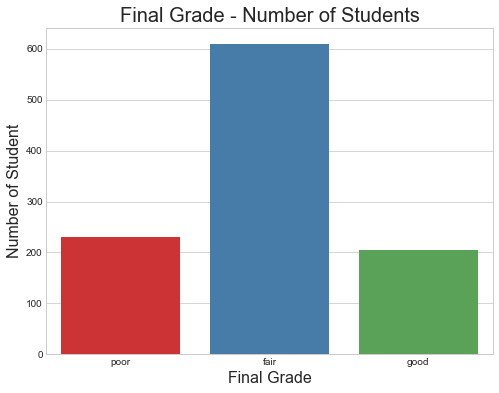
**Một Số Nhận Xét Khi Phân Tích Dữ Liệu**

**I. Phân loại theo các thuộc tính**

1. *Phân loại theo điểm cuối kì*

Trước tiên, ta cần phân loại theo điểm cuối kì để biết được dataset đang phân tích. Số lượng của từng loại học sinh.

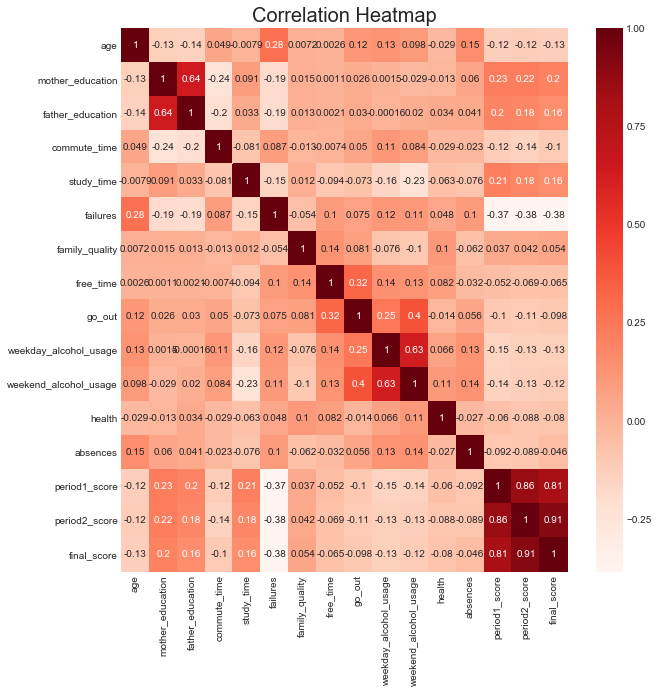


Từ đó nhận ra được khoảng ***600 học sinh*** đạt loại Fair, khoảng 225 học sinh đạt loạt loại Poor, khoảng 200 học sinh đạt loại Good. Phần lớn học sinh đạt loại ***Fair khoảng 60 %*** trên dataset.

*Tóm lại, phần này giúp ta biết khái quát về dataset.*

1. *Vẽ biểu đồ nhiệt các thuộc tính với nhau (thuộc tính = cột dữ liệu)*

Tiếp theo, ta cần xét dataset có các thuộc tính nào có độ tương quan (độ liên quan) cao có khả năng ảnh hưởng đến điểm cuối kì.



Từ biểu đồ nhiệt ta dễ dàng thấy được từ những điểm có màu đậm nhất và nhạt nhất để ***focus*** vào những thuộc tính đó và phân tích theo những thuộc tính đó.

* Kỳ 1, 2 điểm tương quan cao với kết quả cuối cùng (lần lượt là 0,81 và 0,91). - Điểm kỳ 1 tương quan với điểm 2 (0,86).
* Giáo dục của mẹ tương quan với giáo dục của cha (0,64).
* Thường xuyên tiêu thụ rượu vào các ngày trong tuần có nhiều khả năng tiêu thụ rượu vào cuối tuần (0,63).
* Đi chơi thường xuyên có cơ hội tiêu thụ nhiều rượu hơn vào cuối tuần (0,4).
* Có nhiều thời gian rảnh có xu hướng đi ra ngoài thường xuyên hơn (0,32). Tuổi tương quan với thất bại (0,28).
* Học sinh có mẹ có trình độ học vấn cao có xu hướng hoạt động tốt hơn trong giai đoạn 1, 2 và cuối cùng (lần lượt là 0,23, 0,22 và 0,2).
* Dành nhiều thời gian hơn cho việc học mang lại cơ hội tốt hơn để đạt điểm cao hơn trong giai đoạn 1, 2 và cuối cùng (0,21, 0,18 và 0,16)

Từ đó ta cần focus vào những vấn đề có khả năng ảnh hưởng đến điểm cuối kì như romantic relationship, alcohol consumption, parent’s education, frequency of going out, desire to go to college, living area.

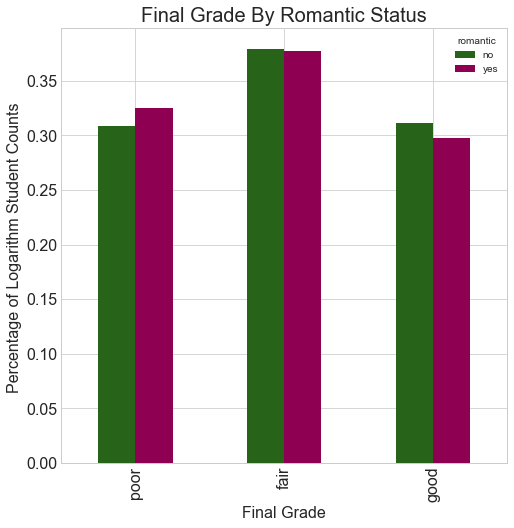
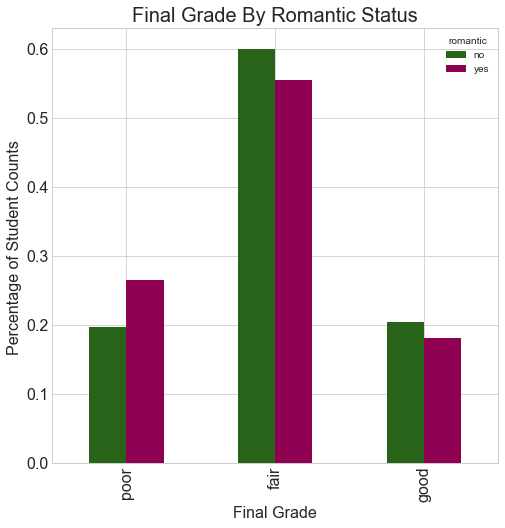
*Tóm lại, ta dùng biểu đồ nhiệt phát hiện các thuộc tính có liên quan với nhau từ đó có thể kết hợp chúng lại để phân tích chúng có sự ảnh hưởng gì đến điểm cuối kì.*

1. *Phân loại điểm cuối kì theo romantic status*

Ở phần này, ta cần focus vào *tỉ lệ của từng loại học sinh*. Có in romantic relationship hoặc không. Ta cần focus vào tỉ lệ do cần đánh giá một tổng thể lớn hơn. Và dataset của ta có thể ***không cân bằng về số lượng người có in relationship hoặc không.*** Ta cần giảm độ dốc của của đồ thị biểu (đối với các biến liên tục), độ chênh lệch chiều cao (đối với các biến rời rạc). Bằng cách lograric lại dataset. *Hàm logarithm là hàm đặc biệt nó vào không thay đổi tính chất của hàm số (đối với các hàm số có miền giá trị dương)*.

|  |  |  |
| --- | --- | --- |
| Relationship | No | Yes |
| Fair | 404 | 206 |
| Good | 137 | 67 |
| Poor | 132 | 98 |

Bảng số lượng phân loại học sinh theo relationship



Xem qua hai biểu đồ ta thấy tỉ lệ học sinh người không in romantic relationship nhiều hơn người có in romantic relationship.

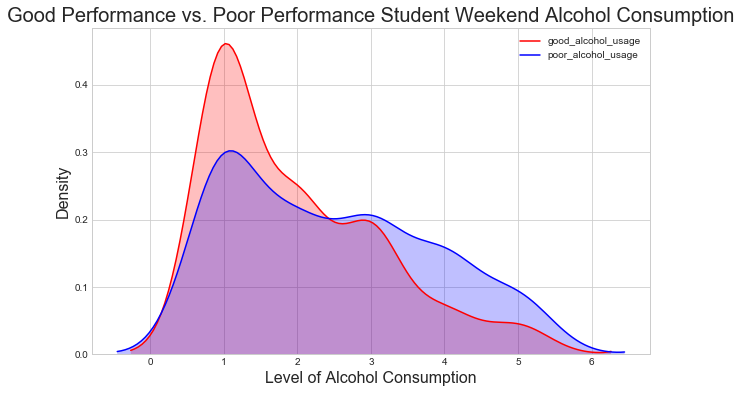
Để chắc chắn hơn với kết luận trên in romantic relationship ảnh hưởng đáng kể đến kết quả cuối kì cần xét thêm phân phối chi bình phương. Phân phối chi bình phương là một phân phối đủ tốt để xét ảnh hưởng 2 tập.

Kết quả chi bình phương: *0.038810128743959726.* Kết quả giúp có cơ sở hơn để khẳng định là in romatic relationship có ảnh hưởng đáng kể đến kết quả cuối kì.

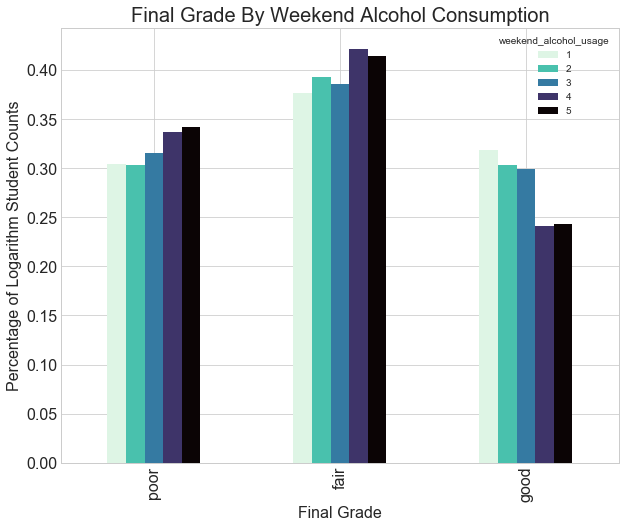
*Tóm lại, phần này giúp biết tại sao phải dùng hàm logaricthm và tỉ lệ logaricthm của các loại để làm gì, phân phối chi bình phương.*

1. *Phân loại điểm cuối kì theo alcohol consumption*

Như trên, để tránh ảnh hưởng do dataset không cần bằng. Ta cần focus vào tỉ lệ. Do dataset này đa phần là học sinh fair và bỏ học sinh fair để xem sự ảnh hưởng do alcohol consumption học sinh poor và good là như thế nào. Do chỉ còn hai loại good, poor nên vẽ biểu đồ mật độ xác suất với biến level of alcohol consumption.



Từ biểu đồ có thể thấy được, những học sinh good hàm mật độ có đỉnh vị trí khoảng level 1. Tức là mật độ học sinh good ở level 1 là cao nhất. Tương tự, học sinh poor hàm mật độ có đỉnh vị trí khoảng level 1. Có nghĩa là mật độ học sinh poor ở level 1 là cao nhất. Giao điểm 2 hàm mật độ là khoảng level 2.5.Tức là những người có alcohol consumption dưới 2.5 có khả năng cao là học sinh Good. Alcohol consumption trên 2.5 có khả năng cao là học sinh Poor.



Từ biểu đồ có thể thấy được, ở học sinh giỏi tỉ lệ lograricthm giảm dần theo chiều tăng của alcohol consumption. Nếu bạn có alcohol consumption ở mức 4, 5 thì ***theo percentage of logaricthm student counts khoảng 75%*** không là học sinh good. Nếu bạn có alcohol consumption ở mức 1, 2, 3 thì ***theo percentage of logaricthm student counts 70%*** là học sinh fair or good. Tỉ lệ này khá cao đối với ***percentage of logaricthm student counts.***

Để chắc chắn hơn và xét thêm hàm chi bình phương. Kết quả chi bình phương là *0.0075922940563682983*.Và kết quả cho thấy, sử dụng rượu bia ảnh hưởng đáng kể đến kết quả cuối cùng.

*Tóm lại, phần này giúp biết tại sao phải dùng hàm mật độ khi nào, và tỉ lệ logaricthm của các loại và phân phối chi bình phương.*