R 語言和商業分析 -
洞悉商業世界中的資料科學

# 探索性資料分析與資料視覺化



黃大維　國立台灣大學統計碩士
沈辰禧　國立台灣大學商研碩士

洞悉商業世界中的資料科學

# 課程大綱

探索性資料分析
與資料視覺化

跨出資料分析的第一步

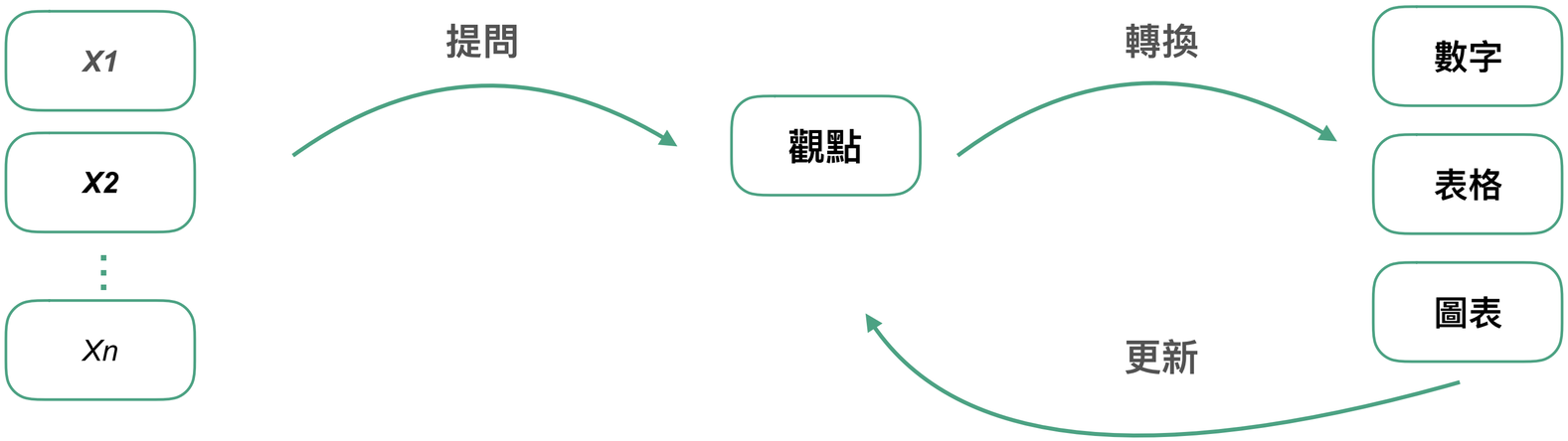
思考架構與解決方案

**探索性資料分析的流程**

資料視覺化的要素與方法

R 實戰：銷售管理視覺化

# 一個提問、轉換、再提問的流程




1. 探索性資料分析(Exploratory Data Analysis, EDA)是一個觀念，並非一套恆常不變的流程
2. 建立一套初步審視、診斷資料的方法，幫助我們快速瞭解一組資料的分配與相關性
3. 透過建立觀點和提問的過程，提出適當的假說幫助日後建立更進階的分析

# 區分連續變數與類別變數

| | 連續變數 | 類別變數 |
|---|---|---|
| 資料 | 資料的值(value)：<br>整數(integer)<br>數值(numeric) | 資料的值(value)：<br>字元(character)<br>類別(factor) |
| 說明 | 通常是一連串的數列，可是在特定或不特定的範圍內出現的任何數字 | 通常用於分組、分類、編號等字串，數字和文字僅具區別的功能，沒有數學上的差異 |
| 範例 | 1. 每日的銷售金額(integer)<br>2. 商品的銷售量(integer)<br>3. 商品的毛利率(numeric) | 1. 不同的日別、月別<br>2. 商品的品項 |

# 快速審視資料的分配與相關性

| | 分配 Distribution | 相關性 Covariance |
|---|---|---|
| 用途 | 探討單一連續變數在各個值的分佈，也可以比較連續變數在不同類別變數中的分佈 | 檢視兩組變數(不限變數類型)的關聯程度 |
| 方法 | 1. 計算機率密度<br>2. 計算次數 | 1. 僅用圖表呈現<br>2. 計算相關係數 |
| 範例 | 1. 直方圖(histogram) or 盒鬚圖(boxplot)<br>2. 長條圖(bar chart) | 1. 連續對連續：散佈圖(scatter plot)<br>2. 連續對類別：多個boxplot<br>3. 類別對類別：熱密度圖 |

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

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma_X}\right)\left(\frac{Y_i - \overline{Y}}{\sigma_Y}\right)$$

探索性資料分析的流程

# 建立標準化的流程，探討資料中隱含的模式

## 檢視資料

- 資料型態
- 變數的summary(見R語言特訓班部份)

## 辨識模式

利用簡單的圖表，探討：

- 重點趨勢
- 分配、相關性、異常值等

## 觀點與假說

- 提出假說
- 引導下一步分析

探索性資料分析的流程

# 建立標準化的流程，探討資料中隱含的模式

## 檢視資料

- 資料型態
- 變數的summary(見R語言特訓班部份)

## 辨識模式

利用簡單的圖表，探討：

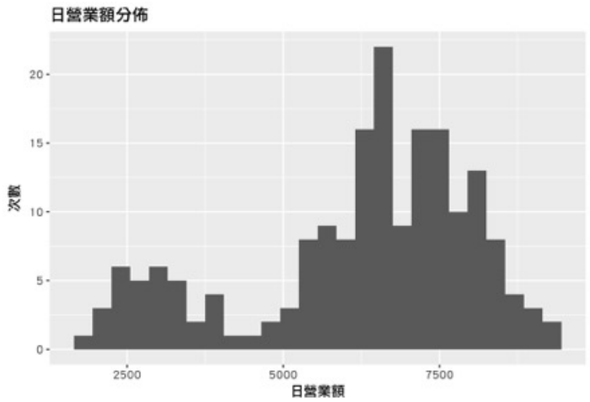
- 重點趨勢
- 分配、相關性、異常值等

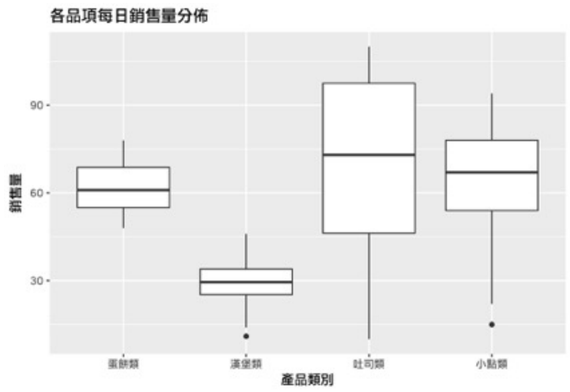
## 觀點與假說

- 提出假說
- 引導下一步分析

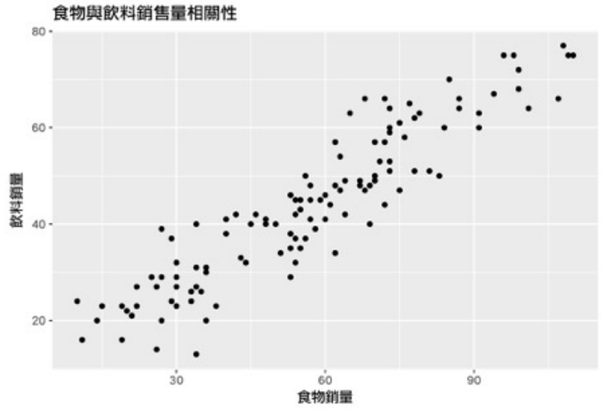
# 呈現分配與相關性的圖表

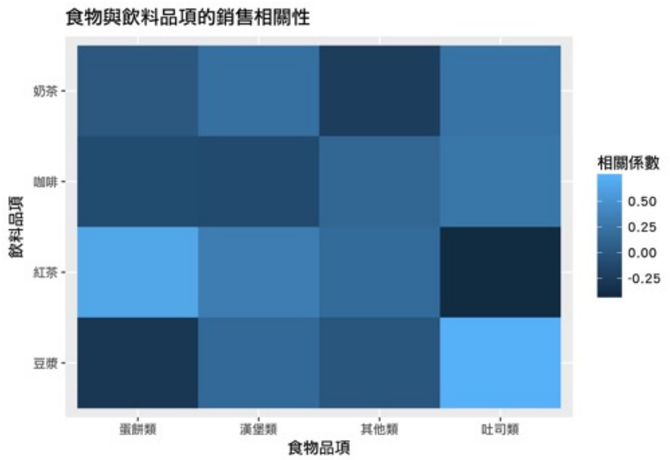
## 分配 Distribution



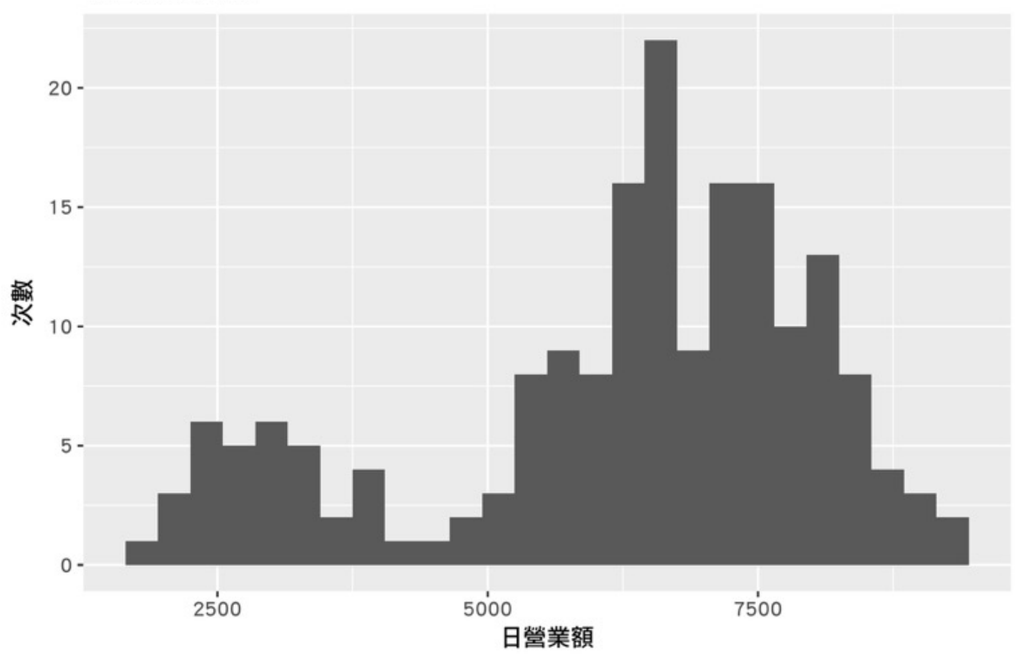

## 相關性 Covariance



# 分配：利用直方圖探討變數在數值間的次數分配

## 早餐店的日營業額分配



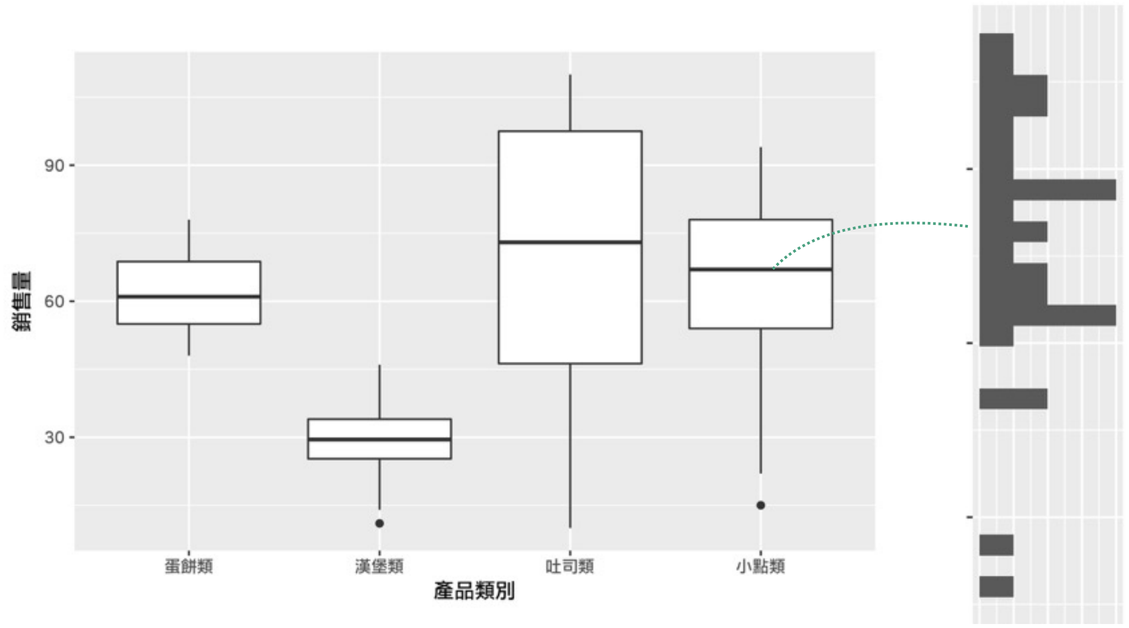

**直方圖 Histogram**

X 軸：每日營業額
Y 軸：次數，出現了幾次

了解營業額的分佈情形，有無集中分佈在那些值、哪些值出現的頻率最高、最低等等

# 分配：利用盒鬚圖比較各品項的銷售量分配

## 早餐店的產品銷售量分配




小點類銷量分配

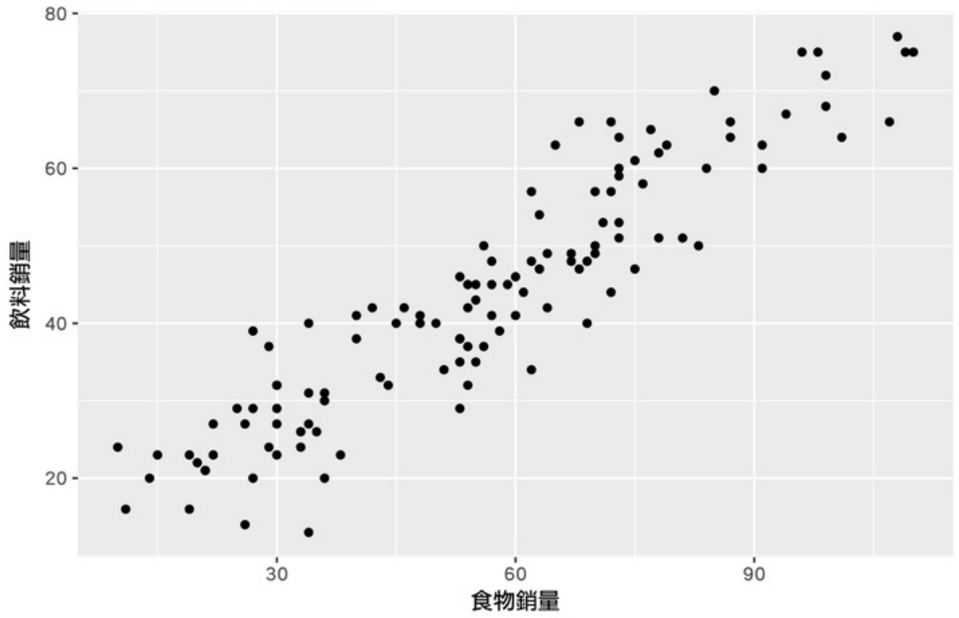
## 盒鬚圖 Boxplot

X 軸：產品類別
Y 軸：銷售量

了解不同產品類別間，銷售量分配會有何不同；同時也能探討銷售量和產品類別的相關性

# 相關性：利用散佈圖探討兩個連續變數的關聯

## 早餐店的食物和飲量銷售相關性



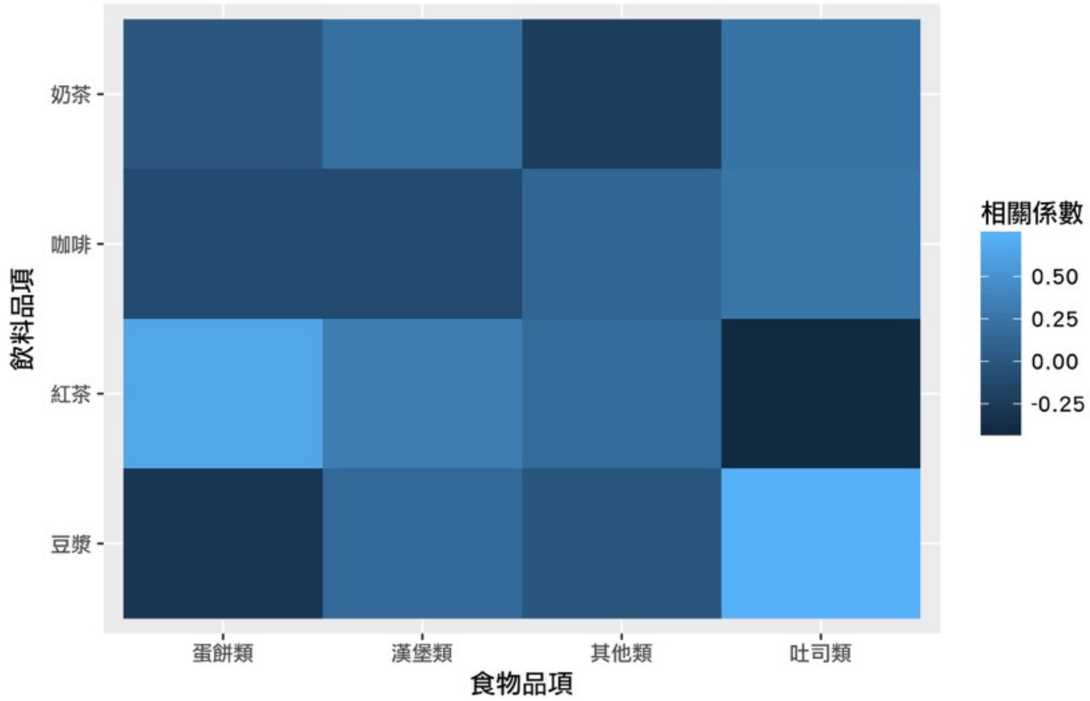

## 散佈圖 Scatter Plot

X 軸：食物銷量
Y 軸：飲料銷量

了解在不同食物和飲料兩種類別中，銷售量有沒有明顯的相關性

# 相關性：熱密度圖探討相關係數

## 早餐店的食物和飲量銷售相關係數



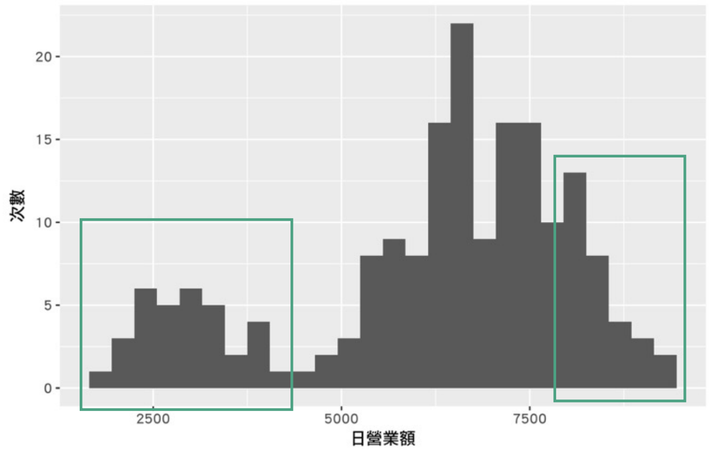

## 散佈圖 Scatter Plot

X 軸：食物品項
Y 軸：飲料品項

計算兩種品項的不同類別之間的相關係數，具體討論彼此的相關性；也可以探討同品項間不同類別的相關係數

# 觀察一：從觀察營業額分配提出增加收入的可能方法

## 早餐店的日營業額分配




## 辨識模式

**模式一**

日營業額明顯有某些日子表現特別差，審視原始資料後發現這些時間多落在週六和週日

**模式二**

日營業額也出現少數特別突出的樣本，比對資料後沒有特別標注原因

## 觀點與假說

**假說一**

週六週日比早起吃一般早餐的人較少，可能可以推出早午餐或其他輕食選擇

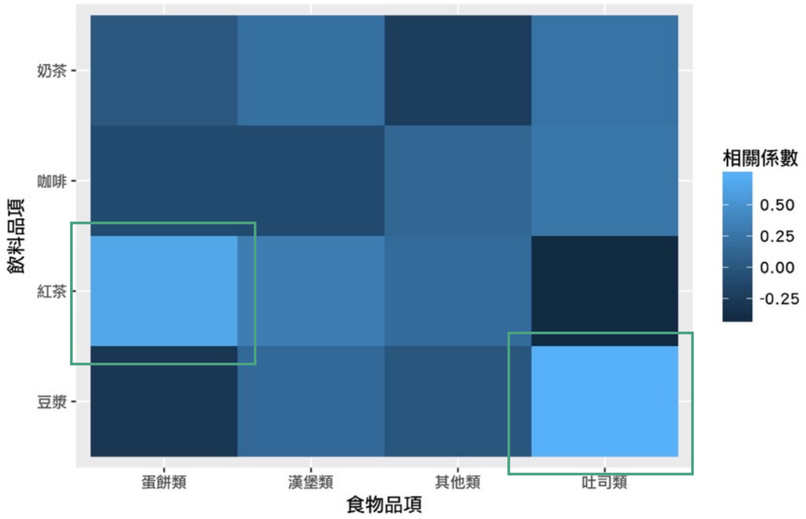
**假說二**

營業額特別高，也許和附近社區、國小的活動有關，找出這些日期能提前準備促銷或合作

探索性資料分析的流程

# 觀察二：從觀察產品相關性提出增加收入的方法

## 早餐店的食物和飲量銷售相關係數




## 辨識模式

**模式一**

吐司類的銷量和豆漿有較高的相關係數

**模式二**

蛋餅和紅茶同樣存在較高的相關係數

## 觀點與假說

**假說一**

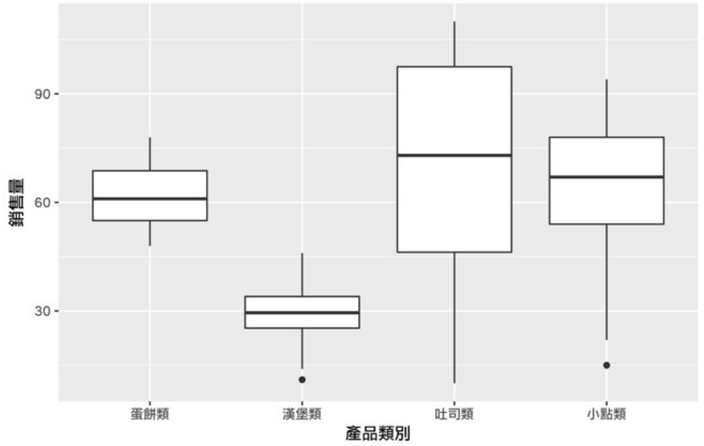
可以推出吐司和豆漿的套餐，促進銷售量

**假說二二**

也可以推出蛋餅和紅茶的套餐，促進銷售量

# 觀察三：從觀察產品銷售分配提出的降低成本的可能方法

## 早餐店的產品銷售量分配




## 辨識模式

**模式一**

漢堡的銷售量的中位數明顯較低，也最為集中

**模式二**

計算成本資料後，發現漢堡的毛利較低

## 觀點與假說

**假說**

由於漢堡的銷量已經較差、毛利又比較低，可以減少在漢堡上的促銷，或是找到成本更低的原料，提升整體銷售毛利率

探索性資料分析的流程

# 提出假說的意義：為進一步的分析提供方向

## 檢視資料

- 資料型態
- 變數的summary(見R語言特訓班部份)

## 辨識模式

利用簡單的圖表，探討：

- 重點趨勢
- 分配、相關性、異常值等

## 觀點與假說

- 提出假說
- 引導下一步分析

## 進一步的分析

驗證假說

- 用實驗設計驗證套餐是否增加銷售
- 用迴歸分析預測每日營業額
- 用集群分析瞭解顧客的喜好

**若沒有進行EDA，直接建模甚至執行商業策略，不僅會沒有效果，更可能帶來難以預期的傷害，之後在各個演算法的章節皆會說明這些錯誤解讀資料後，使用演算法的案例**