**分散式資料庫期末報告**

**OLAP分析葡萄牙學生成績**

|  |  |
| --- | --- |
|  | **指導教授：蔡正發教授**  **組員：B10756014吳湄潔**  **B10756055王諒桓** |

**目錄**

**一、 資料背景2頁**

**二、 資料集分析簡介3頁**

**三、 資料整理4頁**

**四、 OLAP分析7頁**

1. **切片(slice) 8-11頁**
2. **切丁(Dice) 12-14頁**
3. **向上截取(Roll-Up) 15頁**
4. **向下擷取(Drill-Down) 15頁**
5. **轉軸(Pivot) 16頁**

**五、 分析結論17-18頁**

1. **資料背景：**

* **來源：Kaggle資料科學平台**[**https://www.kaggle.com/uciml/student-alcohol-consumption**](https://www.kaggle.com/uciml/student-alcohol-consumption)
* **資料名稱：Student Alcohol Consumption**
* **分析項目：學生葡萄牙語成績分析**
* **更新資料集：最後更新2016年**
* **各表單的資料量：**

**學生資料：649筆 學校：2筆**

**性別：2筆 住址：2筆**

**補習:2筆 父教育程度:5筆**

**讀書時間：4筆 母教育程度：5筆**

**課外活動：2筆 在家是否上網：2筆**

**缺勤紀錄：3筆 家庭關係：5筆**

**自由時間：5筆 與朋友外出頻率：5筆**

**健康狀況：5筆 有男/女朋友：2筆**

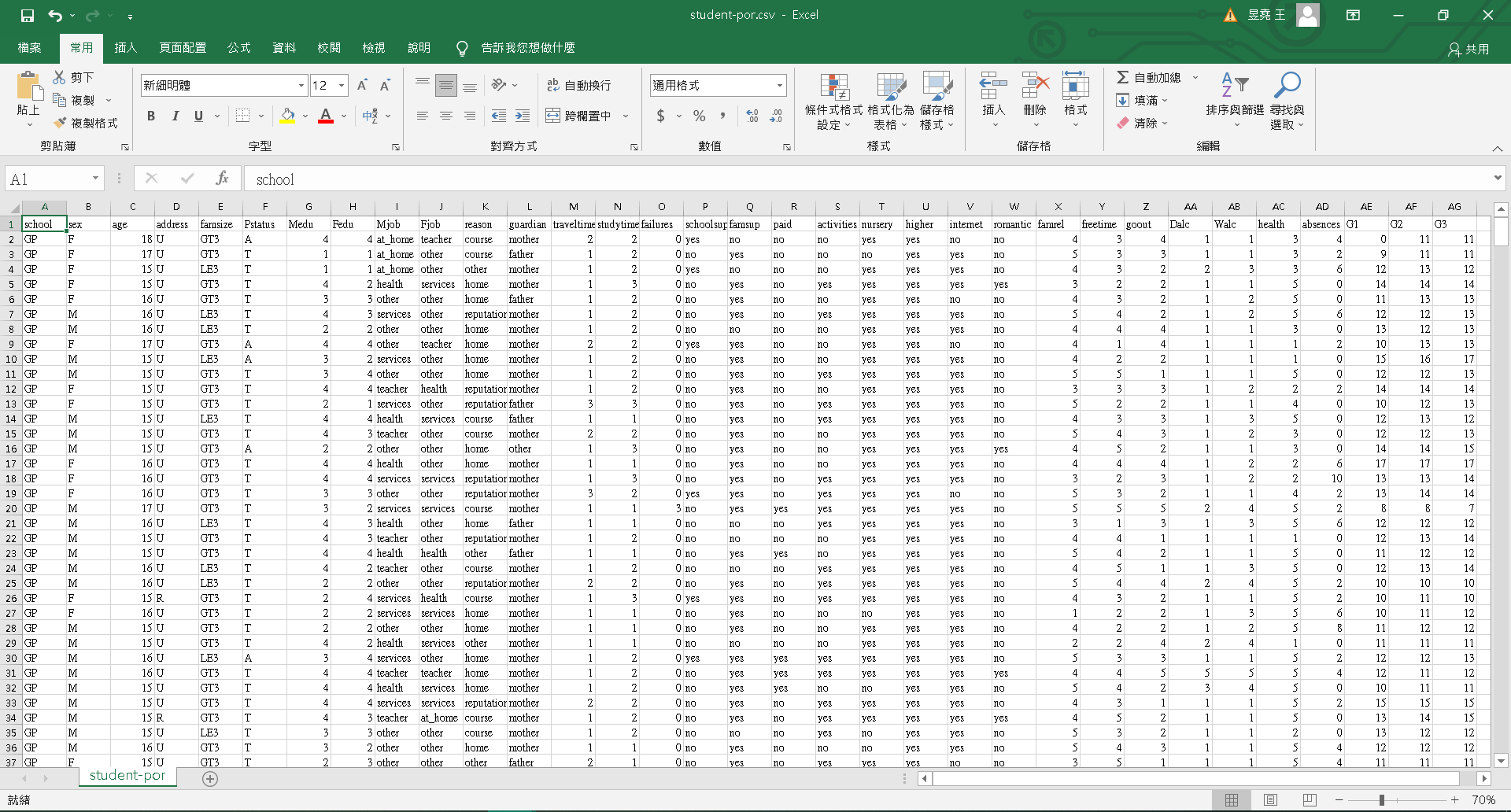
**及格：2筆**

1. **資料集分析簡介**

**原版Kaggle資料集作者原本是想研究學生飲酒習慣，會不會導致葡萄牙語成績的影響，因為葡萄牙法定年齡為16歲即可買酒，但在台灣學生必須達到18歲，所以我們決定稍微修改資料內容，將飲酒習慣欄位去除，改用其他欄位資料來分析是否影響學生成績的因素。**

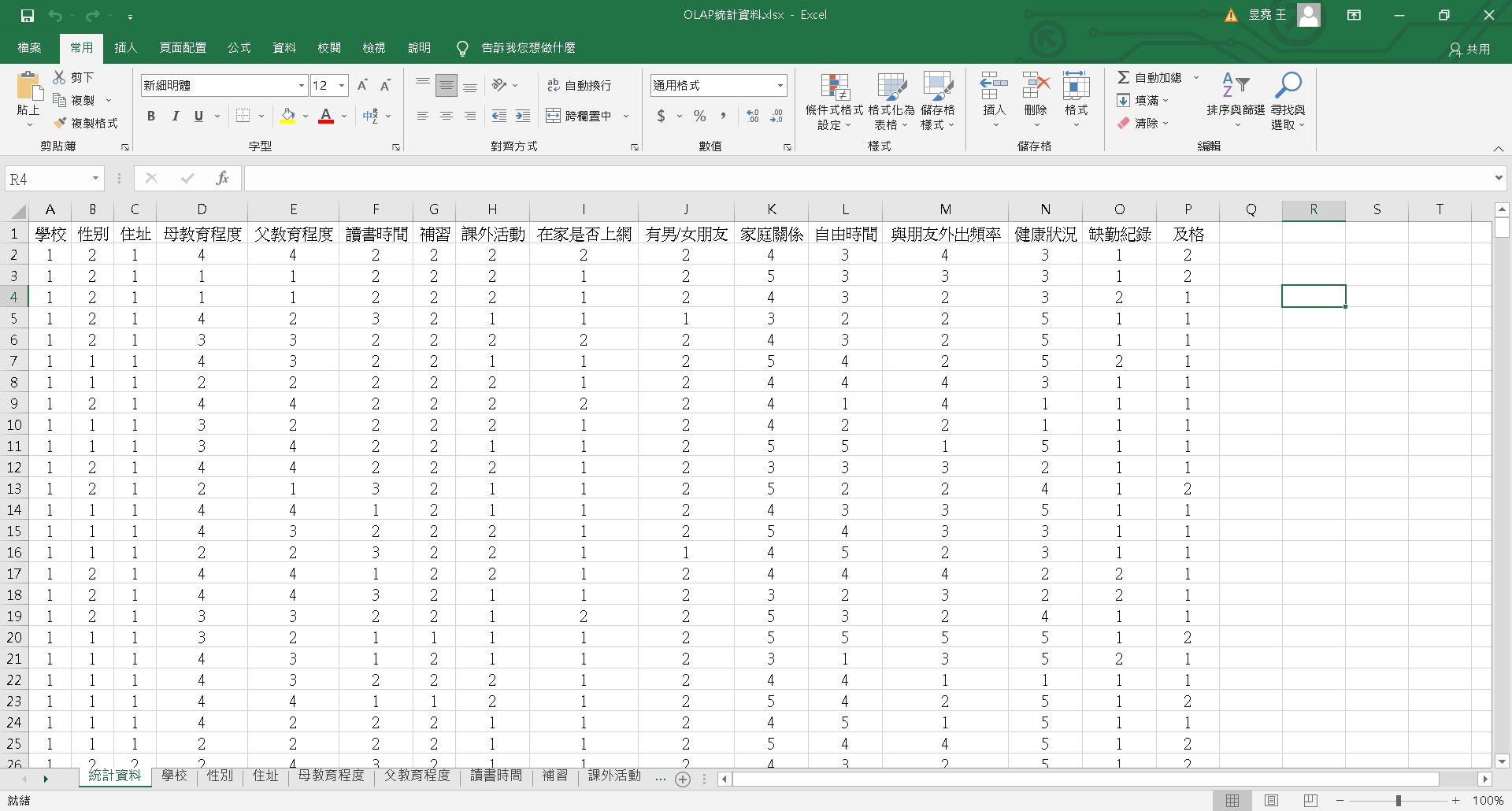
**三、 資料整理**

**原始的資料**

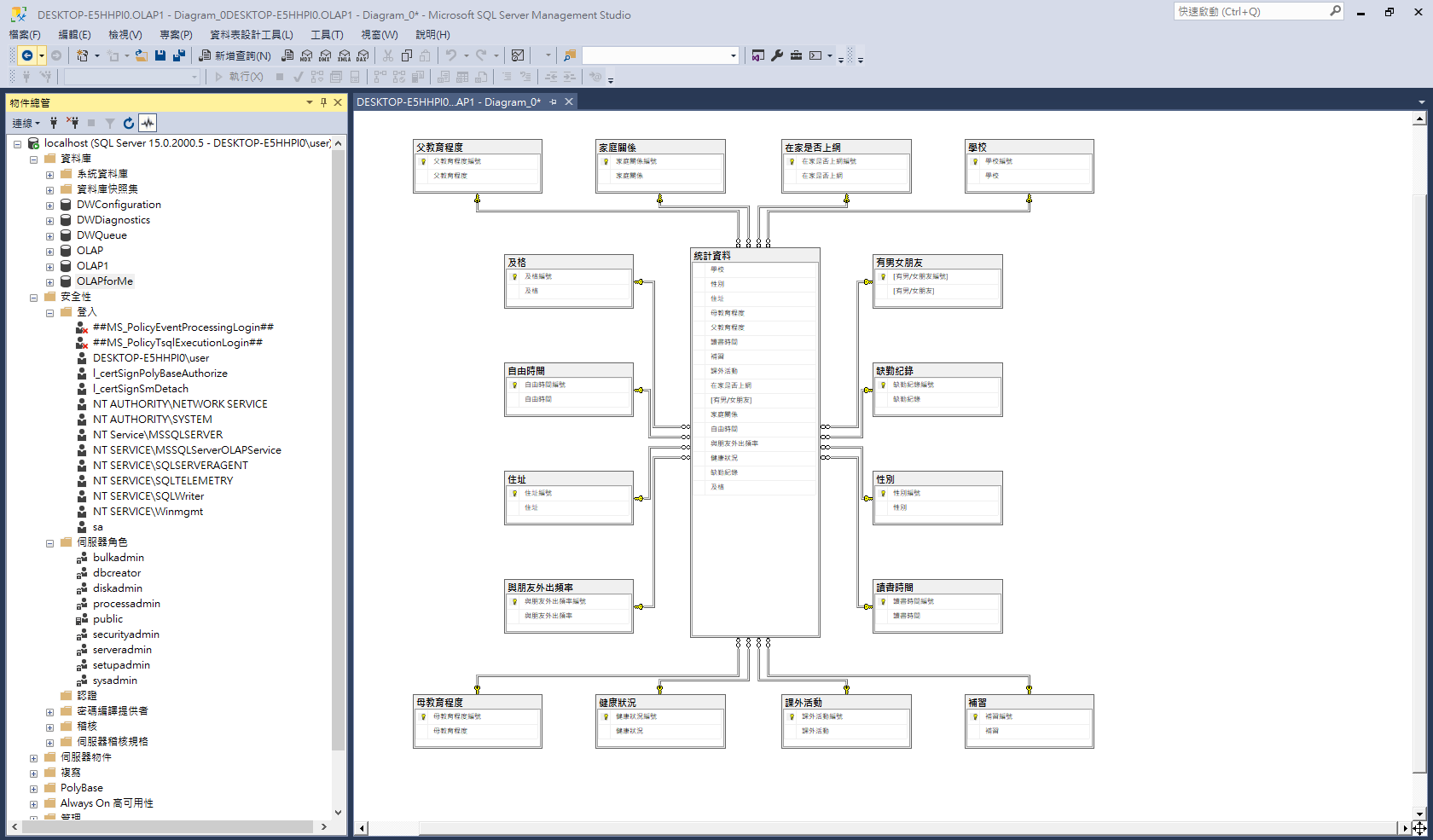


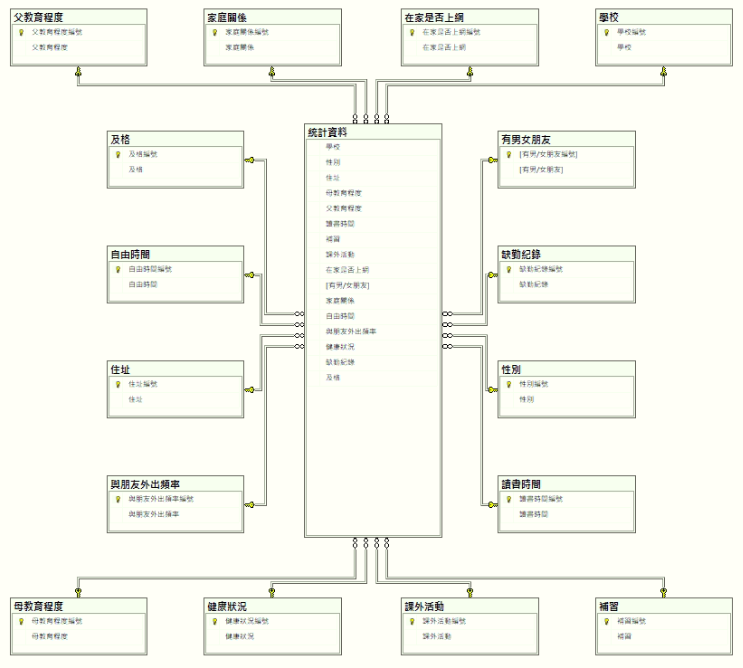
**整理後**

**說明: 在成績方面將三次考試成績平均後合成一個欄位，滿分為20，達到12分為及格，低於則不及格。**

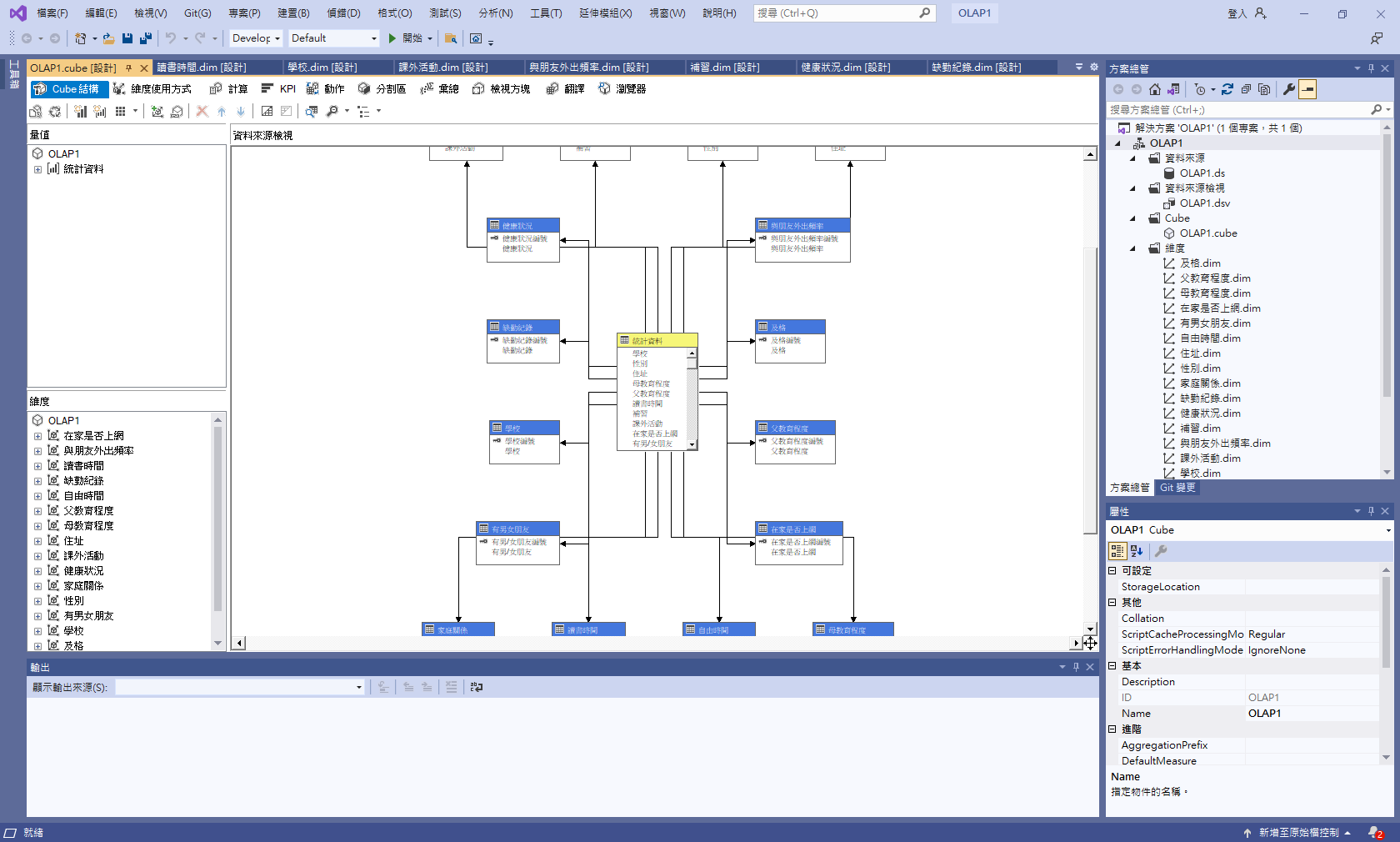


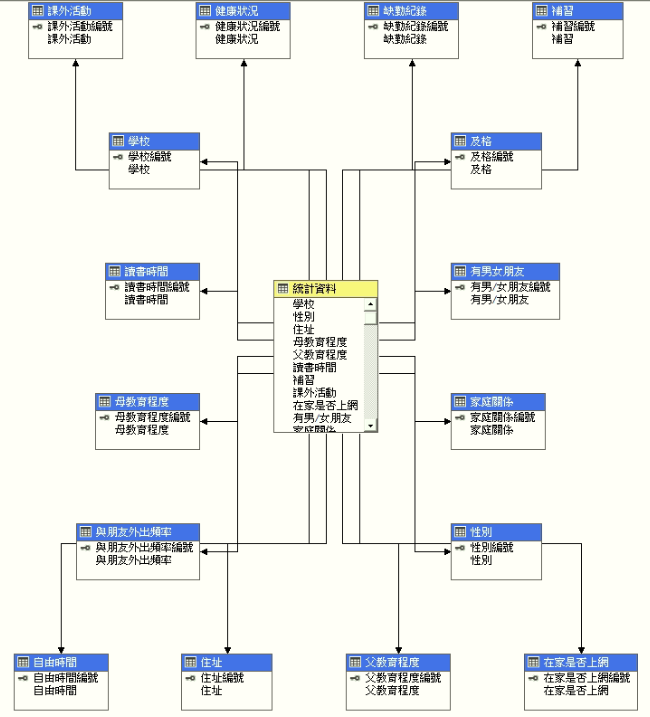
**Star Schema**





**Cube**

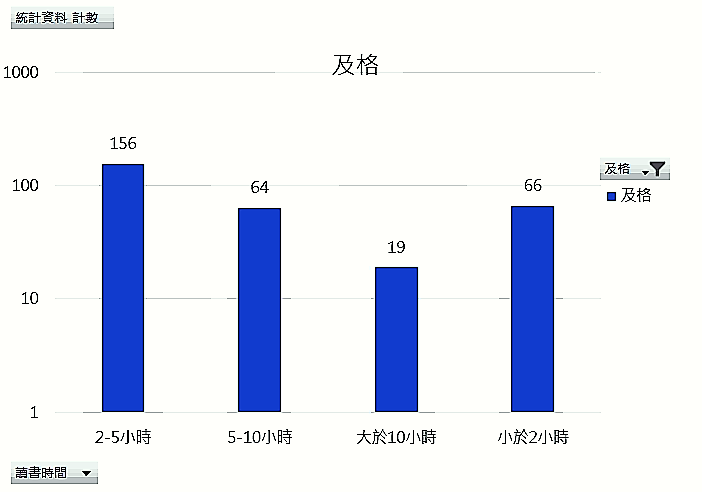




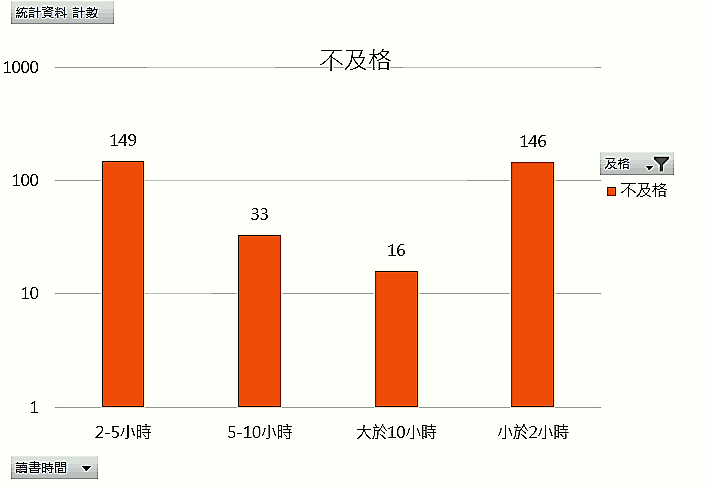
**四、OLAP分析：**

**1、切片(Slice)**

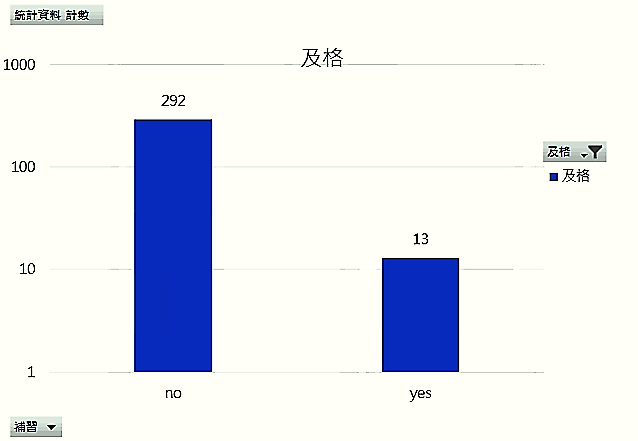
**及格人數的讀書時間**



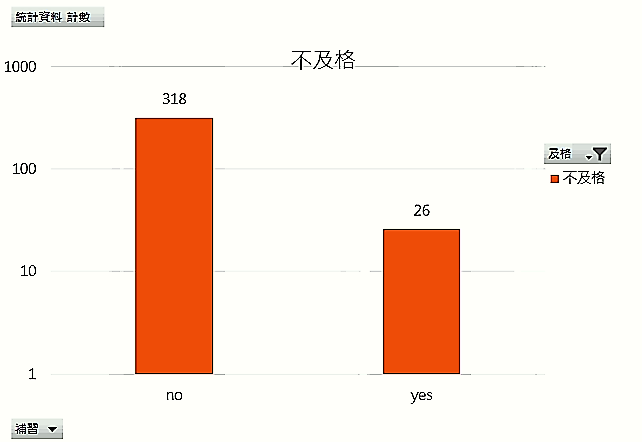
**不及格人數的讀書時間**



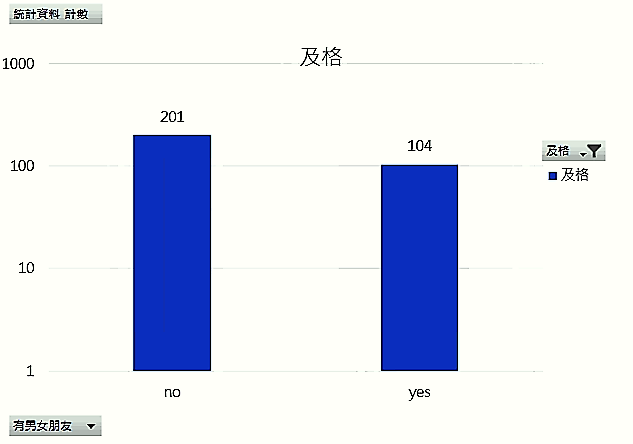
**及格人數的是否補習**



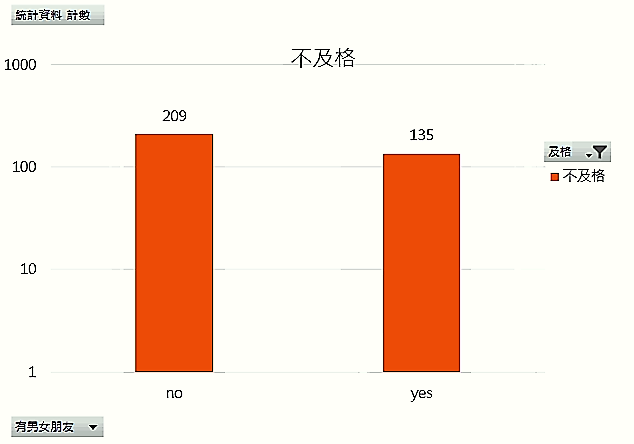
**不及格人數的是否補習**



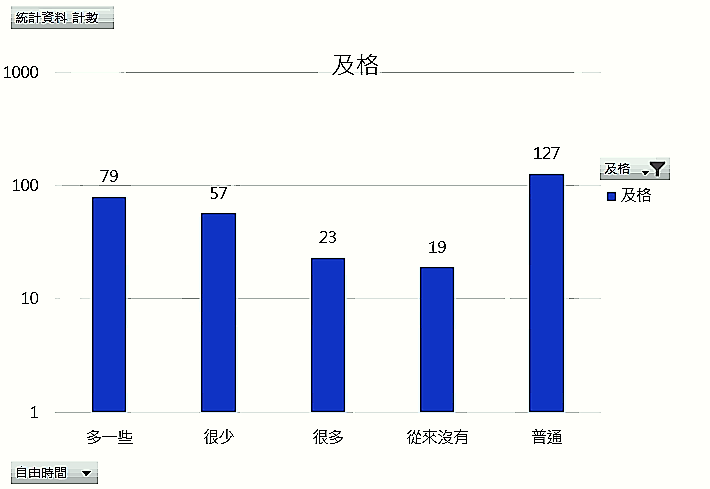
**及格人數的是否有男或女朋友**



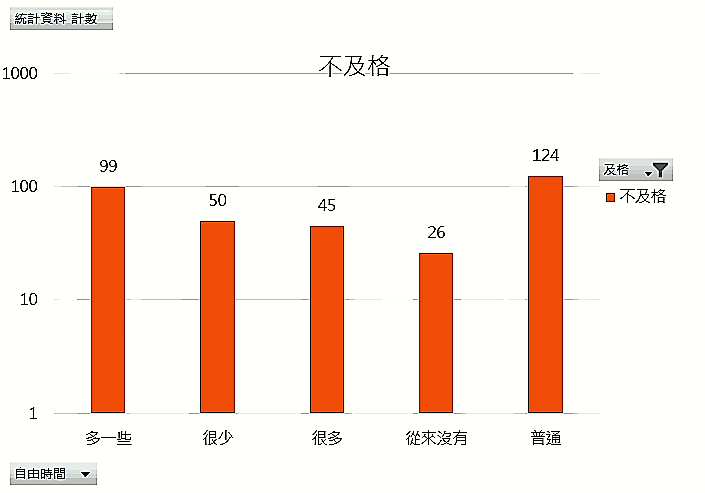
**不及格人數的是否有男或女朋友**



**及格人數的自由時間**



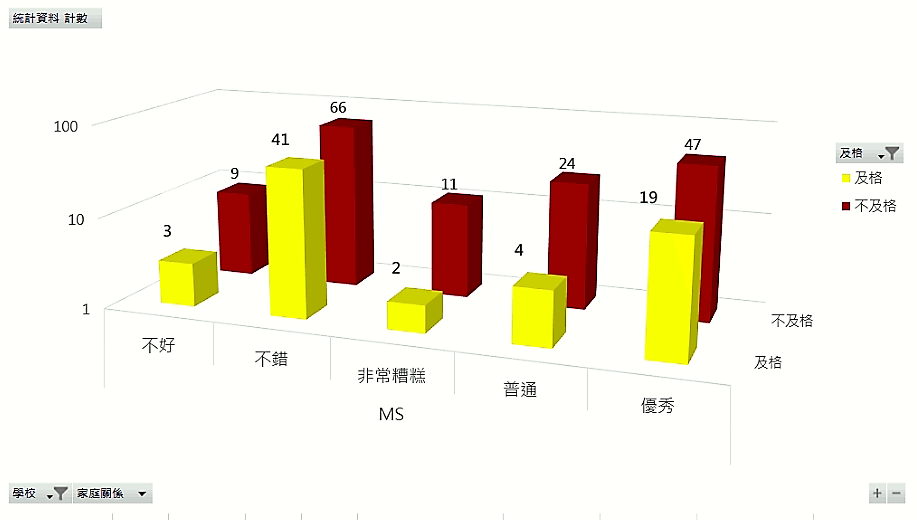
**不及格人數的自由時間**

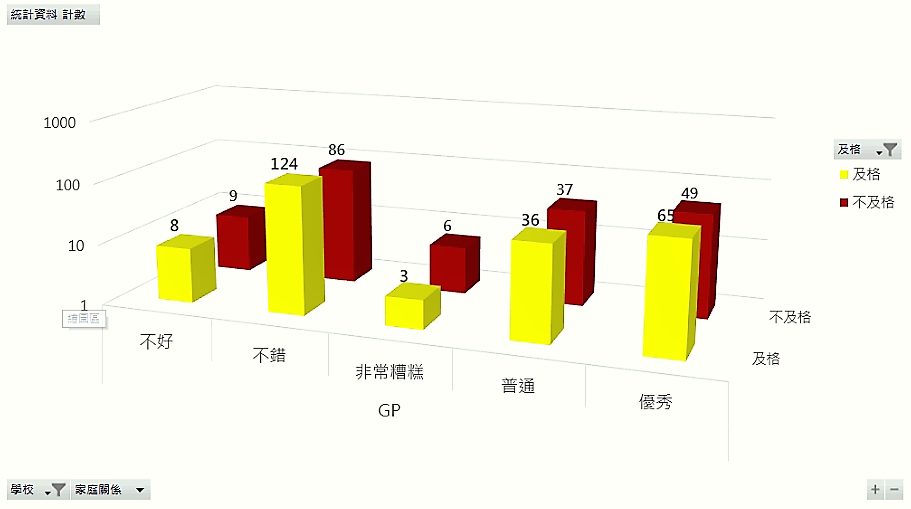


**2、切丁(Dice):**

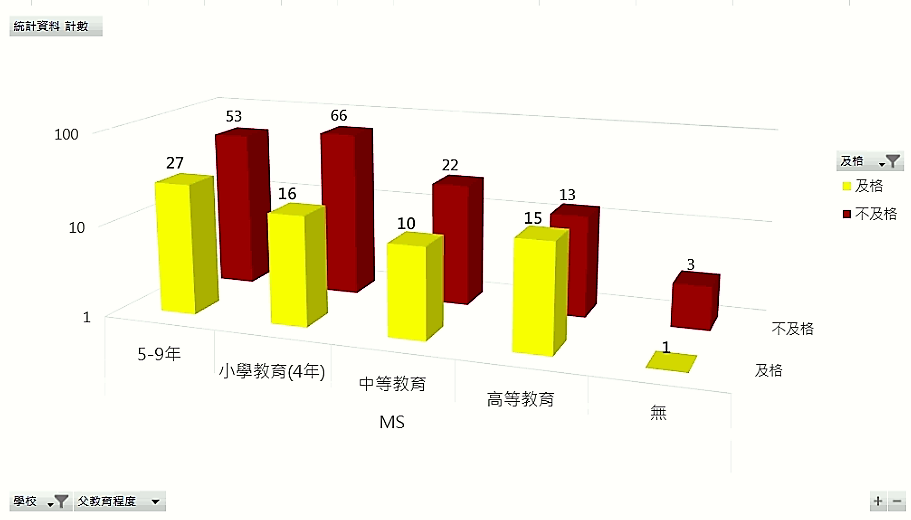
* + 劃分兩個學校去做分析GP和MS。
  + 主要分析兩所學校的家庭關係和父母教育是否和他們成績有關。

**MS學校的家庭關係**

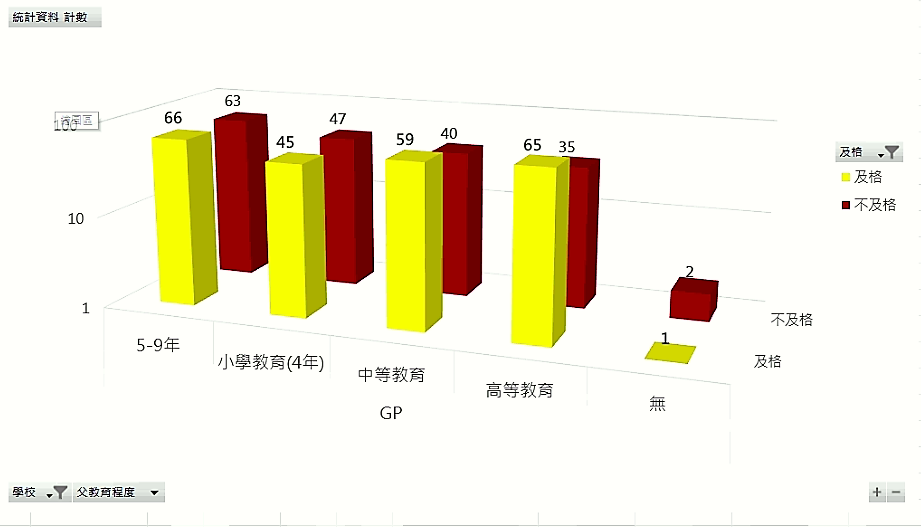


**GP學校的家庭關係**

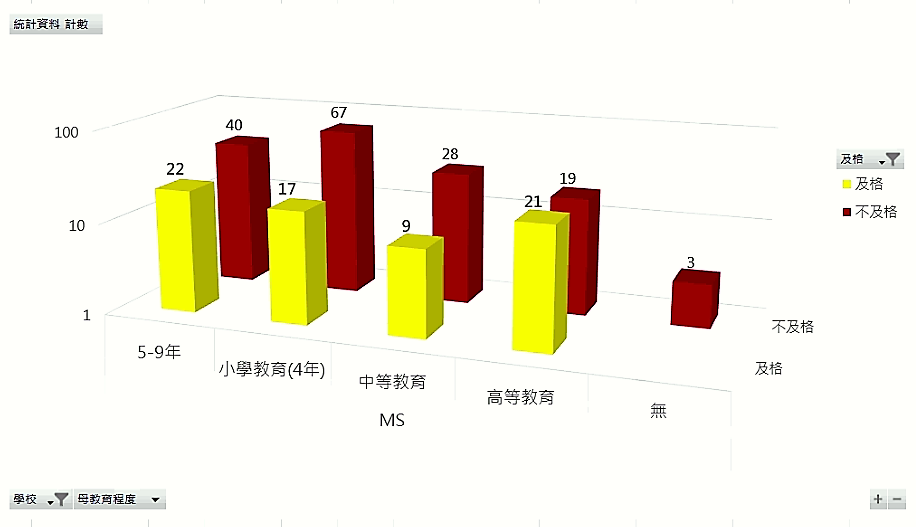
**MS學校的父親教育程度**



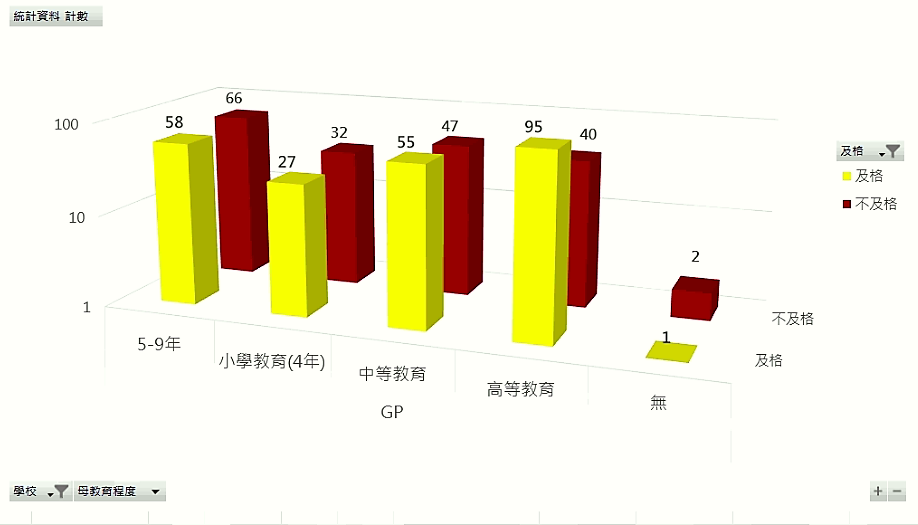
**GP學校的父親教育程度**



**MS學校的母親教育程度**

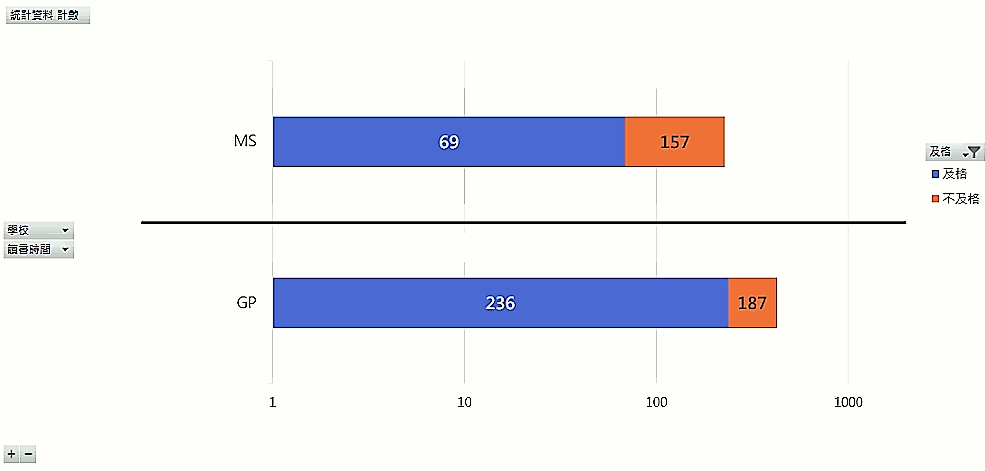


**GP學校的母親教育程度**



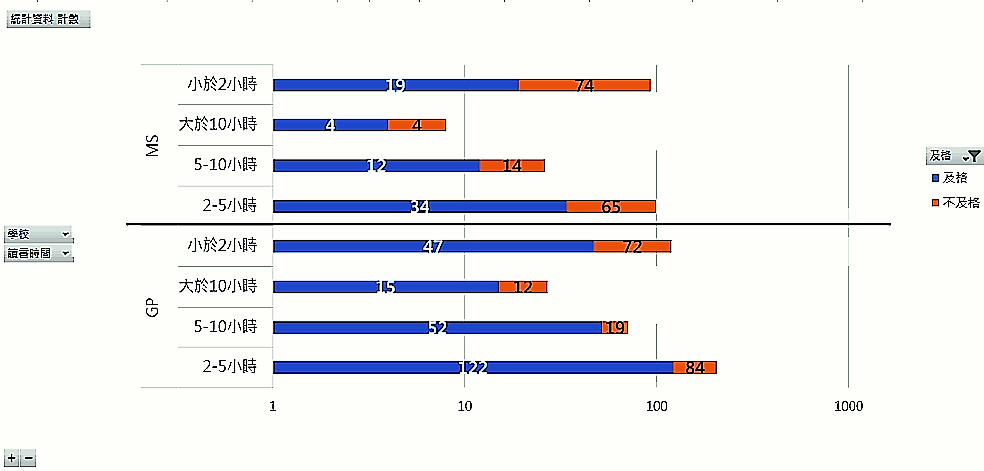
**3、向上截取(Roll-Up) 與 向下擷取(Drill-Down)**

**MS和GP兩學校的成績分布**



**向上截取**

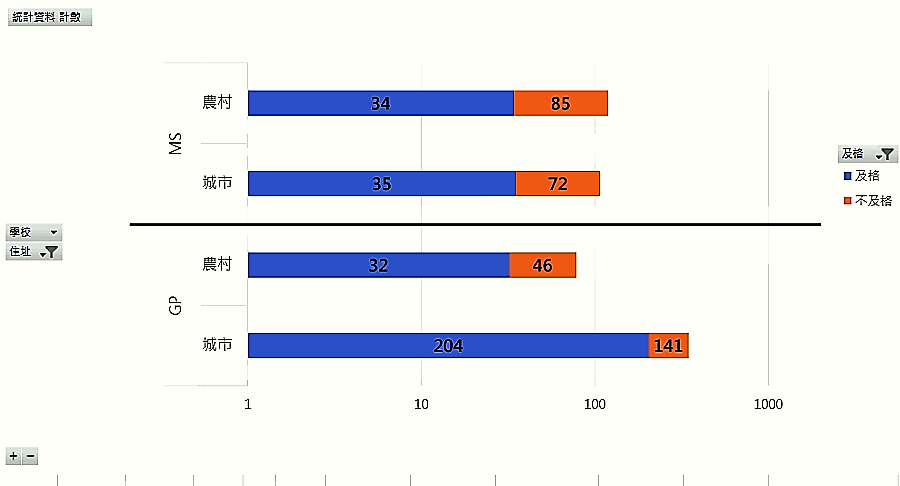
**MS和GP兩學校的讀書時間成績分布**



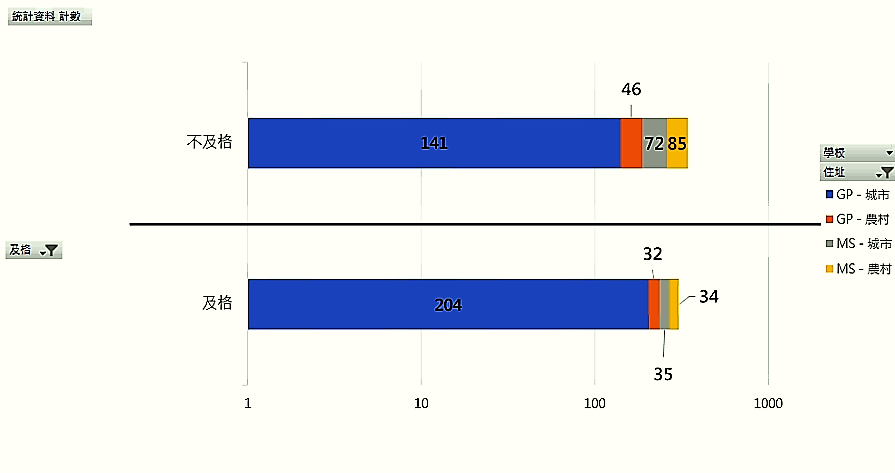
**向下截取**

**3、轉軸(Pivot)**

**MS和GP兩學校學生住址的成績分布**



**原始圖表**



**轉軸後圖表**

**五、分析結論：**

1. **學生多數讀書時間落在2~5小時，但不及格的人數大多都聚集在小於2小時和2~5小時。**
2. **大多的學生都沒有補習，所以補習並不被我們認為是影響成績的主要因素。**
3. **及格和不及格的學生這兩者是否有男女朋友的比數都是2:1的劃分，所以並沒有看出交男女朋友是否為影響成績的主要因素。**
4. **自由時間的部分不及格學生分布都在普通和多一些，也和及格的學生分布相近，但再多一些和很多的人數還是超過及格人數一些。**
5. **兩間學校的家庭關係分佈都在不錯和優秀，及格和不及格人數都很相近，值得注意的地方是兩間學校家庭關係不好和非常糟糕的不及格人數都接近10多人，所以不排除家庭關係也跟成績有所影響。**
6. **推測MS應該是偏鄉的學校，父母的教育程度都在小學程度和5~9年教育，可能也是因素之一讓MS學校的及格率只有30%。**
7. **GP學校就是比較都市的大學校，幾乎父母的教育都落在中等和高等，且及格率也有54%。**
8. **從向下擷取和向上截取的分析圖表中，也能發現比起MS學校，GP學校的讀書時間有蠻多的人數都聚集在5~10小時，可以看的齣都是學校讀書風氣還是高於偏鄉學校的。**
9. **最後的轉軸部分，可以明顯看出GP學校高於79%都是都市人，而MS學校52%的學生都住在農村，所以驗證了上面我們的推測。**
10. **結合上述的分析資料，我們發現在這兩間學校裡，影響學生成績的因素有許多，包括讀書時間、家庭關係和父母教育，還發現了我們預先沒有考量的城鄉差距，在影響學生成績的因素，不能只從單一的面向去看待，也許某些學生正面臨家庭難題或因為需要顧及社團，而耽誤了課業，藉此資料，我們可以從更多不同面向去了解影響成績的各個因素，希望以後學校都能利用這種分析系統，去了解學生的難處，去加以幫助有困難的學生。**