資料分析方法

Methods of Data Analysis

統計分析
Statistical Analysis

描述性統計（Descriptive Statistics）

- 或稱「敘述性統計」
- 集中量數：呈現資料集中的情形，如：(算術)平均、中位數、眾數等
- 變異量數：呈現資料分散的情形，如：全距（最大值 - 最小值）、標準差、四分位數等
  - DataFrame.describe()

	A	B	C	D
count	6.000000	6.000000	6.000000	6.000000
mean	0.473862	0.615370	0.568419	0.622193
std	0.252262	0.312380	0.164988	0.329959
min	0.080301	0.202910	0.317583	0.047279
25%	0.361322	0.365285	0.462811	0.507737
50%	0.474192	0.685850	0.660573	0.725291
75%	0.684415	0.851957	0.677213	0.809227
max	0.736302	0.951857	0.692137	0.962875

幾何平均數（Geometric Mean）

- 適用於計算比率數據的變化率

$$G = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

  - e.g. 營業額成長：12%, 15%, -4%, -10%, 6%

  - scipy.stats.gmean([1.12, 1.15, 0.96, 0.9, 1.06]) => 1.04 (4%)

調和平均數（Harmonic Mean）

- 數值倒數的算術平均數的倒數，又稱為「倒數平均數」。

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}$$

  - e.g. 台北到高雄坐高鐵平均時速300公里、高雄到台北坐台鐵普悠瑪號平均時速150公里，全程平均時速是多少？

  - scipy.stats.hmean([300, 150]) => 200

截尾平均數（Trimmed Mean）

- 平均數容易受到極端值影響
- 截尾平均數會將極端值去除後再取算術平均
  - 自訂上下限：scipy.stats.tmean(*array-like data*, (*lower limit, upper limit*))
    - 截尾後的標準差（tstd）、變異數（tvar）、最大值（tmax）、最小值（tmin）
  - 依比例去除：scipy.stats.trim_mean(*array-like data, proportiontocut*)

四分位數（Quartile）

- 將數據從小到大排列
  - 第一四分位數（Q1）：在1/4位置的數，又稱「較小四分位數」
  - 第二四分位數（Q2）：在1/2位置的數，又稱「中位數」
  - 第三四分位數（Q3）：在3/4位置的數，又稱「較大四分位數」
  - 四分位距（IQR）= Q3 - Q1
    - e.g. 1, 2, 3, 4, 5, 6, 7, 8
    - 內插法：Q1 = 2.75、Q2 = 4.5、Q3 = 6.25

Notes

四分位數的計算方法爭議

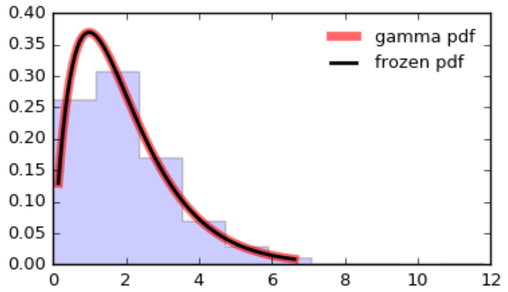
四分位數確切的數值計算方法仍具爭議，Scipy和Pandas計算出的值有少許誤差。

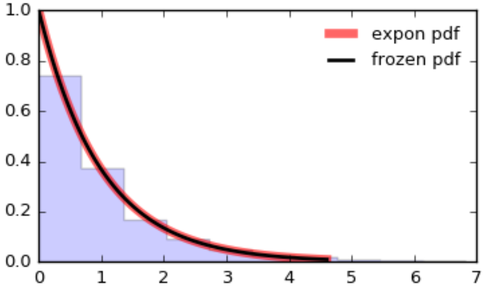
基本統計函式表

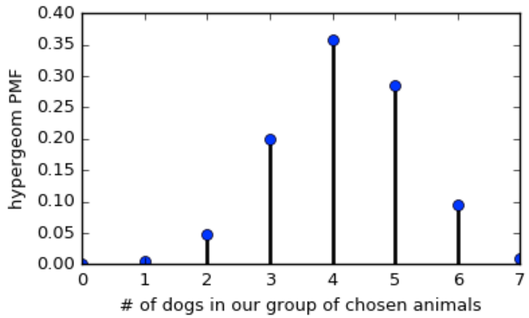
統計函式		Pandas DataFrame	Scipy.stats	Numpy
敘述統計		DataFrame.describe()	describe(*data*)	x
算術	平均數	DataFrame.mean()	x	mean(*data*)
幾何		x	gmean(*data*)	x
調和		x	hmean(*data*)	x
截尾		x	trim_mean(*data, proportiontocut*) tmean(*data, (lower limit, upper limit)*)	x
加權		x	x	average(*data, weights*)
截尾	最大值	x	tmax(data, (lower limit, upper limit))	x
	最小值	x	tmin(data, (lower limit, upper limit))	x
	標準差	x	tstd(data, (lower limit, upper limit))	x
	變異數	x	tvar(data, (lower limit, upper limit))	x
最大值		DataFrame.max()	x	max(*data*)
最小值		DataFrame.min()	x	min(*data*)
中位數		DataFrame.median()	x	median(*data)*
標準差		DataFrame.std()	x	std(*data)*
變異數		DataFrame.var()	x	var(*data*)
四分位數		DataFrame.quantile(*quantile*)	mstats.mquantiles(*data*)	x
眾數		DataFrame.mode() (0.19.1版)	mode(*data*) (0.18.1版)	x (1.11版)

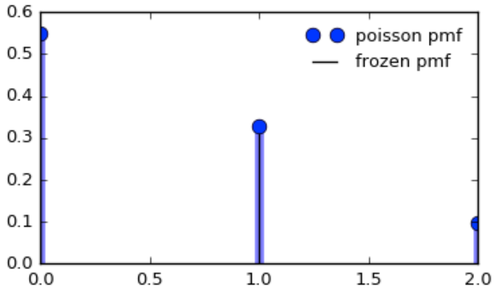
機率分佈

- 連續機率分佈（Continuous Distributions）：
  - 伽瑪分佈（Gamma Distribution）
  - 指數分佈（Exponential Distribution）
  - 常態分佈（Normal Distribution）
  - 均勻分佈（Uniform Distribution）
  - 卡方分佈（Chi-square Distribution）
- 間斷機率分佈（Discrete Distributions）：
  - 白努力分佈（Bernoulli Distribution）
  - 二項式分佈（Binomial Distribution）
  - 負二項式分佈（Negative Binomial Distribution）
  - 波式分佈（Poisson Distribution）
  - 超幾何分佈（Hypergeometric Distribution）









Scipy.stats

- Scipy 統計函式：https://docs.scipy.org/doc/scipy/reference/stats.html

Continuous distributions

alpha	An alpha continuous random variable.
anglit	An anglit continuous random variable.
arcsine	An arcsine continuous random variable.
beta	A beta continuous random variable.
betaprime	A beta prime continuous random variable.
bradford	A Bradford continuous random variable.
burr	A Burr (Type III) continuous random variable.
burr12	A Burr (Type XII) continuous random variable.
cauchy	A Cauchy continuous random variable.
chi	A chi continuous random variable.
chi2	A chi-squared continuous random variable.
cosine	A cosine continuous random variable.
dgamma	A double gamma continuous random variable.
dweibull	A double Weibull continuous random variable.
erlang	An Erlang continuous random variable.
expon	An exponential continuous random variable.

Discrete distributions

bernoulli	A Bernoulli discrete random variable.
binom	A binomial discrete random variable.
boltzmann	A Boltzmann (Truncated Discrete Exponential) random variable.
dlaplace	A Laplacian discrete random variable.
geom	A geometric discrete random variable.
hypergeom	A hypergeometric discrete random variable.
logser	A Logarithmic (Log-Series, Series) discrete random variable.
nbinom	A negative binomial discrete random variable.
planck	A Planck discrete exponential random variable.
poisson	A Poisson discrete random variable.
randint	A uniform discrete random variable.
skellam	A Skellam discrete random variable.
zipf	A Zipf discrete random variable.

其他常用統計函式

- One-way ANOVA
- F value
- T-test
- 相關性（Correlation）
- 峰度（kurtosis）
- 偏態（skewness）
- 共變數（Covariance）
- …

References

- Numpy 統計函式：https://docs.scipy.org/doc/numpy/reference/routines.statistics.html
- Scipy 統計函式：https://docs.scipy.org/doc/scipy/reference/stats.html
- Pandas DataFrame 統計函式：http://pandas.pydata.org/pandas-docs/stable/api.html#api-dataframe-stats

推薦學習資源

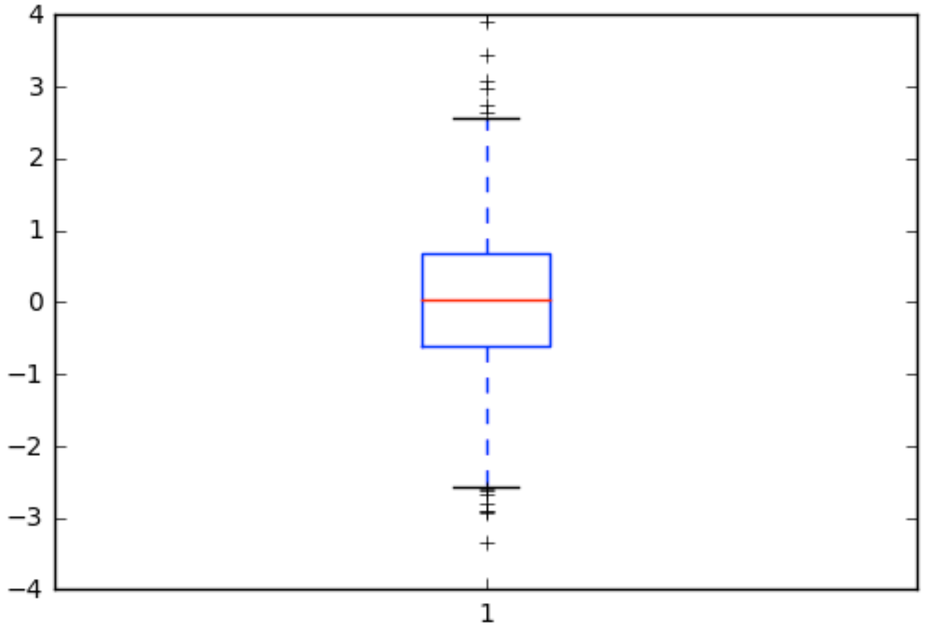
- 台大電機系 葉丙成教授 機率

- Youtube：https://www.youtube.com/watch?v=YGpKcJdrp5A&list=PLw9fh2FrjAqu1Gj_WznO-humCJT-OB2zF

異常値偵測

Anomaly Detection

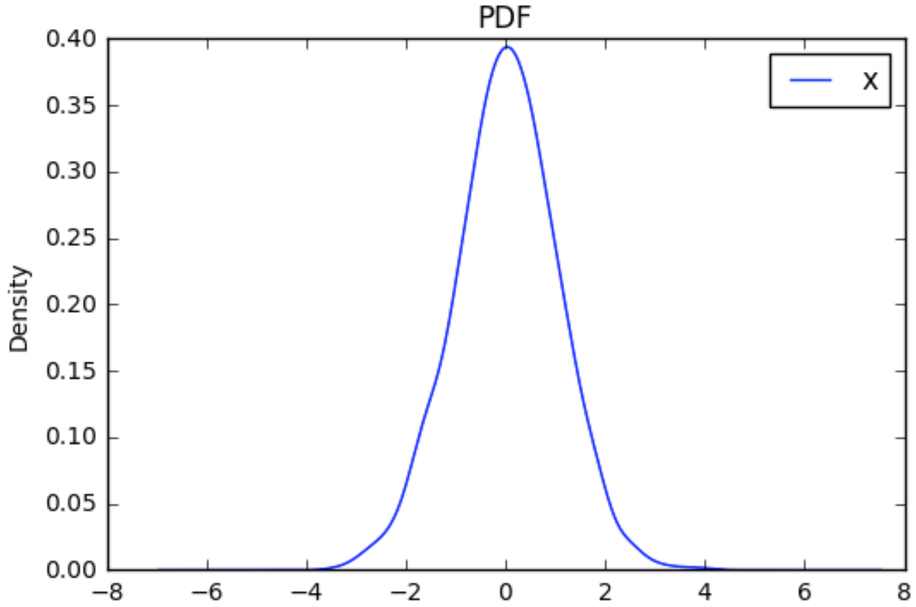
異常值偵測(1) - 四分位數與箱形圖



plt.boxplot(x, showfliers=True)

- 四分位間距(InterQuartile Range, IQR)
  - Q3 - Q1
- 最大值：Q3 + 1.5 * IQR
- 最小值：Q1 - 1.5 * IQR
- 異常值：高於最大值、低於最小值

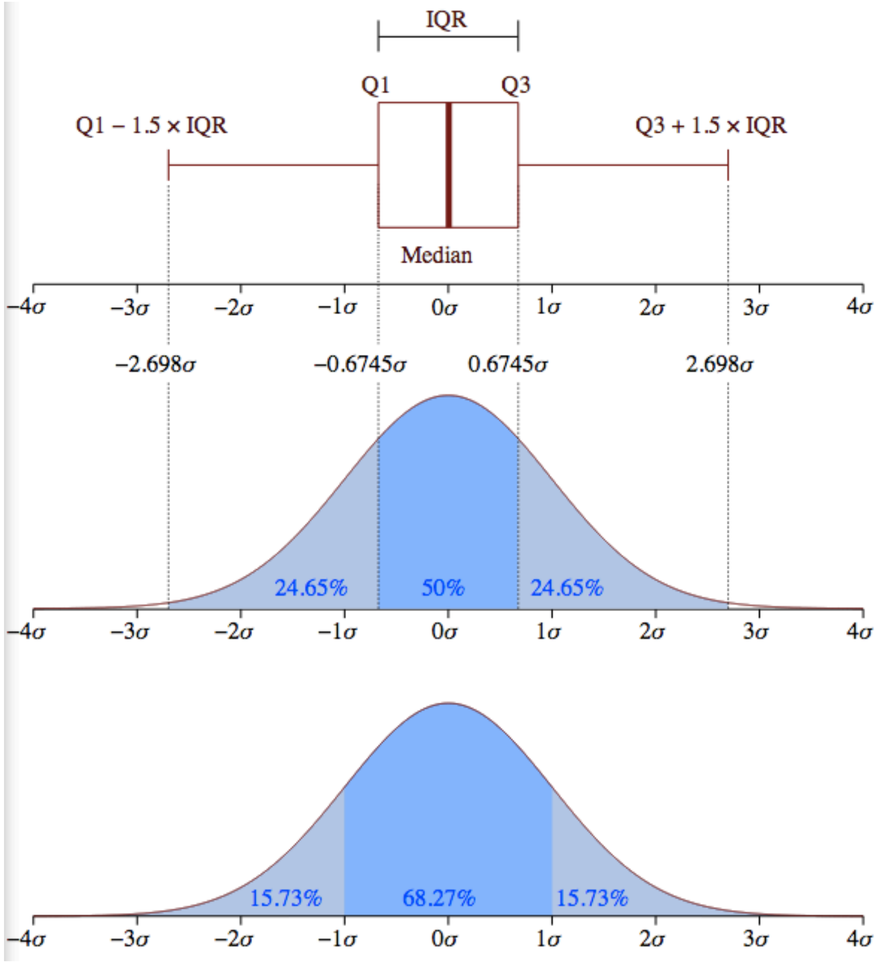
異常值偵測(2) - 常態分佈與標準差



df.plot(kind='kde')

- 平均（mean）、標準差（std, σ）
- 上限：mean + 3 * σ (sigma)
- 下限：mean - 3 * σ (sigma)
- 異常值：高於上限、低於下限

比較



- 若資料型態傾向常態分佈（例如：身高、體重、成績），適合使用標準差的判定方式
- 若異常值過大或過小，容易過度影響標準差，則建議使用四分位數和箱型圖，因為大於Q_3和小於Q_1的值不論離多遠都不會影響四分位數的值，所以在判定異常值效果好

相關性分析

Correlation Analysis

Pearson 相關係數

- 最常用的相關係數（-1~+1）

共變數

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

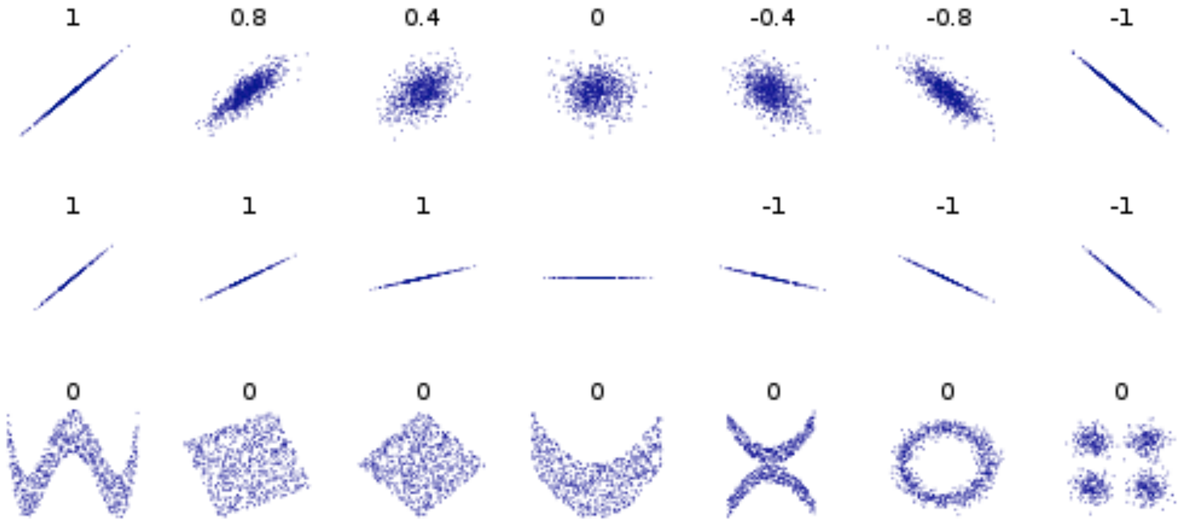
標準差

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

(Pearson, 1917)

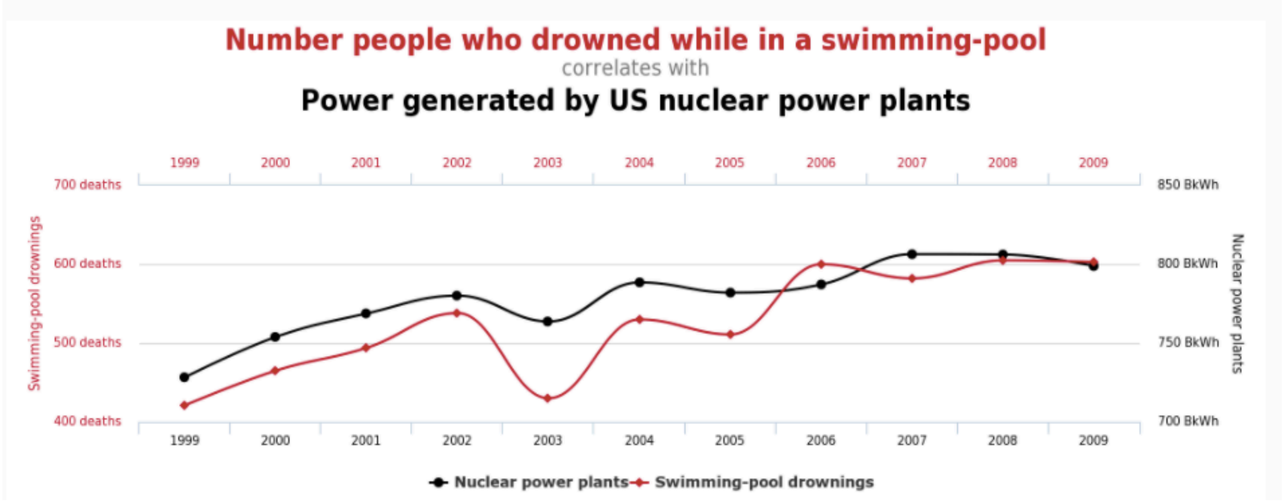
相關程度

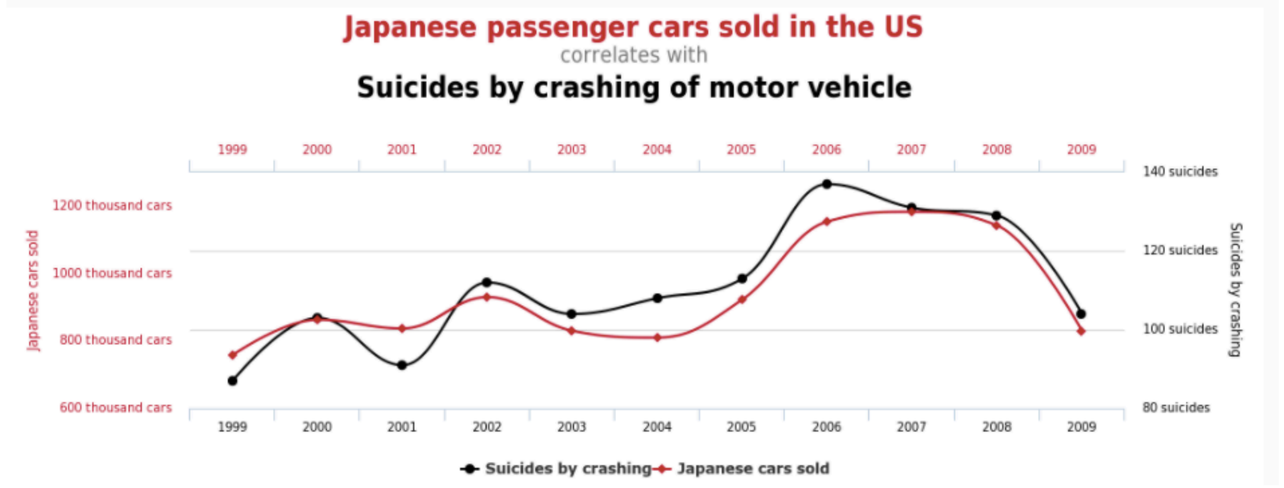
- 相關程度
  - +1：完全正相關
  - -1：完全負相關
  - 0.3至 -0.3：低度 (正/負)相關
  - (+-) 0.3至0.6：中度 (正/負)相關
  - (+-) 0.6至0.9：高度 (正/負)相關
    - DataFrame.corr()



(from wikipedia)

相關程度不等於有因果





(http://tylervigen.com/)

購物籃分析

Market Basket Analysis

購物籃分析（Market Basket Analysis）

- 針對消費行為中的購物籃項目做分析
- 項目（item）：購物項目
- 項目集（itemset）：購物籃項目組合
- 支持度（Support）：A 和 B 在所有購物籃紀錄中同時被購買的機率
  - $\text{Support}\ (A \Rightarrow B) = P(A \cap B)$
- 信賴度（Confidence）：A 被買的情況下 B也被買的機率
  - $\text{Confidence}\ (A \Rightarrow B) = P(B \mid A) = P(A \cap B) / P(A)$（條件機率）

Example

- **支持度 Support (A ⇒ B) = P(A ∩ B)**
  - Support (apple) = 4/8 = 0.5
  - Support (apple ⇒ beer) = 3/8 = 0.375
  - ➡ 有37.5%的機率會同時購買蘋果和啤酒
- **信賴度 Confidence (A ⇒ B) = P(B | A)**
  - Confidence (apple ⇒ beer)

  = P(apple ∩ beer) / P(apple)

  = (3/8) / (1/2)

  = 0.75

  ➡ 購買蘋果時，有75%的機率也會買啤酒

- e.g. 購物籃紀錄
  - Basket 1: apple, beer, rice, chicken
  - Basket 2: apple, beer, rice
  - Basket 3: apple, beer
  - Basket 4: apple, mango
  - Basket 5: milk, beer, rice, chicken
  - Basket 6: milk, beer, rice
  - Basket 7: milk, beer
  - Basket 8: milk, mango

Apriori 演算法 (1/3)

- 找尋頻繁項目集（Frequent Itemsets）和關聯規則（Association Rules）最常用的演算法

- Basket 1: apple, beer, rice, chicken
- Basket 2: apple, beer, rice
- Basket 3: apple, beer
- Basket 4: apple, mango
- Basket 5: milk, beer, rice, chicken
- Basket 6: milk, beer, rice
- Basket 7: milk, beer
- Basket 8: milk, mango

假設給定：
Support threshold = 0.3
Confidence threshold = 0.5

項目集(i=1)	支持度
{apple}	0.5
{beer}	0.75
{rice}	0.5
~~{chicken}~~	~~0.25~~
~~{mango}~~	~~0.25~~
{milk}	0.5

Apriori 演算法 (2/3)

項目集(i=1)	支持度
{apple}	0.5
{beer}	0.75
{rice}	0.5
~~{chicken}~~	~~0.25~~
~~{mango}~~	~~0.25~~
{milk}	0.5

項目集(i=2)	支持度
{apple, beer}	0.375
~~{apple, rice}~~	~~0.25~~
~~{apple, milk}~~	~~0~~
{beer, rice}	0.5
{beer, milk}	0.375
~~{rice, milk}~~	~~0.25~~

項目集(i=3)	支持度
~~{apple, beer, rice}~~	~~0.25~~
~~{apple, beer, milk}~~	~~0~~
~~{bear, rice, milk}~~	~~0.25~~

- Basket 1: apple, beer, rice, chicken
- Basket 2: apple, beer, rice
- Basket 3: apple, beer
- Basket 4: apple, mango
- Basket 5: milk, beer, rice, chicken
- Basket 6: milk, beer, rice
- Basket 7: milk, beer
- Basket 8: milk, mango

假設給定：
Support threshold = 0.3
Confidence threshold = 0.5

Apriori 演算法 (3/3)

最高itemsets(i=2)的頻繁項目集

頻繁項目集(i=2)	支持度
{apple, beer}	0.375
{beer, rice}	0.5
{beer, milk}	0.375

產生關聯規則

計算信賴度

關聯規則	支持度	信賴度
{apple} ⇒ {beer}	0.375	0.75
{beer} ⇒ {apple}	0.375	0.5
{beer} ⇒ {rice}	0.5	0.667
{rice} ⇒ {beer}	0.5	1
{beer} ⇒ {milk}	0.375	0.5
{milk} ⇒ {beer}	0.375	0.75

- 關聯規則產生方法：從每個itemsets裡的item做所有排列組合
  - e.g. 若最高頻繁項目集為 {A, B, C}
  - 則關聯規則為：{A ⇒ B}、{A ⇒ C}、{B ⇒ A}、{B ⇒ C}、{C ⇒ A}、{C ⇒ B}、{A, B ⇒ C}、{A, C ⇒ B}、{B, C ⇒ A}、{A ⇒ B, C}、{B ⇒ A, C}、{C ⇒ A, B}

實作

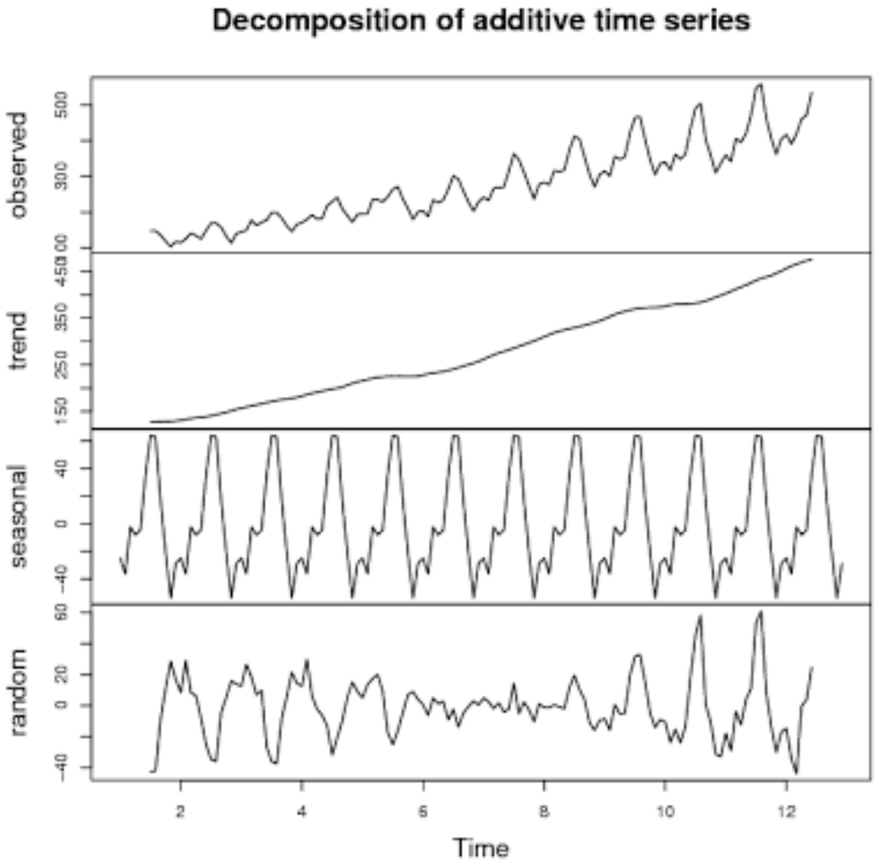
- Apriori.py on Github（原始碼已附於範例程式.zip中，並已修改為Python 3版本）
  - https://github.com/asaini/Apriori
- 建議值
  - Support：0.1~0.2
  - Confidence：0.5~0.7

時間序列處理與週期性分析

Time Series Processing & Analysis

時間序列分析（Time Series Analysis）

- 主要目的：
  - 分析現象
  - 預測未來
- 時間序列分解：
  - 季節性（Seasonality）
  - 不規則（Irregular / Random）
  - 趨勢（Trend）



Additive Decomposition
Observed series =
Trend + Seasonal + Irregular

$$O_t = T_t + S_t + I_t$$

時間序列分析有更多複雜的模型和分析方法...

(Australian Bureau of Statistics, 2005; Yanchang Z., 2015)

Python 時間資料型態

- time, calendar
- datetime
  - datetime.date : (year, month, day)
  - datetime.time : (hour, minute, second, microsecond)
  - datetime.datetime: (year, month, day, hour, minute, second, microsecond)
  - datetime.timedelta: (days, seconds, microseconds)
- Documents : https://docs.python.org/3/library/datetime.html

String & Datetime 轉換

- string to datetime: datetime.strptime(*str, format*)
- datetime to string: datetime.strftime(*datetime, format*)

Directive	Meaning	Example
%a	Weekday as locale's abbreviated name.	So, Mo, ..., Sa (de_DE)
%A	Weekday as locale's full name.	Sonntag, Montag, ..., Samstag (de_DE)
%w	Weekday as a decimal number, where 0 is Sunday and 6 is Saturday.	0, 1, ..., 6
%d	Day of the month as a zero-padded decimal number.	01, 02, ..., 31
%b	Month as locale's abbreviated name.	Jan, Feb, ..., Dez (de_DE)
%B	Month as locale's full name.	Januar, Februar, ..., Dezember (de_DE)
%m	Month as a zero-padded decimal number.	01, 02, ..., 12
%Y	Year with century as a decimal number.	0001, 0002, ..., 2013, 2014, ..., 9999
%H	Hour (24-hour clock) as a zero-padded decimal number.	00, 01, ..., 23
%I	Hour (12-hour clock) as a zero-padded decimal number.	01, 02, ..., 12
%p	Locale's equivalent of either AM or PM.	am, pm (de_DE)
%M	Minute as a zero-padded decimal number.	00, 01, ..., 59
%S	Second as a zero-padded decimal number.	00, 01, ..., 59
%f	Microsecond as a decimal number, zero-padded on the left.	000000, 000001, ..., 999999

- 節錄自：https://docs.python.org/3/library/datetime.html

Pandas DatetimeIndex

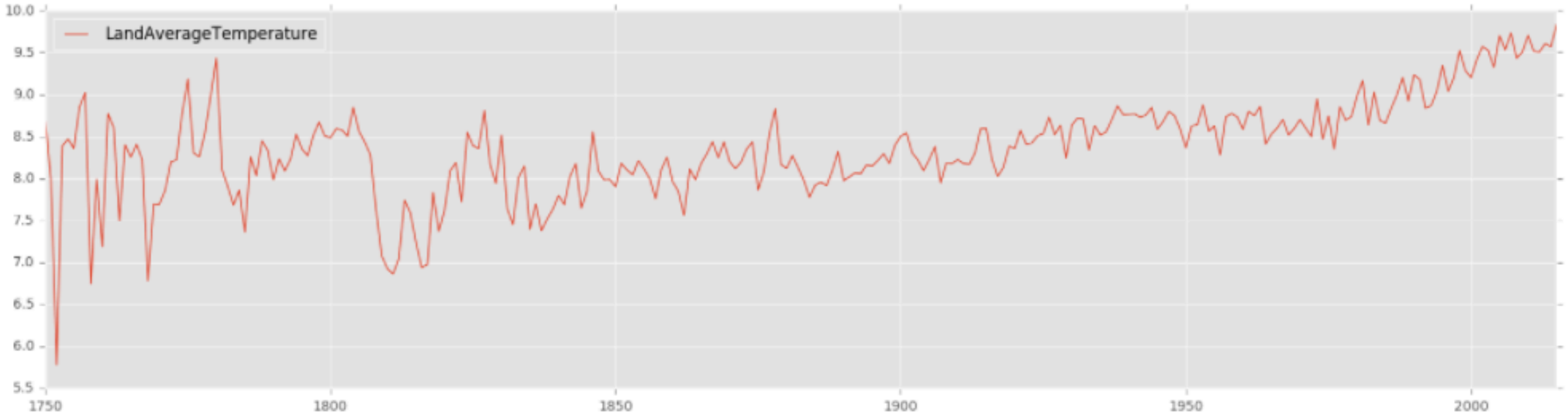
- 把DataFrame的index轉為DatetimeIndex型態
  - e.g df.index = pd.to_datetime(df.index,format='%Y-%m-%d')
- DatetimeIndex 時間分割/聚合
  - e.g. DataFrame.groupby([*columns*]).agg_func()
  - e.g. DataFrame.resample('M').agg_func()
  - e.g. DataFrame.resample('Q-NOV').agg_func() #Q-EndMonth

M	1	2	3	4	5	6	7	8	9	10	11	12
Q-DEC	Q1			Q2			Q3			Q4		
Q-NOV	Q1			Q2			Q3			Q4		Q1

- Document（DatetimeIndex的attributes、methods）: http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DatetimeIndex.html

範例：全球地表溫度變化

- 資料集：Climate Change: Earth Surface Temperature Data. https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data



References

- Australian Bureau of Statistics. (2005). Time Series Analysis: The Basics. Retrieved from: http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Time+Series+Analysis:+The+Basics#WHAT%20IS%20AN%20IRREGULAR%3F

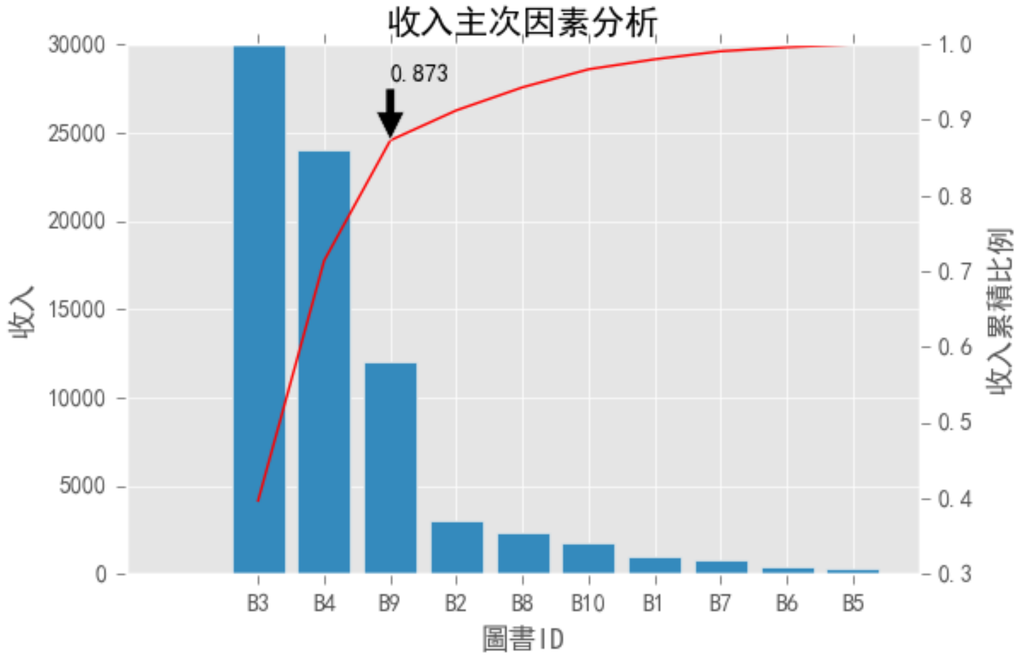
- Yanchang Z. (2015). R and Data Mining: Examples and Case Studies. Retrieved from: http://www.rdatamining.com/docs

主次因素分析

Pareto Analysis

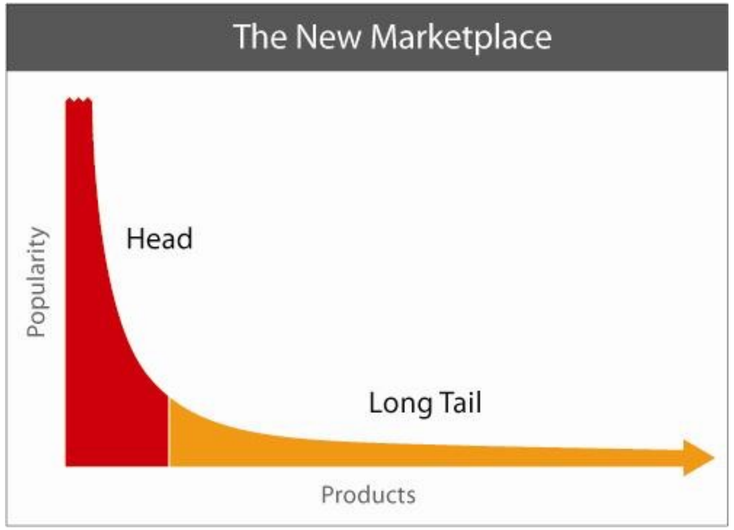
80/20 法則

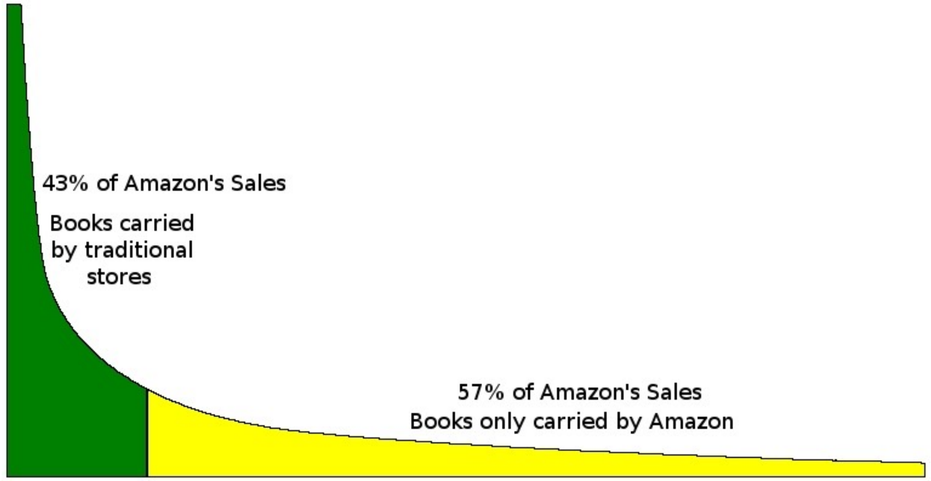
- 又稱為「Pareto 法則」
- 義大利經濟學家Pareto(1906) 觀察到，當時義大利20%的人擁有80%的財富
- 80%（多數）的結果來自於20%（少數）的原因
  - e.g. 80%的收益來自20%的產品
- Pareto Chart (柏拉圖、主次因素分析)



長尾理論（Long-tail）

- Chris Anderson (2004)：非熱門的商品(80%)加起來的總銷量巨大，超過熱門商品(20%)
- 網路上的情況（如電子商務）更為明顯
- 改變80/20法則的思維
  - e.g. Amazon 57%的銷售來自非熱門商品





References

- Chris A. (October, 2004) The Long Tail. Wired. Retrieved from: https://www.wired.com/2004/10/tail/

- Chris A. The Long Tail, in a nutshell. Retrieved from: http://www.thelongtail.com/about.html

- Neil, P. (2015, December 22). 7 Brilliant Examples of Brands Driving Long-Tail Organic Traffic. Retrieved from: http://neilpatel.com/2015/12/22/7-brilliant-examples-of-brands-driving-long-tail-organic-traffic/