**Introduction to Big Data Analytics**

**Term Project Report**

**Topic: 學生學業成績分析**

張育祿 二資陸生二 107AEA002

李子健 二資陸生二 108AEA001

王翔 二資陸生二 108AEA008

**December 2020**

**動機：**

學業成績對於一個學生來說是非常重要的，但我們並不知道什麼是影響成績最大的因素。我們會通過對資料集的分析，建立模型來預測學生成績。

**計畫摘要：**

我們利用資料中的受教育程度，班級，選擇課程，成績，出勤特徵，以及家長參與等信息，來進行資料分析及視覺化，通過分析資料並建立模型預測學生成績。

**研究步驟：**

資料预处理->資料視覺化->模型建立分析->参数调优->预测效果

**環境：**

Python 3.7 + Jupyter Notebook

**Python 所需套件：**

pandas、sklearn、seaborn、matplotlib、numpy

**參考資料：**

Students' Academic Performance Dataset (xAPI-Edu-Data) <https://www.kaggle.com/brianvancil/xapi-edu-data1>

Seaborn API Website

<https://seaborn.pydata.org/api.html>

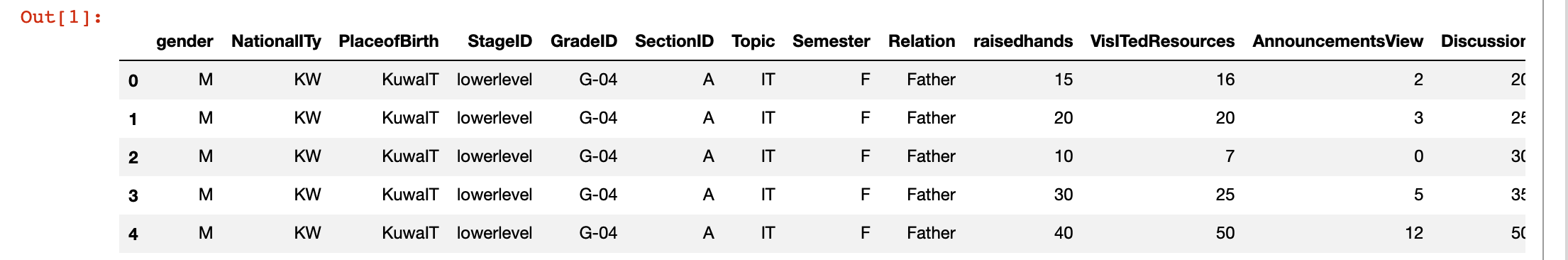
**資料集栏位介紹：**

1. gender-学生性别(“M”或“FM”)
2. National-學生國籍(’ Kuwait’,’ Lebanon’,’ Egypt’,’ SaudiArabia’,’ USA’,’ Jordan’,’ Venezuela’,’ Iran’,’ Tunis’,’ Morocco’,’ Syria’,’ Palestine’,’ Iraq’,’ Lybia)
3. PlaceofBirth-學生出生地(“KuwaIT”、“Jordan”、“Iraq ” 、“lebanon”、“SaudiArabia ”、“USA ”、“Palestine”、“Egypt”、“Tunis”、“Iran”、“Lybia ”、“Syria ”、“Morocco”、“venzuela”)
4. StageID-學生所屬教育級別(“lowerlevel”、“MiddleSchool ”、“HighSchool”)
5. GradeID-年級(“G-01”、“G-02”、“G-03”、“G-04”、“G-05”、“G-06”、“G-07”、“G-08”、“G-09”、“G-10”、“G-11”、“G-12”)
6. SectionID-學生所屬教室(‘A’，‘B’，‘C’)
7. Topic—課程('IT'、 'Math'、 'Arabic' 、'Science'、'English'、 'Quran' 、'Spanish' 、'French' 、'History'、 'Biology' 'Chemistry' 、'Geology')
8. Semester-學年(“F”、“S”)
9. Relation-家長與學生之關係(’mom’,’father’)
10. raisedhands-學生在課堂上有舉手的次數(数字：0-100)
11. VisITedResources-學生訪問課程內容的次數(数字：0-100)
12. AnnouncementsView-學生檢查新公告的次數(数字：0-100)
13. Discussion-學生參與討論小組的次數(数字：0-100)
14. ParentAnsweringSurvey-家長是否回答學校提供的調查(’Yes’,’No’)
15. ParentschoolSatisfaction-家長對學校的滿意度(’Yes’,’No’)
16. StudentAbsenceDays-每名學生缺席天數(above-7, under-7)
17. Class-學生成績分類(L、M、H)

**分析過程：**

**1. 載入套件及資料**

1. **import** pandas as pd
2. **import** numpy as np
3. **import** seaborn as sns
4. **import** matplotlib.pyplot as plt
6. **from** sklearn **import** preprocessing, svm
7. **from** sklearn.linear\_model **import** Perceptron
8. **from** sklearn.tree **import** DecisionTreeClassifier
10. data = pd.read\_csv('xAPI-Edu-Data.csv')
11. data.head()



查看資料

**2. 資料預處理**

**2.1查看資料規格**

**Input:** data.info()

**Output:**

1. <**class** 'pandas.core.frame.DataFrame'>
2. RangeIndex: 480 entries, 0 to 479
3. Data columns (total 17 columns):
4. #   Column                    Non-Null Count  Dtype
5. ---  ------                    --------------  -----
6. 0   gender                    480 non-null    object
7. 1   NationalITy               480 non-null    object
8. 2   PlaceofBirth              480 non-null    object
9. 3   StageID                   480 non-null    object
10. 4   GradeID                   480 non-null    object
11. 5   SectionID                 480 non-null    object
12. 6   Topic                     480 non-null    object
13. 7   Semester                  480 non-null    object
14. 8   Relation                  480 non-null    object
15. 9   raisedhands               480 non-null    int64
16. 10  VisITedResources          480 non-null    int64
17. 11  AnnouncementsView         480 non-null    int64
18. 12  Discussion                480 non-null    int64
19. 13  ParentAnsweringSurvey     480 non-null    object
20. 14  ParentschoolSatisfaction  480 non-null    object
21. 15  StudentAbsenceDays        480 non-null    object
22. 16  Class                     480 non-null    object

根據上表分析可知，不存在空值，資料不需要進行預處理。

**2.2查看學生成績類別**

Input: data.Class.unique()

Output: array(['M', 'L', 'H'], dtype=object)

學生成績一共分為三類['L', 'M', 'H']，這將作為評判學生的標準。

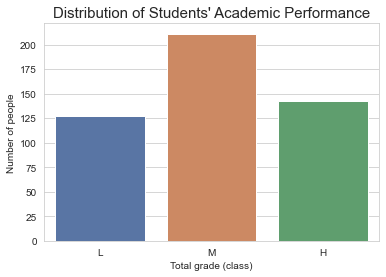
L:0-59 不及格；

M:60-89 中等；

H:90-100 高分。

**3. 資料視覺化**

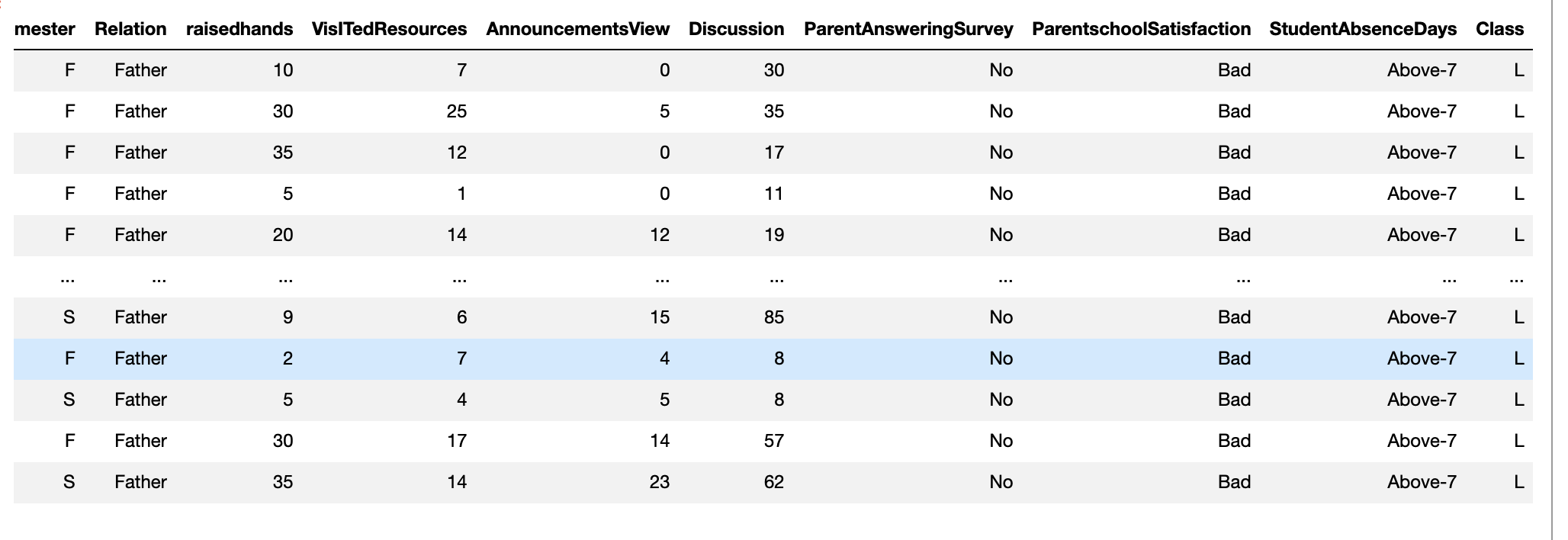
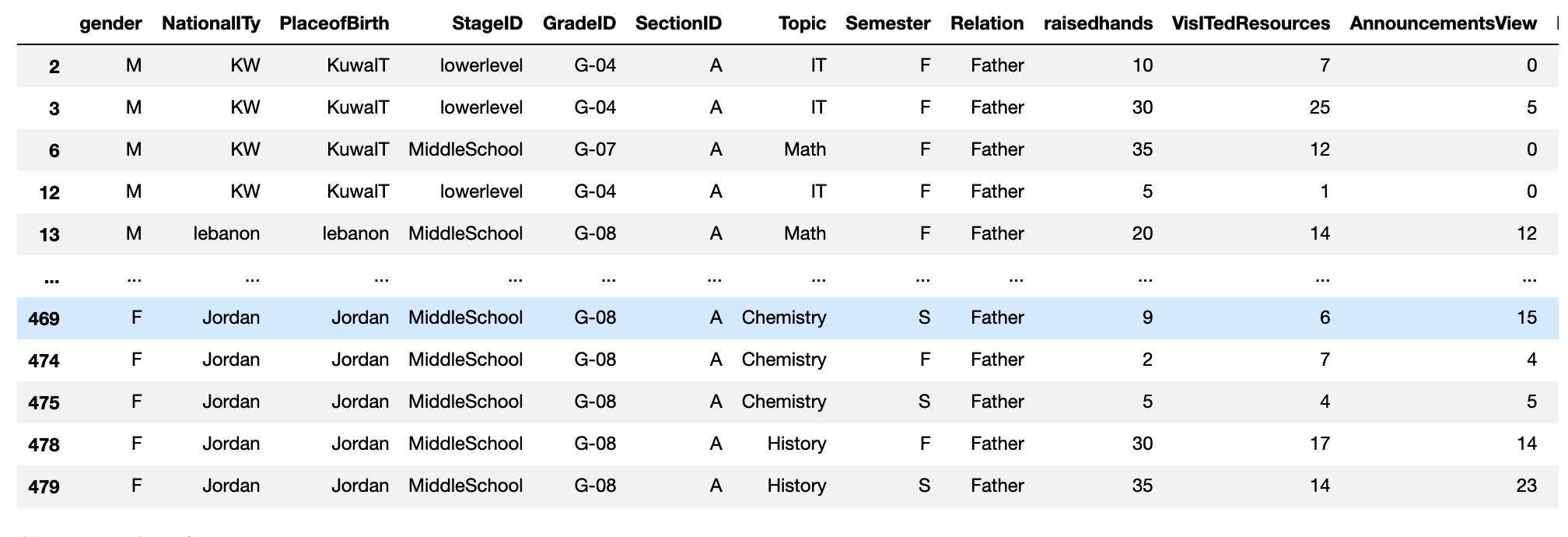
1. sns.set\_style("whitegrid")
2. ax = sns.countplot(x='Class', data=data, order=['L', 'M', 'H'], palette="deep")
3. plt.xlabel('Total grade (class)')
4. plt.ylabel('Number of people')
5. plt.title("Distribution of Students' Academic Performance", size=15)
6. plt.show()



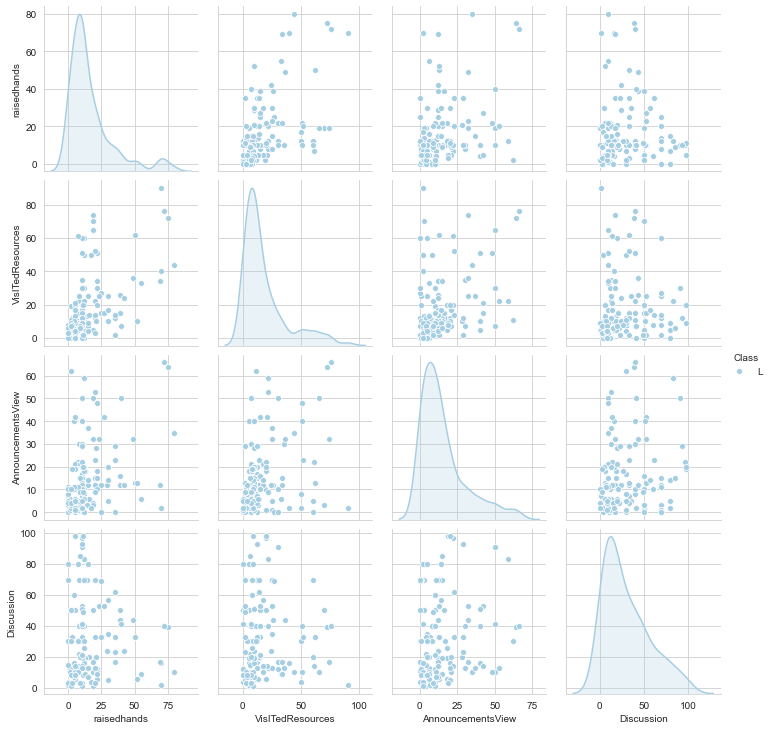
根據上圖得知，大部分學生都處於中等成績，高分其次，不及格的人數最少。

**3.1查看不及格學生的信息**

1. data.loc[data["Class"]=='L']



1. low\_class=data.loc[data["Class"]=='L']
2. sns.pairplot(low\_class, hue="Class",
3. diag\_kind="kde", palette="Paired")
4. plt.show()

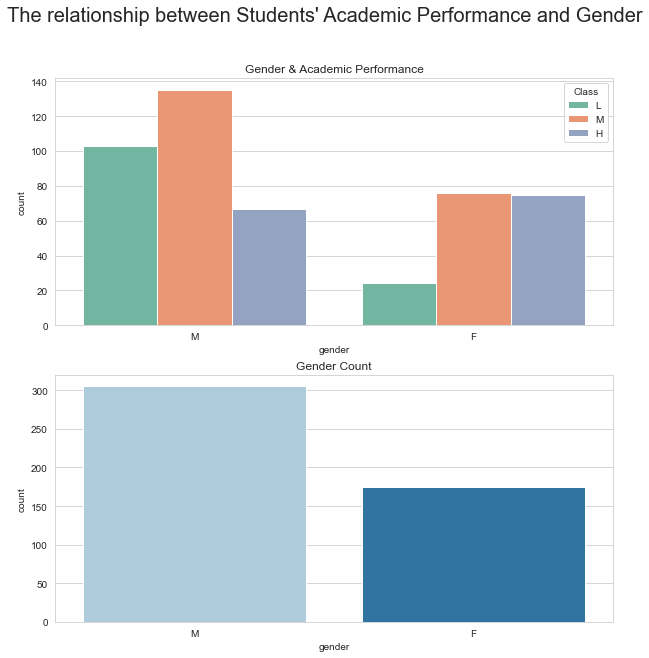


根據以上兩個表格和圖分析得知，似乎不及格的學生缺課天數都超過七天，各項數值都集中在一個很低的區域，例如舉手少的人，一般也不會參與討論（由左下圖得知）。

我們還能觀察到不及格的學生一般監護人是父親，且他們沒有接受學校調查、對學校的滿意度不高。

**3.2學生成績與性別的關係**

1. fig, axarr  = plt.subplots(2,figsize=(10,10))
2. sns.countplot(x='gender', hue='Class', data=data, order=['M', 'F'],hue\_order = ['L', 'M', 'H'], ax=axarr[0], palette="Set2")
3. sns.countplot(x='gender', data=data, order=['M','F'], ax=axarr[1], palette="Paired")
4. axarr[0].set\_title('Gender & Academic Performance')
5. axarr[1].set\_title('Gender Count')
6. fig.suptitle("The relationship between Students' Academic Performance and Gender", size=20)
7. plt.show()

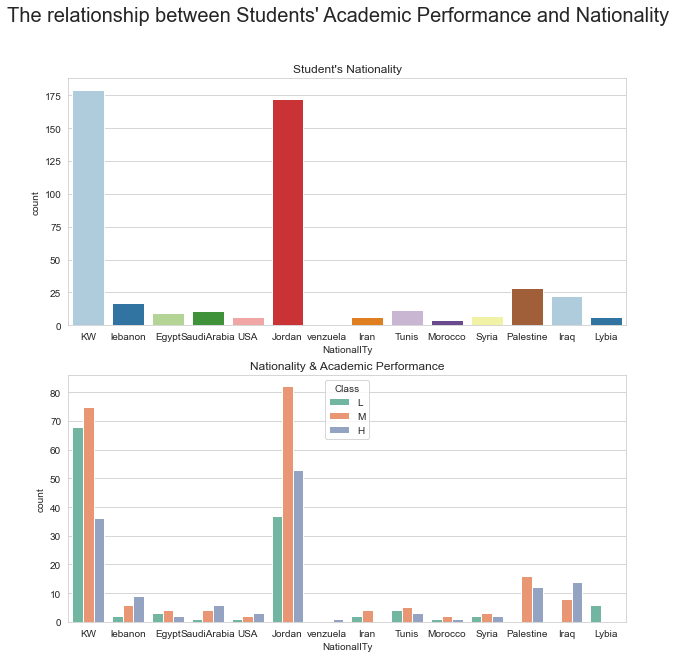


根據此圖分析，我們可以明顯看出女學生的不及格人數要遠遠少於男學生，且女學生的中等和高分段學生人數基本持平，男女學生在高分段人數相差不大。

由此我們可以推測：性別可能影響學生的成績。

**3.3學生成績與國籍的關係**

1. fig, axarr  = plt.subplots(2,figsize=(10,10))
2. axarr[0].set\_title("Student's Nationality")
3. axarr[1].set\_title('Nationality & Academic Performance')
4. fig.suptitle("The relationship between Students' Academic Performance and Nationality", size=20)
5. sns.countplot(x='NationalITy', data=data, ax=axarr[0], palette="Paired")
6. sns.countplot(x='NationalITy', hue='Class', data=data,hue\_order = ['L', 'M', 'H'], ax=axarr[1], palette="Set2")
7. plt.show()



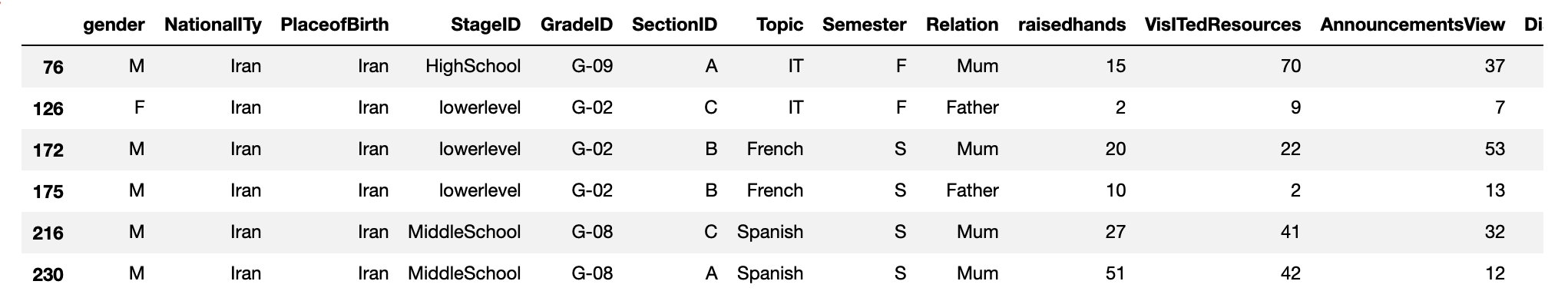
根據此圖分析，除了KW和Jordan這兩個國籍的學生之外，其餘國籍的學生可分析的樣本數量都比較少，並不能因此推斷出任何有效的信息。

我們能得出的信息有：相比於KW的學生，Jordan學生不及格人數為KW學生的一半，總體成績更好一些，而 Iran和Lybia國籍的學生沒有取得高分的。

我們來查看 Iran 和 Lybia 國籍的學生樣本，嘗試分析影響他們學業成績的因素。

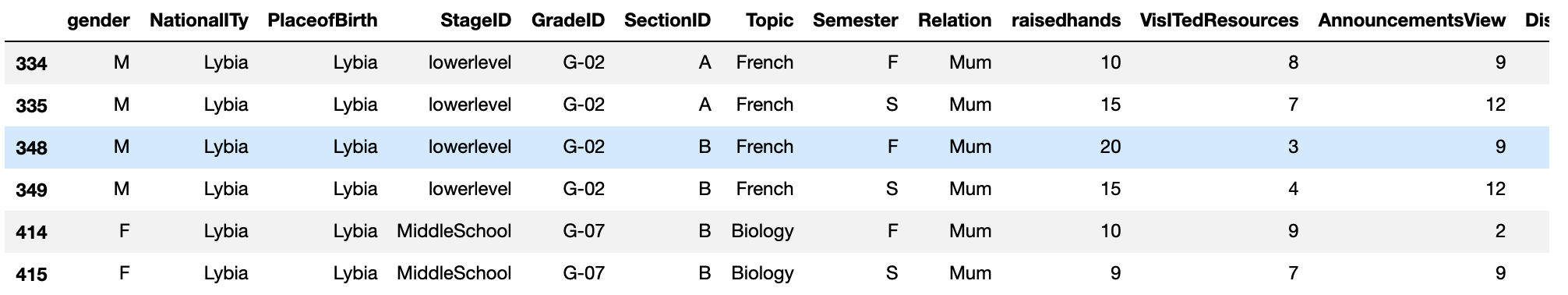
查看 Iran 國籍的學生樣本：

1. data.loc[data['NationalITy'] == 'Iran']



查看 Lybia 國籍的學生的樣本：

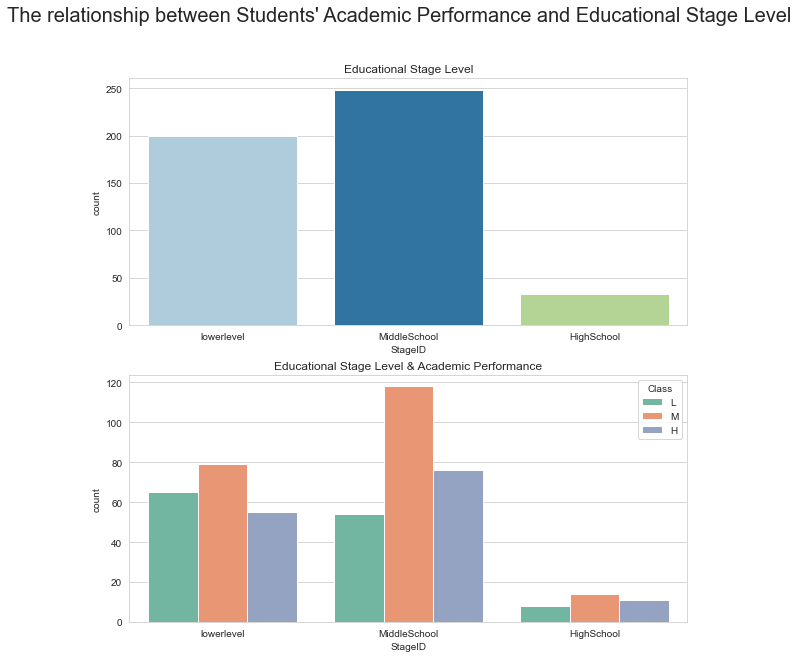
1. data.loc[data['NationalITy'] == 'Lybia']



通過觀察，我們可以得知 Lybia 國籍的學生似乎與所有未通過考試的學生的數據有高度的重合(缺課超過7天，數值偏低，沒有學校調查等)。

**3.4學生所屬教育級別與成績的關係**

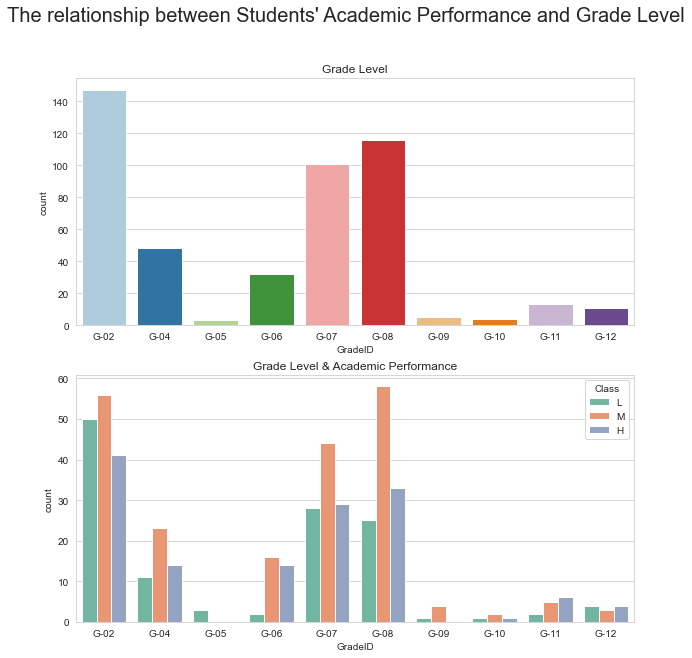
1. fig, axarr  = plt.subplots(2,figsize=(10,10))
2. axarr[0].set\_title('Educational Stage Level')
3. axarr[1].set\_title('Educational Stage Level & Academic Performance')
4. fig.suptitle("The relationship between Students' Academic Performance and Educational Stage Level", size=20)
5. sns.countplot(x='StageID', data=data, ax=axarr[0], palette="Paired")
6. sns.countplot(x='StageID', hue='Class', data=data, hue\_order = ['L', 'M', 'H'], ax=axarr[1], palette="Set2")
7. plt.show()



根據此圖分析，我們可以得知不管學生處於哪個教育級別，都是中等成績的人數偏多。

**3.5學生所屬年級與成績的關係**

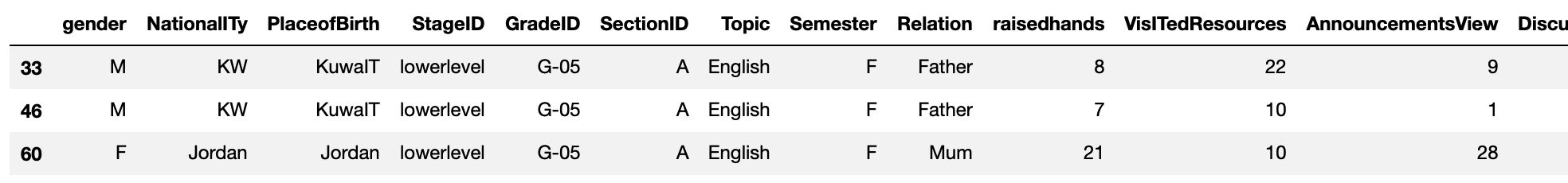
1. fig, axarr  = plt.subplots(2,figsize=(10,10))
2. axarr[0].set\_title('Grade Level')
3. axarr[1].set\_title('Grade Level & Academic Performance')
4. fig.suptitle("The relationship between Students' Academic Performance and Grade Level", size=20)
5. sns.countplot(x='GradeID',
6. data=data,
7. order=['G-02', 'G-04', 'G-05', 'G-06', 'G-07', 'G-08', 'G-09', 'G-10', 'G-11', 'G-12'],
8. ax=axarr[0], palette="Paired")
9. sns.countplot(x='GradeID',
10. hue='Class',
11. data=data,
12. order=['G-02', 'G-04', 'G-05', 'G-06', 'G-07', 'G-08', 'G-09', 'G-10', 'G-11', 'G-12'],
13. hue\_order = ['L', 'M', 'H'],
14. ax=axarr[1], palette="Set2")
15. plt.show()



根據此圖分析，我們可以得知五年級、九年級、十年級的學生人數很少。除此之外，沒有五年級學生及格，也沒有九年級學生取得高分。

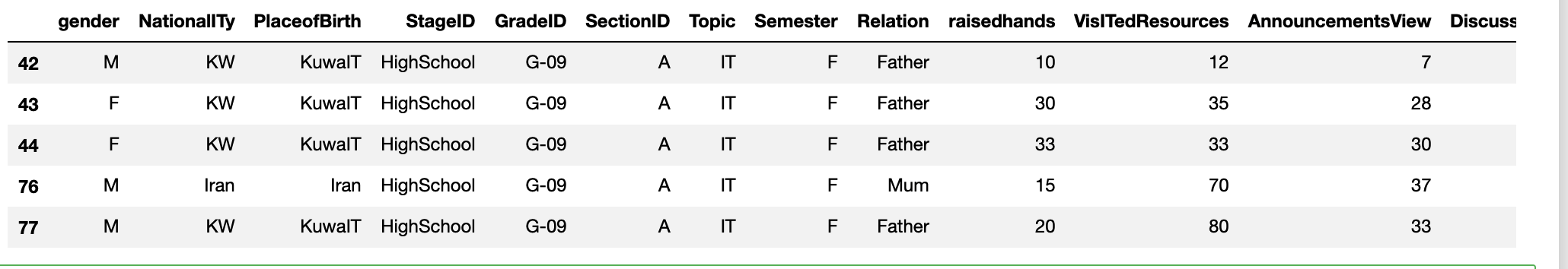
查看五年級學生的樣本：

1. data.loc[data['GradeID']=='G-05']



查看九年級學生的樣本：

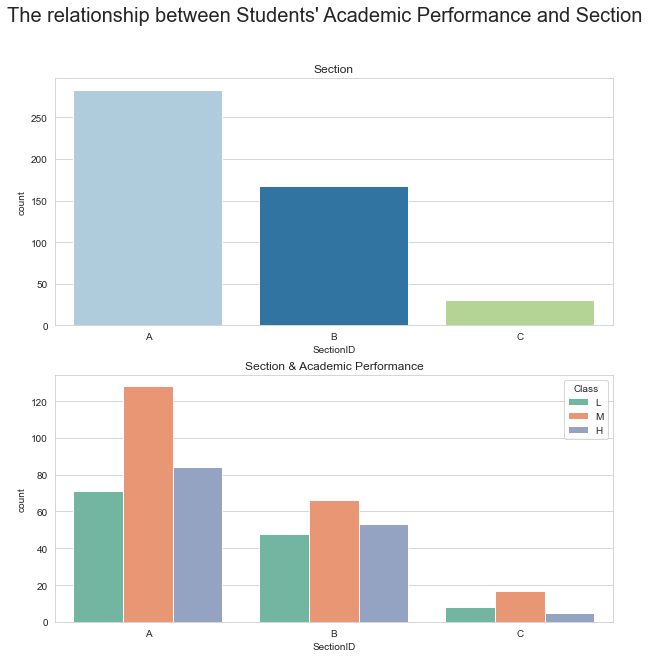
1. data.loc[data['GradeID']=='G-09']



在觀察後，發現五年級和九年級學生似乎與所有未通過考試的學生的數據有高度的重合(缺課超過7天，數值偏低，沒有學校調查等) 。

**3.6學生所屬教室與成績的關係**

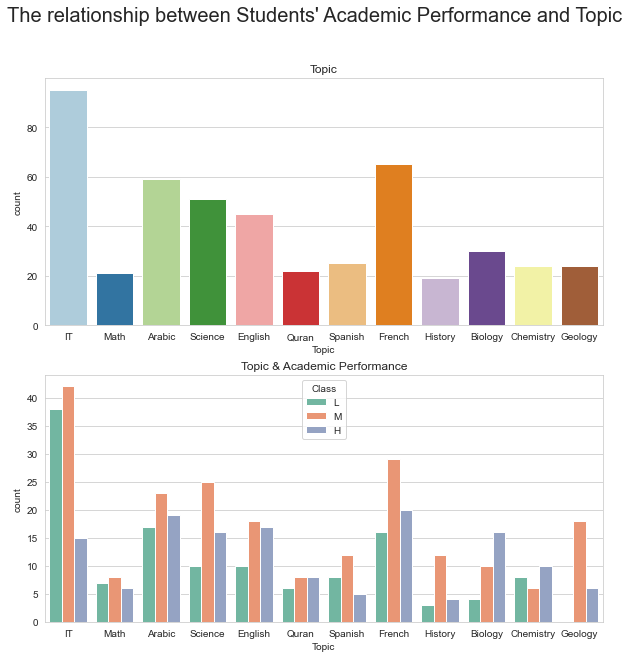
1. fig, axarr  = plt.subplots(2,figsize=(10,10))
2. axarr[0].set\_title('Section')
3. axarr[1].set\_title('Section & Academic Performance')
4. fig.suptitle("The relationship between Students' Academic Performance and Section", size=20)
5. sns.countplot(x='SectionID', data=data,
6. order=['A', 'B', 'C'], ax = axarr[0], palette="Paired")
7. sns.countplot(x='SectionID', hue='Class',
8. data=data, order=['A', 'B', 'C'],
9. hue\_order = ['L', 'M', 'H'], ax = axarr[1], palette="Set2")
10. plt.show()



根據此圖分析，我們得知三個班的總體趨勢都差不多，我們並不能得出什麼有效信息。

**3.7學生所選課程與成績的關係**

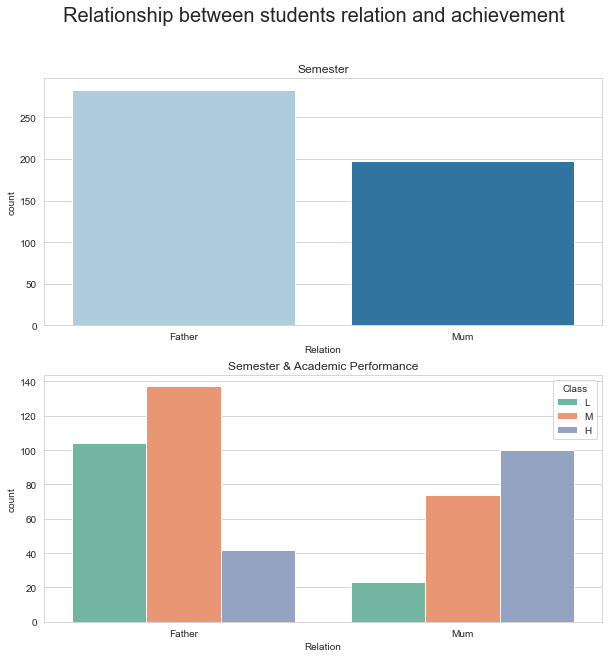
1. fig, axarr  = plt.subplots(2,figsize=(10,10))
2. axarr[0].set\_title('Topic')
3. axarr[1].set\_title('Topic & Academic Performance')
4. fig.suptitle("The relationship between Students' Academic Performance and Topic", size=20)
5. sns.countplot(x='Topic', data=data, ax = axarr[0], palette="Paired")
6. sns.countplot(x='Topic', hue='Class', data=data,hue\_order = ['L', 'M', 'H'], ax = axarr[1], palette="Set2")
7. plt.show()



根據此圖分析，我們可以看到一個有趣的現象，Geology 課程沒有不及格的學生。這是為什麼呢？

**3.8不同學年學生與成績的關係**

1. fig, axarr  = plt.subplots(2,figsize=(10,10))
2. axarr[0].set\_title('Semester')
3. axarr[1].set\_title('Semester & Academic Performance')
4. fig.suptitle("The relationship between Students' Academic Performance and Semester", size=20)
5. sns.countplot(x='Semester', data=data, ax = axarr[0], palette="Paired")
6. sns.countplot(x='Semester', hue='Class', data=data,hue\_order = ['L', 'M', 'H'], ax = axarr[1], palette="Set2")
7. plt.show()

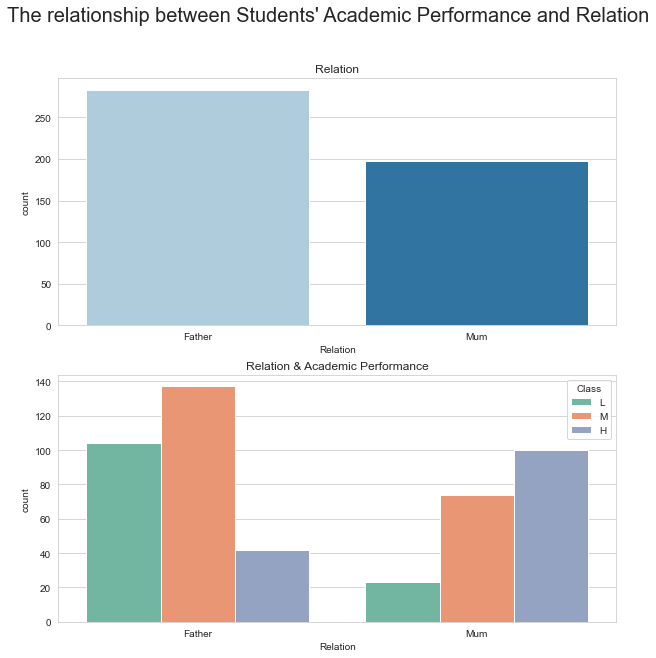


根據此圖分析，第二學年的不及格人數比第一學年少，高分人數都比第一學年多。

由此我們可以推測：學年可能會影響成績。

**3.9學生監護人與成績的關係**

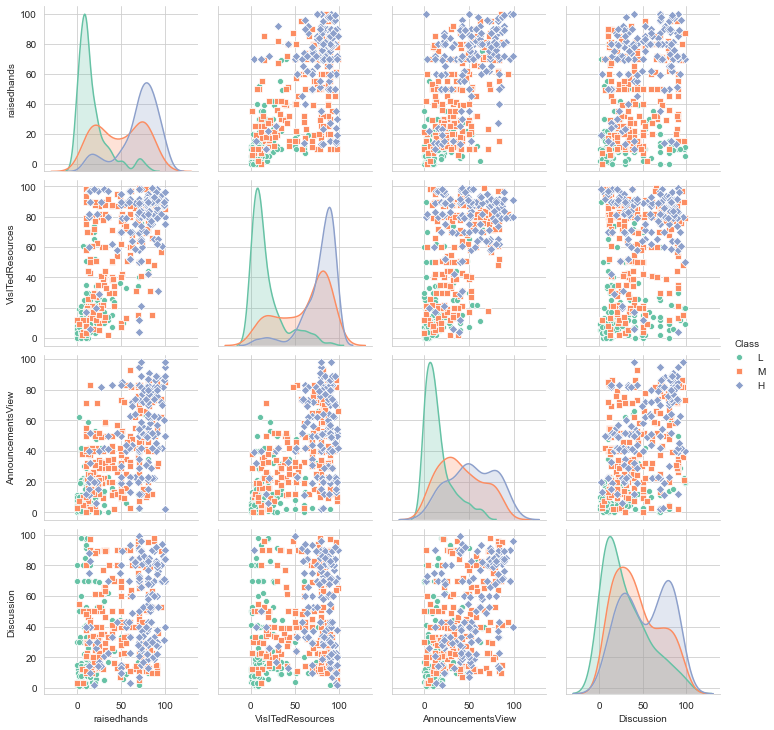
1. fig, axarr  = plt.subplots(2,figsize=(10,10))
2. axarr[0].set\_title('Relation')
3. axarr[1].set\_title('Relation & Academic Performance')
4. fig.suptitle("The relationship between Students' Academic Performance and Relation", size=20)
5. sns.countplot(x='Relation', data=data, ax = axarr[0], palette="Paired")
6. sns.countplot(x='Relation', hue='Class', data=data,hue\_order = ['L', 'M', 'H'], ax = axarr[1], palette="Set2")
7. plt.show()



根據此圖分析，母親作為監護人和學生及格之間似乎有關聯，父親作為監護人和學生不及格之間似乎有關聯。

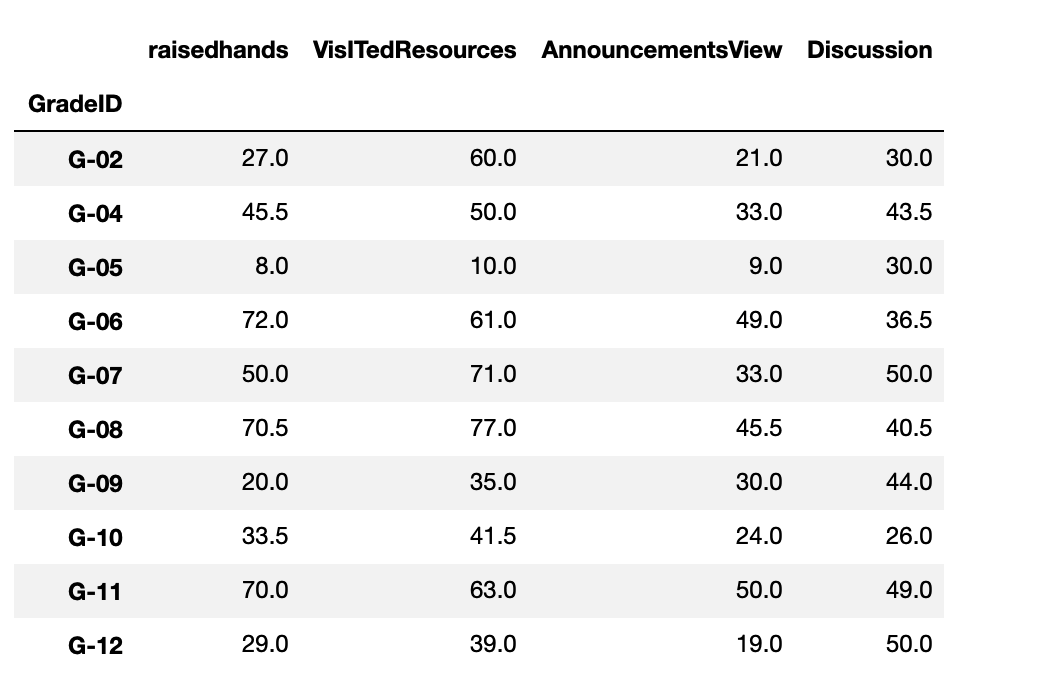
**3.10學生在課堂舉手次數、訪問課程內容次數、檢查新公告次數、參加討論次數與成績的關係**

1. sns.pairplot(data, hue="Class",
2. diag\_kind="kde",
3. hue\_order = ['L', 'M', 'H'],
4. markers=["o", "s", "D"], palette="Set2")
5. plt.show()



查看不同教育層次的raisedhands、VisITedResources、AnnouncementsView、Discussion，獲得此處中位數：

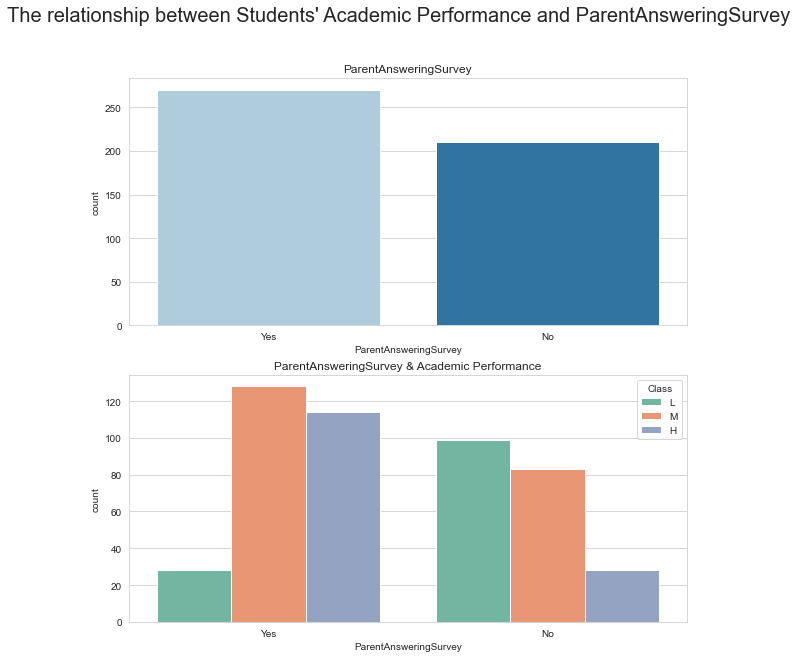
1. data.groupby('GradeID').median()



在這裡我們可以看出五年級和九年級的數據比其他大多數年級少上許多。

**3.11家長是否回答學校提供的調查與學生成績的關係**

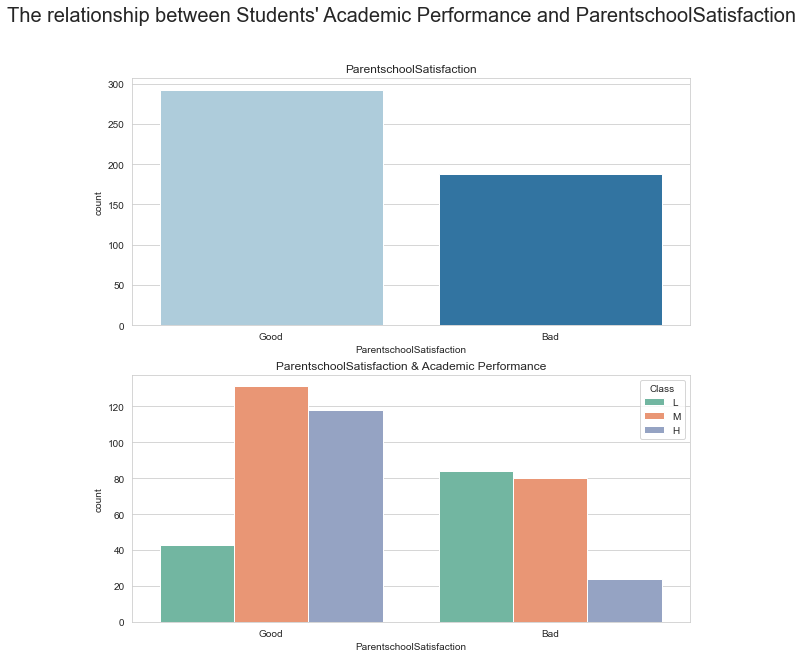
1. fig, axarr  = plt.subplots(2,figsize=(10,10))
2. axarr[0].set\_title('ParentAnsweringSurvey')
3. axarr[1].set\_title('ParentAnsweringSurvey & Academic Performance')
4. fig.suptitle("The relationship between Students' Academic Performance and ParentAnsweringSurvey", size=20)
5. sns.countplot(x='ParentAnsweringSurvey', data=data,
6. order=['Yes', 'No'], ax = axarr[0], palette="Paired")
7. sns.countplot(x='ParentAnsweringSurvey', hue='Class',
8. data=data, order=['Yes', 'No'], hue\_order = ['L', 'M', 'H'],
9. ax = axarr[1], palette="Set2")
10. plt.show()



觀察這兩張圖可能會覺得家長是否回答學校調查與學生成績有關，但是我們並不知道是不是學生的成績導致他們是否回應學校的調查，這無法從資料得出。

**3.12家長對於學校滿意度與學生成績的關係**

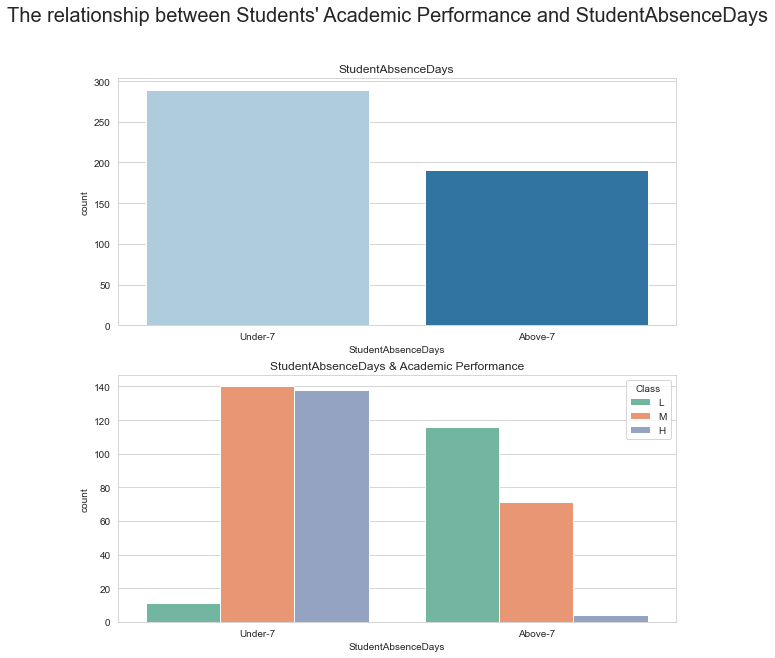
1. fig, axarr  = plt.subplots(2,figsize=(10,10))
2. axarr[0].set\_title('ParentschoolSatisfaction')
3. axarr[1].set\_title('ParentschoolSatisfaction & Academic Performance')
4. fig.suptitle("The relationship between Students' Academic Performance and ParentschoolSatisfaction", size=20)
5. sns.countplot(x='ParentschoolSatisfaction', data=data,
6. order=['Good', 'Bad'], ax = axarr[0], palette="Paired")
7. sns.countplot(x='ParentschoolSatisfaction', hue='Class',
8. data=data, order=['Good', 'Bad'],
9. hue\_order = ['L', 'M', 'H'], ax = axarr[1], palette="Set2")
10. plt.show()



與 3.11 觀察結果相同，不再敘述。

**3.13學生缺勤天數與學生成績的關係**

1. fig, axarr  = plt.subplots(2,figsize=(10,10))
2. axarr[0].set\_title('StudentAbsenceDays')
3. axarr[1].set\_title('StudentAbsenceDays & Academic Performance')
4. fig.suptitle("The relationship between Students' Academic Performance and StudentAbsenceDays", size=20)
5. sns.countplot(x='StudentAbsenceDays', data=data,
6. order=['Under-7', 'Above-7'],
7. ax = axarr[0], palette="Paired")
8. sns.countplot(x='StudentAbsenceDays', hue='Class',
9. data=data, order=['Under-7', 'Above-7'],
10. hue\_order = ['L', 'M', 'H'],
11. ax = axarr[1], palette="Set2")
12. plt.show()



根據此圖分析，學習時間與學生成績有很強的相關性，缺課查過七天的學生很少取得高分，缺課少於七天的學生很少不及格。

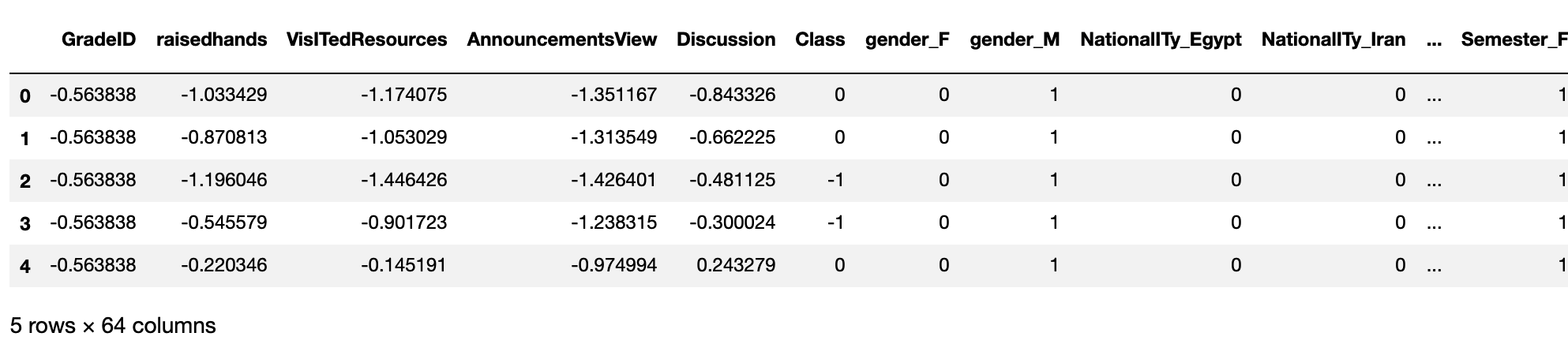
綜上分析，學生成績與訪問課程內容的次數、缺席天數、在課上有舉手的次數、檢查新公告的次數、是否參加討論、性別和學期這些屬性有關。

4. 建立模型預測與學生成績的關係

**4.1處理資料**

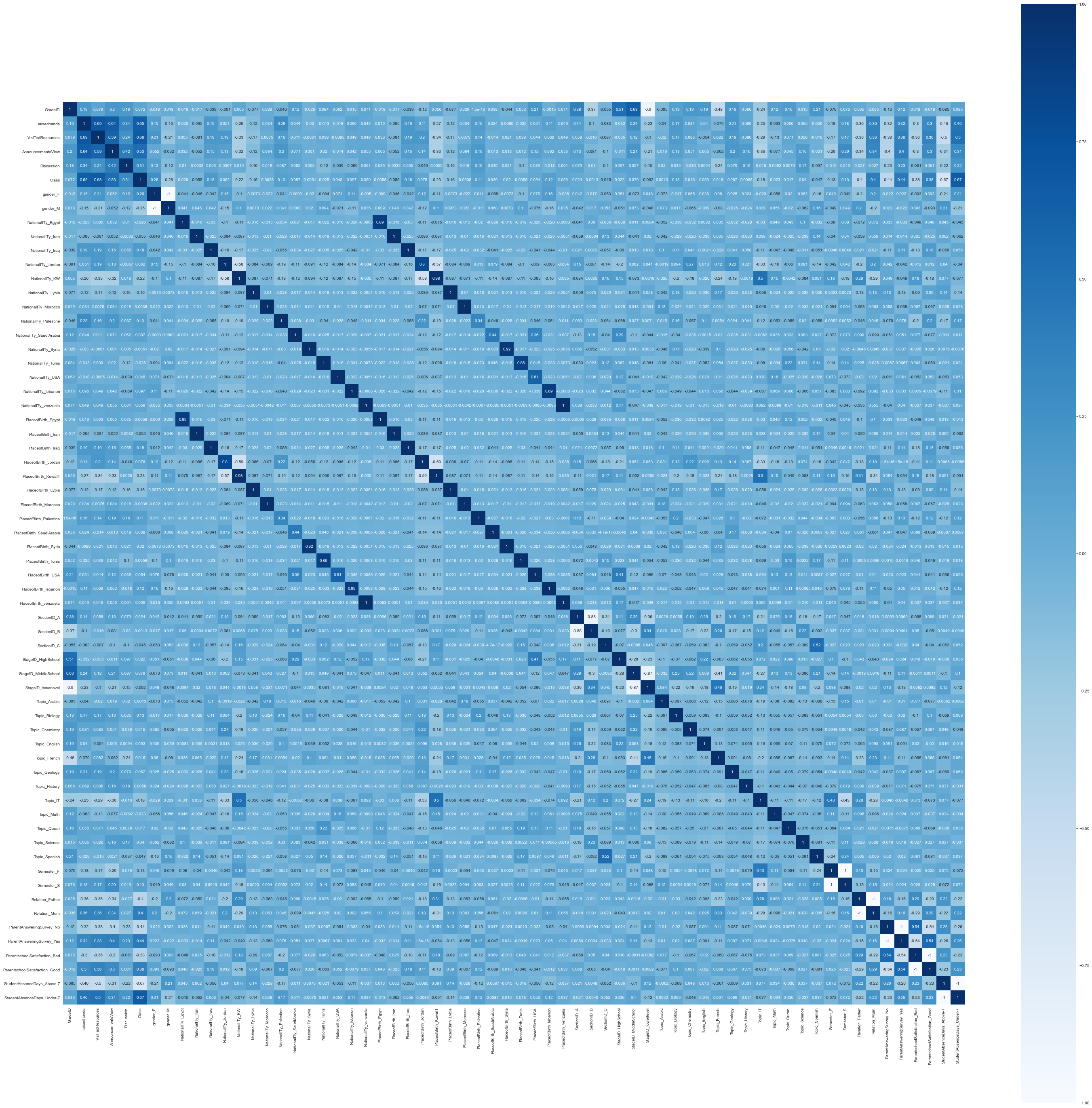
1. # Convert grades into data
2. gradeID\_dict = {"G-01" : 1,
3. "G-02" : 2,
4. "G-03" : 3,
5. "G-04" : 4,
6. "G-05" : 5,
7. "G-06" : 6,
8. "G-07" : 7,
9. "G-08" : 8,
10. "G-09" : 9,
11. "G-10" : 10,
12. "G-11" : 11,
13. "G-12" : 12}
15. data = data.replace({"GradeID" : gradeID\_dict})
16. # Convert scores into data
17. class\_dict = {"L" : -1,
18. "M" : 0,
19. "H" : 1}
20. data = data.replace({"Class" : class\_dict})
22. # Convert to Scale data
23. data["GradeID"] = preprocessing.scale(data["GradeID"])
24. data["raisedhands"] = preprocessing.scale(data["raisedhands"])
25. data["VisITedResources"] = preprocessing.scale(data["VisITedResources"])
26. data["AnnouncementsView"] = preprocessing.scale(data["AnnouncementsView"])
27. data["Discussion"] = preprocessing.scale(data["Discussion"])
29. # Use virtual code conversion to convert 11 columns into 64 columns
30. data = pd.get\_dummies(data, columns=["gender",
31. "NationalITy",
32. "PlaceofBirth",
33. "SectionID",
34. "StageID",
35. "Topic",
36. "Semester",
37. "Relation",
38. "ParentAnsweringSurvey",
39. "ParentschoolSatisfaction",
40. "StudentAbsenceDays"])

43. data.head()

****

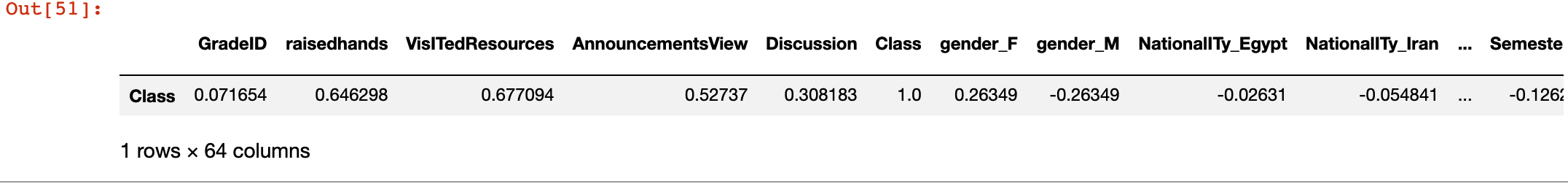
**4.2列出成績與其他屬性的相關性**

1. corr = data.corr()
2. mask = np.triu(np.ones\_like(corr, dtype=bool))
3. f, ax = plt.subplots(figsize=(11, 9))
4. cmap = sns.diverging\_palette(230, 20, as\_cmap=True)
5. sns.heatmap(corr, mask=mask, cmap=cmap, vmax=.3, center=0,
6. square=True, linewidths=.5, cbar\_kws={"shrink": .5})



我們單獨查看 Class 與其他列的關係：

1. corr = data.corr()
2. corr.iloc[[5]]

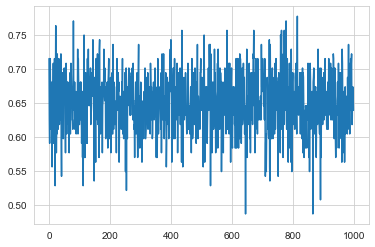


根據上圖和表格，我們可以看出訪問課程內容的次數、缺席天數、在課上有舉手的次數、檢查新公告的次數、是否參加討論、性別和學期都與 Class 有很強的相關性，這和我們之前的分析一樣。

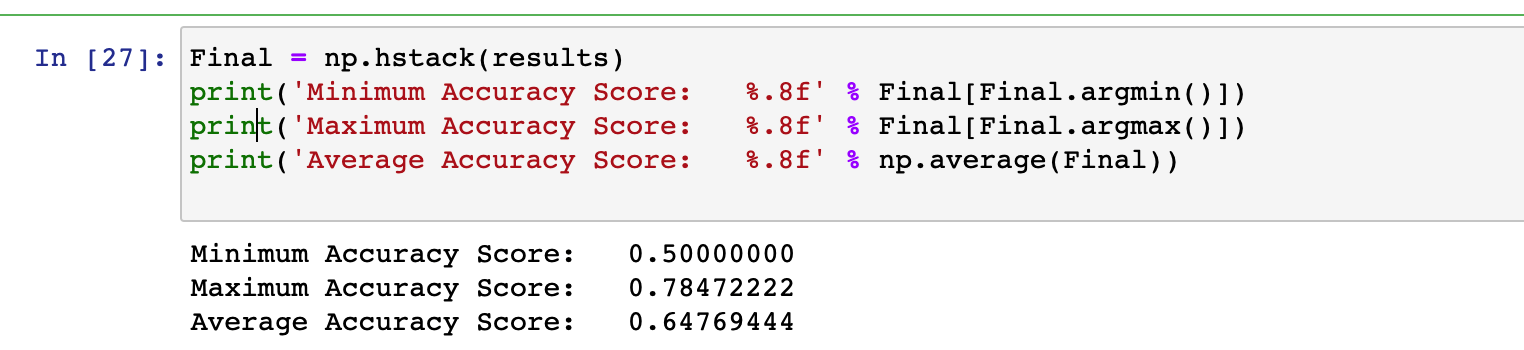
**5. 訓練與預測**

**5.1使用Perception分類器**

1. perc = Perceptron(eta0=0.1, random\_state=15)
3. # The ratio of data\_train to data\_test is 7:3
4. results=[]
6. # Make multiple predictions
7. **for** \_ **in** range(1000):
8. # Randomly generate a data set of 0.7
9. data\_train = data.sample(frac=0.7)
10. # Dataset label
11. data\_train\_X = data\_train.loc[:, **lambda** x: [l **for** l **in** data **if** l != "Class"]]
12. data\_train\_Y = data\_train.loc[:, **lambda** x: "Class"]
14. # The rest is the test dataset
15. data\_test = data.loc[~data.index.isin(data\_train.index)]
16. # Test dataset label
17. data\_test\_X = data\_test.loc[:, **lambda** x: [l **for** l **in** data **if** l != "Class"]]
18. data\_test\_Y = data\_test.loc[:, **lambda** x: "Class"]
20. # Use the .fit() for training
21. perc.fit(data\_train\_X, data\_train\_Y)
22. # Accuracy of prediction results
23. results.append(perc.score(data\_test\_X, data\_test\_Y))
25. plt.plot([\*range(0,1000)],results)



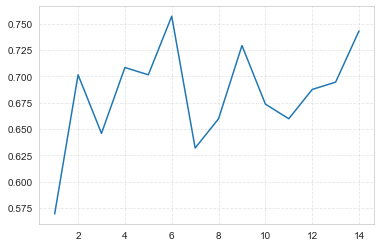
1. Final = np.hstack(results)
2. **print**('Minimum Accuracy Score:   %.8f' % Final[Final.argmin()])
3. **print**('Maximum Accuracy Score:   %.8f' % Final[Final.argmax()])
4. **print**('Average Accuracy Score:   %.8f' % np.average(Final))



Perception 分類器平均預測結果準確率是0.65。

**5.2使用決策樹分類器並繪製不同深度決策樹準確率折線圖**

1. # The ratio of data\_train to data\_test is 7:3
3. results2 = []
5. **for** i **in** range(1,15):
6. # Training dataset
7. data\_train = data.sample(frac=0.7)
8. # label
9. data\_train\_X = data\_train.loc[:, **lambda** x: [l **for** l **in** data **if** l != "Class"]]
10. data\_train\_Y = data\_train.loc[:, **lambda** x: "Class"]
12. # Test dataset
13. data\_test = data.loc[~data.index.isin(data\_train.index)]
14. # label
15. data\_test\_X = data\_test.loc[:, **lambda** x: [l **for** l **in** data **if** l != "Class"]]
16. data\_test\_Y = data\_test.loc[:, **lambda** x: "Class"]
18. # Build decision trees of different depths
19. tree = DecisionTreeClassifier(random\_state=56, criterion='gini', max\_depth=i)
20. # Use the .fit() for training
21. tree.fit(data\_train\_X, data\_train\_Y)
22. # result
23. results2.append(tree.score(data\_test\_X, data\_test\_Y))
25. plt.plot([\*range(1,15)],results2)
26. plt.grid(True, linestyle='--', alpha=0.5)



輸出最大值及對應深度：

1. **print**(max(results2),results2.index(max(results2)))



綜上，對比兩個不同分類器得出的結果，深度為 6 的決策樹分類器經過訓練之後預測的結果更準確。

**結論：**

從我們的得出的結果來看，訪問課程內容的次數、缺席天數、在課上有舉手的次數、檢查新公告的次數、是否參加討論、性別和學期確確實實是影響學生學業成績的因素。