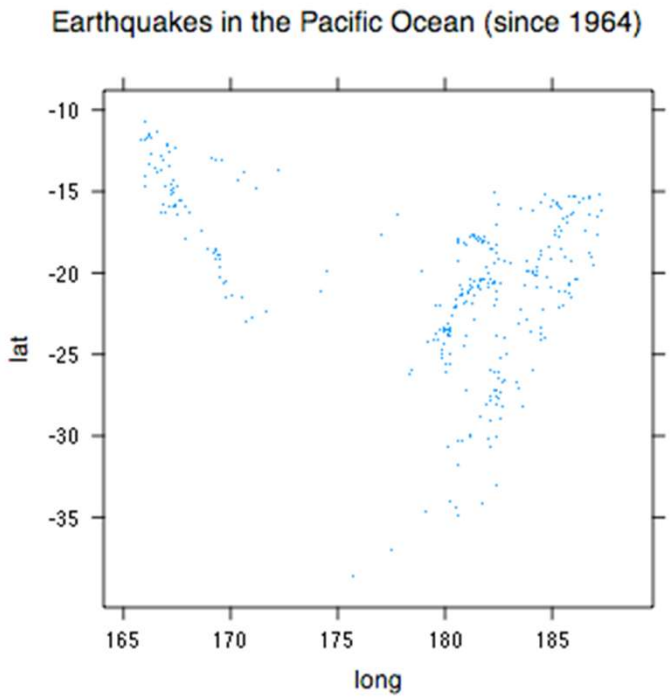
Scatter plot

- Speed and Stopping Distances of Cars -



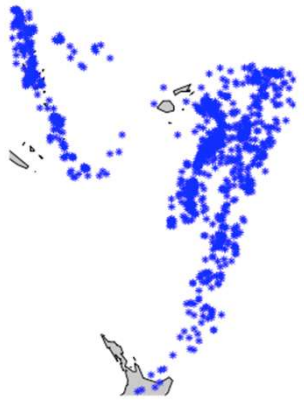
Spatial Data

- If your data has a geographic component, be sure to exploit it
- Data from cities/states/zip cods – easy to get lat/long
- Can plot as scatterplot

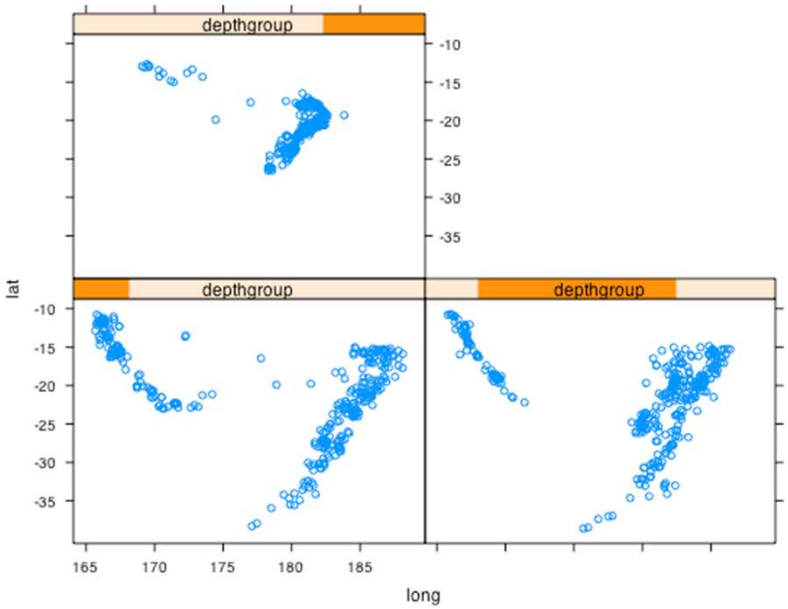


Multivariate: More than two variables

- Get creative!
- Conditioning on variables
  - trellis or lattice plots
  - Infinite possibilities

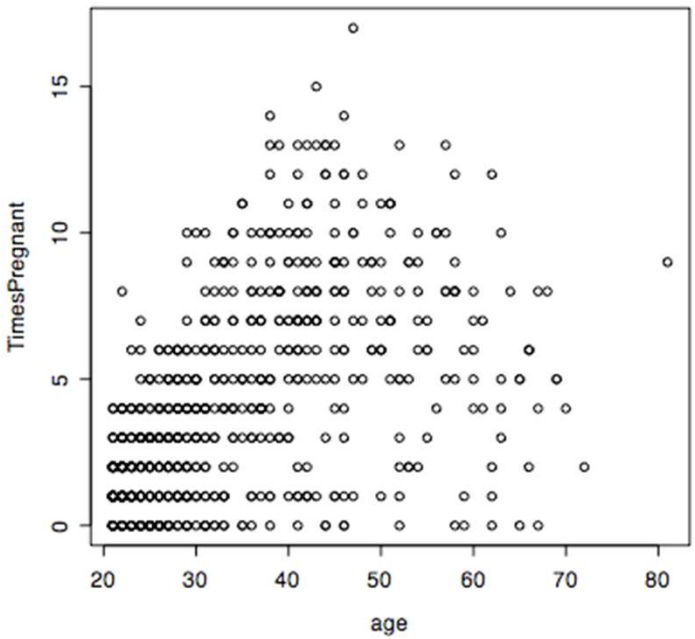


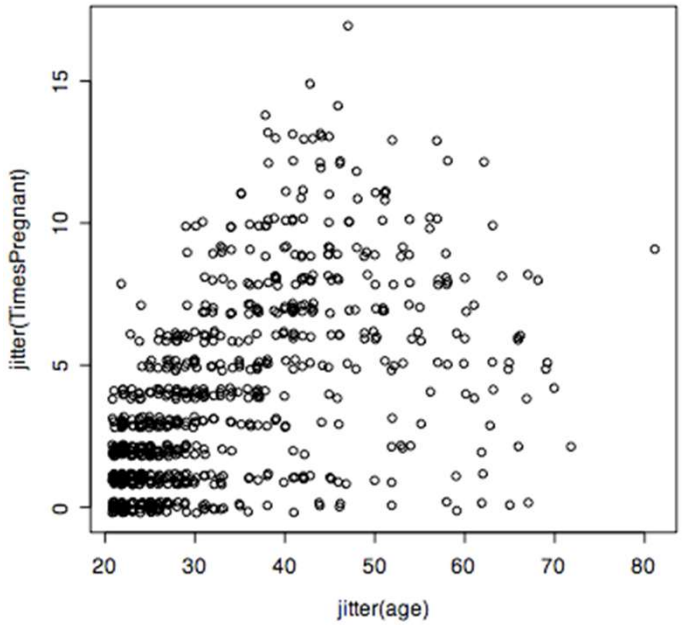
- Earthquake data:
  - locations of 1000 seismic events of MB > 4.0. The events occurred in a cube near Fiji since 1964
  - Data collected on the severity of the earthquake



Jittering

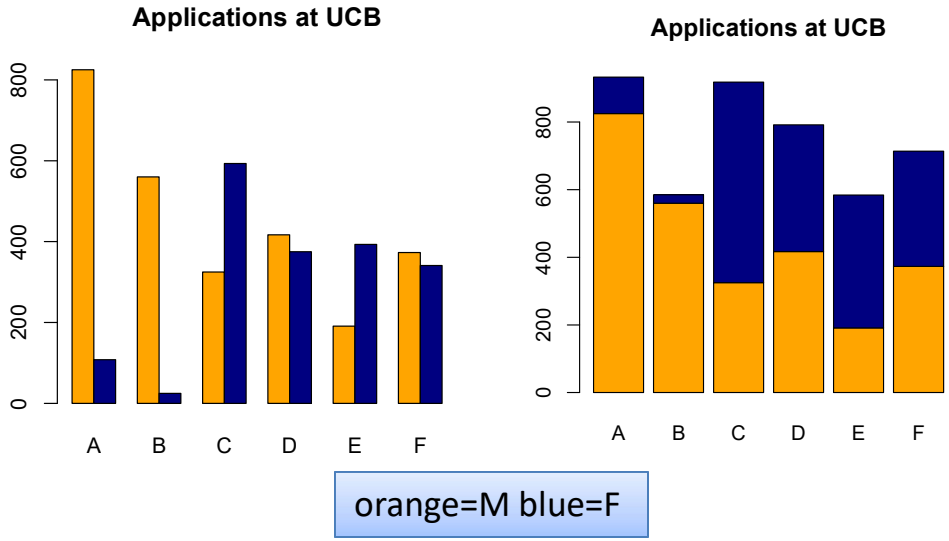
- Jittering points helps too



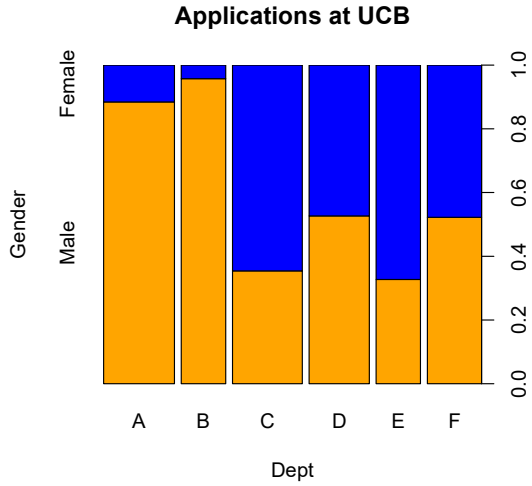


Barcharts and Spineplots

stacked barcharts can be used to compare continuous values across two or more categorical ones.

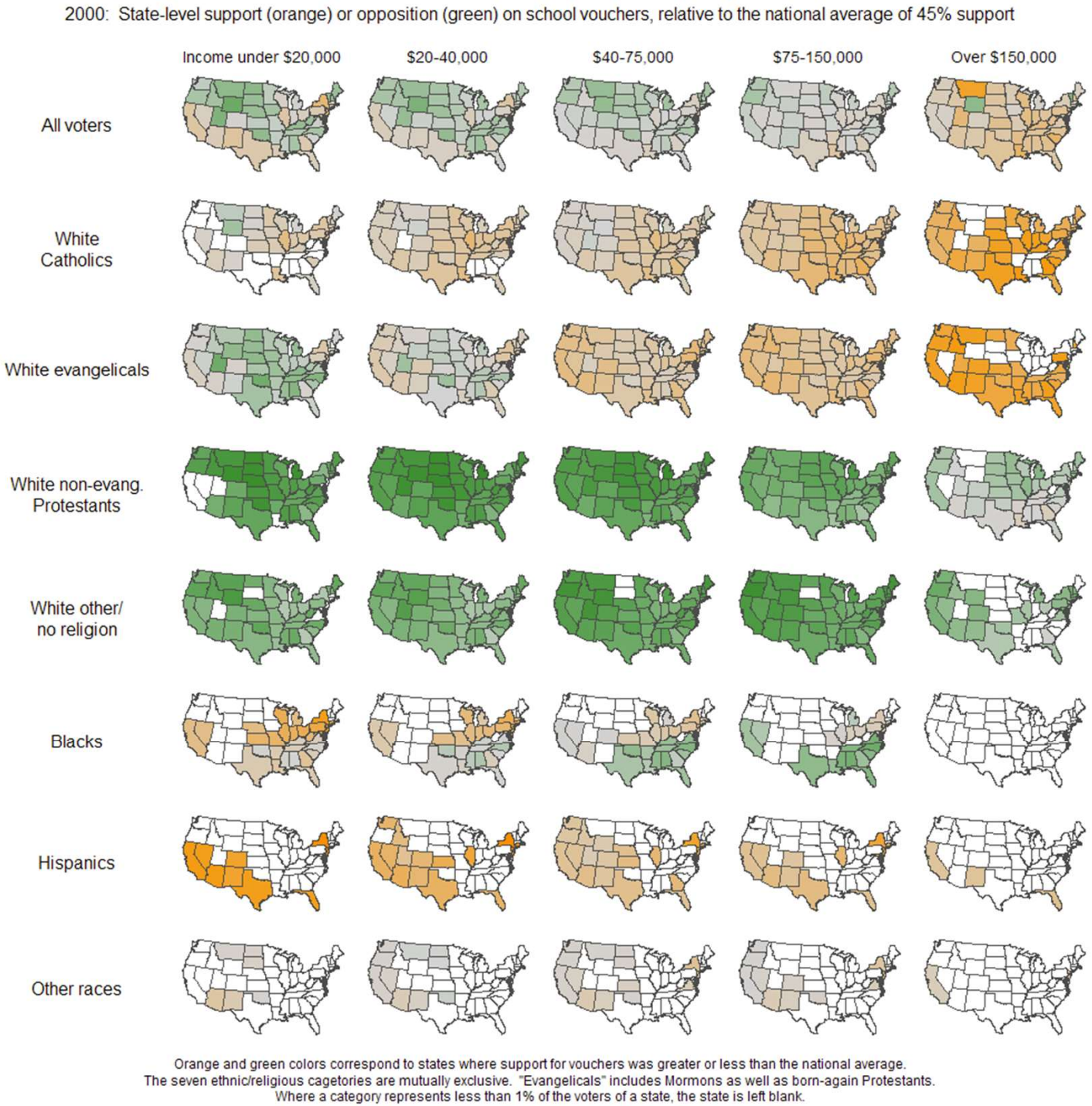


spineplots show proportions well, but can be hard to interpret



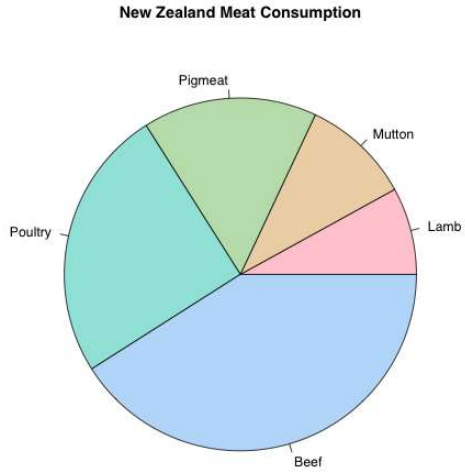
How many dimensions are represented here?

Andrew Gelman blog 7/1

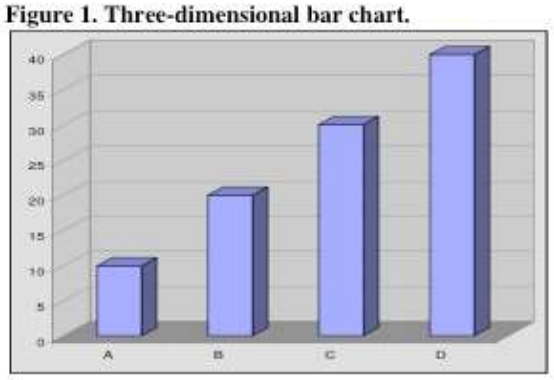


What's missing?

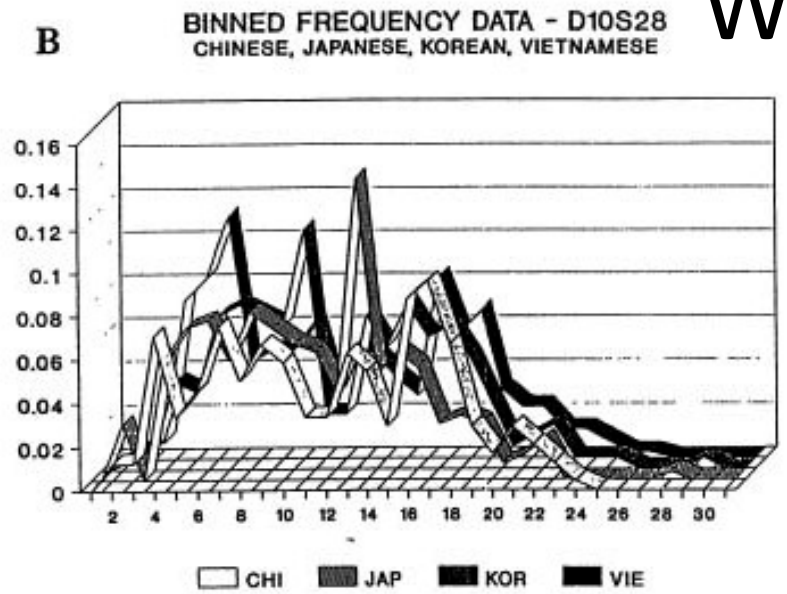
- pie charts
  - very popular
  - good for showing simple relations of proportions
  - Human perception not good at comparing arcs
  - barplots, histograms usually better (but less pretty)

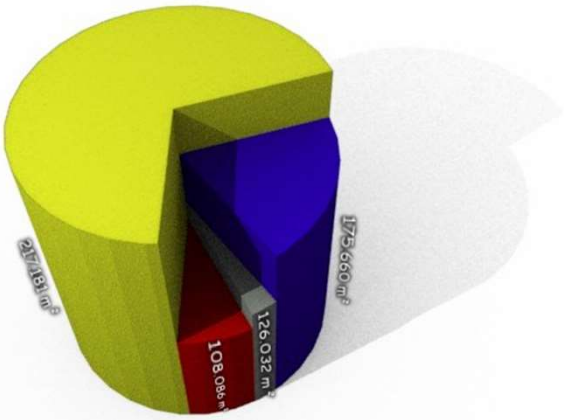


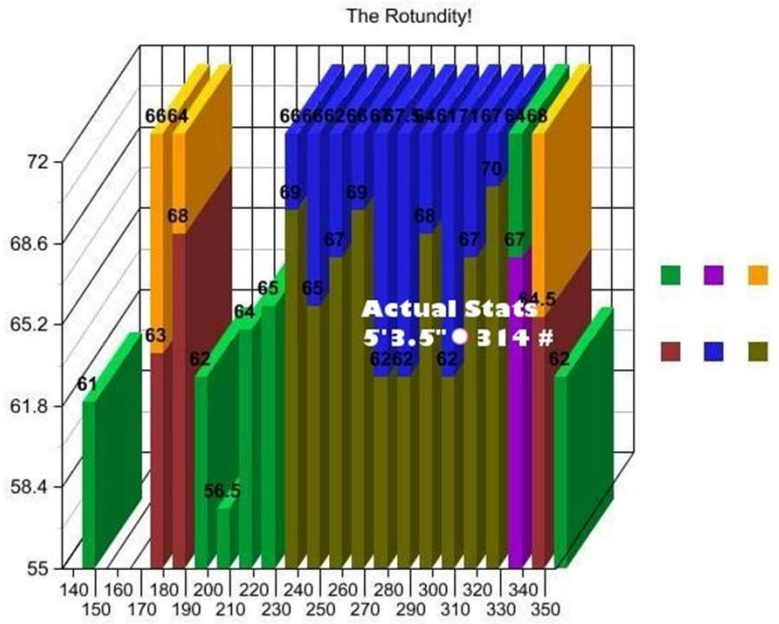
- 3D
  - nice to be able to show three dimensions
  - hard to do well
  - often done poorly
  - 3d best shown through "spinning" in 2D
    - uses various types of projecting into 2D
    - http://www.stat.tamu.edu/~west/bradley/

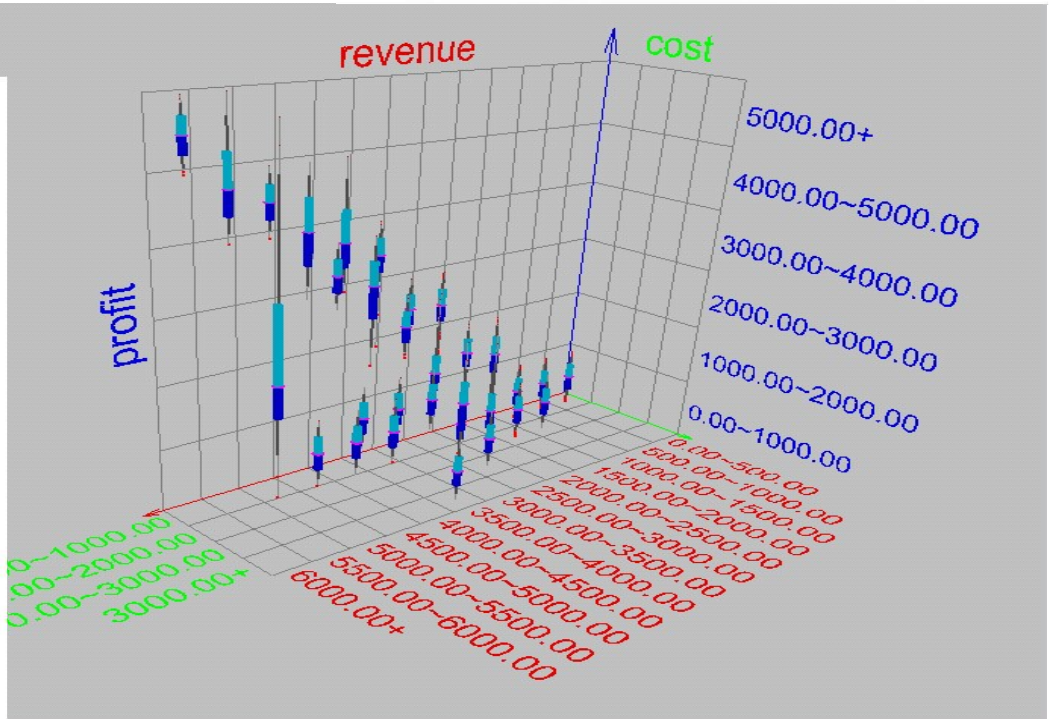


Worst graphic in the world?









Dimension Reduction

- One way to visualize high dimensional data is to reduce it to 2 or 3 dimensions
  - Variable selection
    - e.g. stepwise
  - Principle Components
    - find linear projection onto p-space with maximal variance
  - Multi-dimensional scaling, t-SNE
    - takes a matrix of (dis)similarities and embeds the points in p-dimensional space to retain those similarities

(More on this later)

Visualization done right

- Hans Rosling @ TED

- http://www.youtube.com/watch?v=jbkSRLYSojo