Báo cáo 15/03/24

Mục tiêu tuần vừa qua:

Tìm mối tương quan của yếu tố môn học với quyết định nghỉ học của sinh viên

Công việc đã làm

1. Tìm và xử lý dữ liệu để lấy được bộ dữ liệu mới
2. Plot 1 vài chart tìm sự tương quan giữa các yếu tố và quyết định nghỉ học

Chi tiết kết quả các công việc đã thực hiện

Toàn bộ các dữ liệu được tập trung vào các đặc điểm sau:

- Sinh viên bắt đầu nhập học từ năm 2018 (Do trước đó dữ liệu không đầy đủ để sử dụng)
- Sinh viên thuộc các ngành CNTT (ngành ứng dụng phần mềm, ngành lập trình web, xử lý dữ liệu,... đều sẽ được xếp vào ngành CNTT)
- Chỉ cắt dữ liệu của kỳ 1, 2 và kỳ 3 (dữ liệu kỳ 1,2 dự đoán trạng thái kỳ 3)

Dữ liệu:

Bộ dữ liệu gồm có:

- Dữ liệu lịch sử trạng thái sinh viên
- Dữ liệu điểm trung bình của sinh viên của cả kỳ 1 + 2
- Dữ liệu % điểm danh của cả kỳ 1 + 2
- Dữ liệu điểm trung bình theo từng nhóm môn (tiền tố, ví dụ COM101, COM211 sẽ gom chung thành nhóm môn COM)
- Dữ liệu về môn học theo các kỳ và khung chương trình

Dữ liệu lịch sử trạng thái sinh viên (student_status_history)

Các cột dữ liệu bao gồm:

- **student_code:** Mã sinh viên (mỗi sinh viên sẽ có 1 mã duy nhất, trừ trường hợp sinh viên chuyển ngành với mã số mới hoặc nhập học từ đầu sẽ được cấp mã số mới)
- **semester**: Kỳ thứ của sinh viên (kỳ học). Tại Poly có nhiều khung chương trình đào tạo, có thể có khung chỉ có 2 kỳ là hoàn thành chương trình, có khung có 6 kỳ, có

khung có 7 kỳ (Cùng 1 ngành nhưng có thể có 2 khung chương trình khác nhau được triển khai với số lượng kỳ là 6 và 7) . Trong dữ liệu lịch sử có thể thấy có sinh viên đạt đến trên kỳ 7 vì đó là sinh viên bị nợ môn nên kỳ 8 được định nghĩa với sinh viên ở trạng thái học quá thời gian quy định nhưng chưa thể tốt nghiệp.

- **major_name:** Tên ngành học của sinh viên
- **campus_code**: Cơ sở đào tạo
- **term_name**: Kỳ triển khai
- **status**: Trạng thái học của sinh viên (THO là nghỉ học, HDI là trạng thái lên kỳ bình thường (đủ điều kiện lên kỳ) ngoài ra có TN là tạm ngưng (bảo lưu), ngoài ra còn có HL, CHO – chờ xếp lớp). Tạm thời chỉ lấy THO là nghỉ học còn lại sẽ là đi học nói chung (trong báo cáo này sẽ gọi chung là ĐH)

student_status_history_raw1.csv

File này là dữ liệu toàn bộ sinh viên chỉ cắt ra trong khoảng năm 2018 đến nay và chỉ cắt từ kỳ 1 đến kỳ 3

→ Sau khi quan sát thì nhận thấy rằng 1 sinh viên có thể chuyển nhiều ngành với cùng 1 mã số sinh viên, có thể chuyển nhiều cơ sở khác nhau trong quá trình học.

→Nhận thấy rằng yếu tố trên có thể ảnh hưởng đến quyết định nghỉ học của sinh viên, nên đã thực hiện cắt bỏ những sinh viên chuyển ngành khác CNTT trong quá trình học và những sinh viên có chuyển cơ sở học

student_status_history.csv

Đây là file sau khi đã thực hiện lọc đầy đủ và sẽ sử dụng để plot tìm mối tương quan

Dữ liệu này chưa xử lý được trường hợp sinh viên có sự chuyển ngành nhưng đổi mã số mới hoặc nhập học lại nhưng đổi mã số mới.

Dữ liệu này có tổng cộng 10073 sinh viên trong đó có 877 sinh viên có quyết định nghỉ học ở kỳ 3

Vấn đề nhận thấy nhưng chưa xử lý được ở dữ liệu:

- Có trường hợp sinh viên học xong kỳ 3 (ở ngành CNTT sau đó nghỉ học 1 năm rồi sau đó nhập học lại ĐH với trạng thái HDI ở ngành khác ví dụ PH07811)
- Có TH có 2 bản ghi kỳ 3, 1 bản ghi đầu là HDI, bản ghi sau là THO → Giải thích đối với TH này như sau: sinh viên sau khi học xong kỳ 2 được xét đủ điều kiện lên kỳ 3 nên sinh viên học kỳ 3 với trạng thái chính thức HDI nhưng sau khi học kỳ 3, sinh viên không đủ điều kiện lên kỳ 4 nên vẫn ở kỳ 3, trong thời gian này, sinh viên đăng ký nghỉ

học → kỳ 3 thứ 2 của sinh viên là THO. Với TH này vì dự đoán kỳ liền tiếp sau kỳ 2 của sinh viên nên sẽ tính là sinh viên ĐH.

Dữ liệu điểm trung bình kỳ 1 + 2 (avg_grade.csv)

Là dữ liệu điểm trung bình từ khi nhập học của sinh viên cho đến khi sinh viên lên kỳ 3

Dữ liệu này được tính theo tổng toàn bộ điểm trung bình mỗi môn của sinh viên với số tín chỉ rồi chia cho toàn bộ số tín chỉ đã học của sinh viên

Các cột dữ liệu bao gồm

- user_code: Mã sinh viên
- total_grade: Tổng điểm
- total_credit: Tổng tín chỉ
- average_grade: Điểm trung bình
- *Các cột khác lấy ra để check và map dữ liệu*

Dữ liệu điểm trung bình theo nhóm môn (tiền tố) (avg_grade_by_prefix.csv)

Nhóm môn này là những môn có chung tiền tố như WEB, COM, ENT. Các môn trong 1 nhóm môn là các môn về cùng 1 lĩnh vực ví dụ tiền tố web đều là làm web nhưng có thể công nghệ, mảng web khác nhau (frontend, backend), cũng có thể là các level trong cùng 1 môn.

Các cột dữ liệu bao gồm

- user_code: Mã sinh viên
- semester: Kỳ học
- average_grade: Điểm trung bình
- prefix_subject: Tên nhóm môn
- *Các cột khác lấy ra chỉ để check và map dữ liệu*

Nhận thấy rằng các nhóm môn có thể được học ở kỳ 1, có nhóm học ở kỳ 2 cũng có nhóm học ở cả 2 kỳ

Dữ liệu điểm danh (attendance.csv)

Là dữ liệu % điểm danh của cả 2 kỳ 1 và 2 của sinh viên (tổng điểm danh có mặt/tổng thời buổi học)

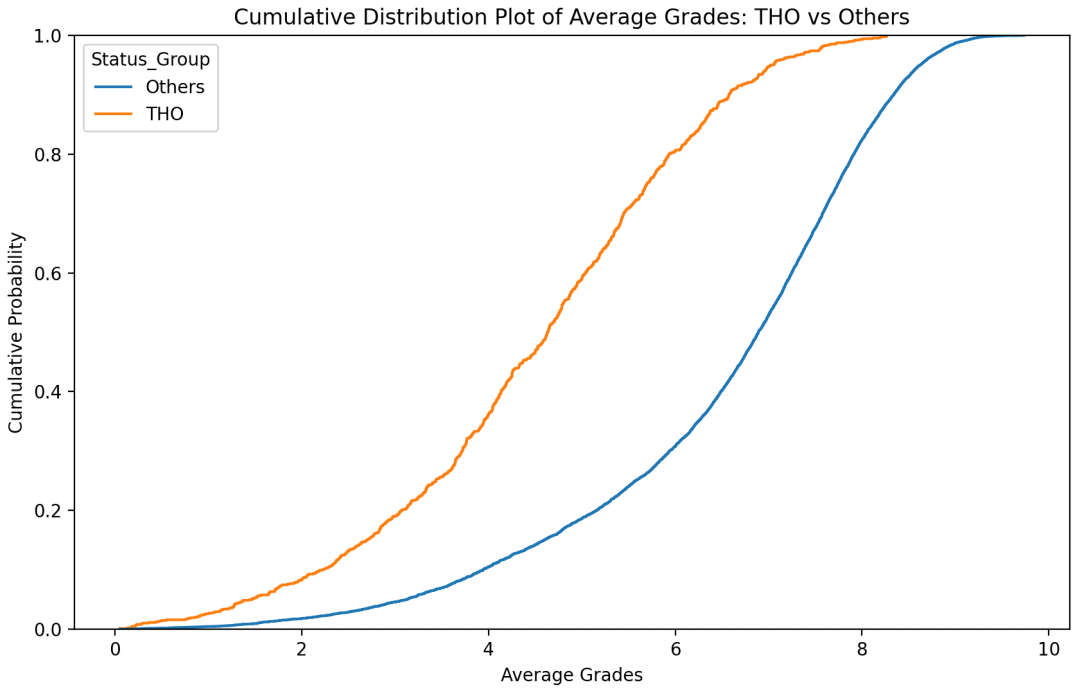
Các cột dữ liệu bao gồm

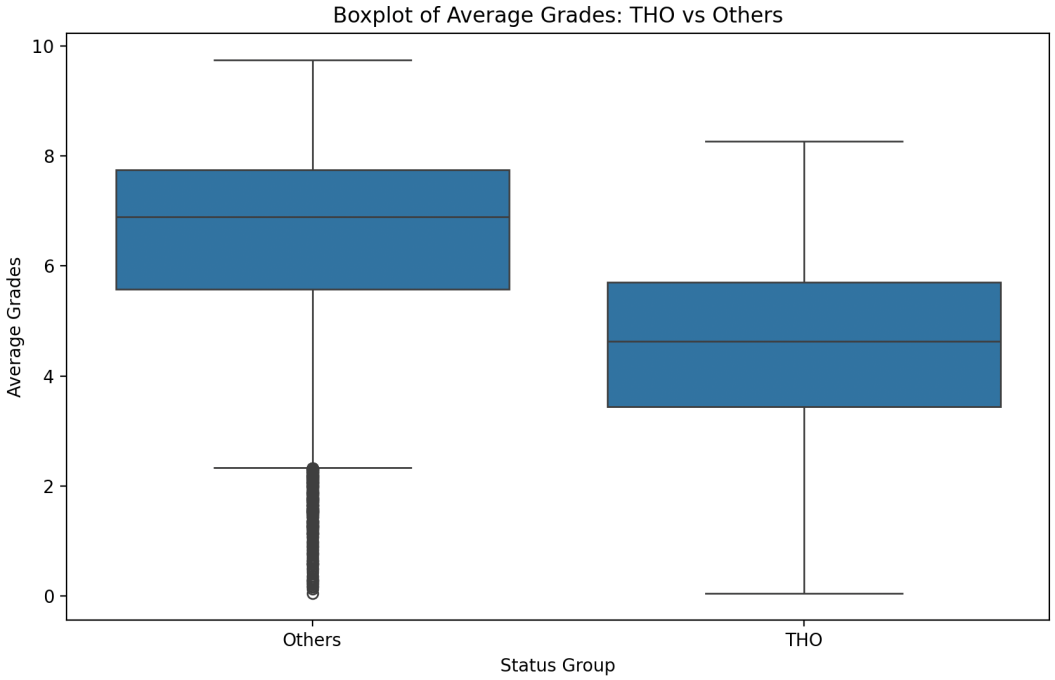
- User_code: Mã sinh viên
- Total_attendance: tổng có mặt
- Total_activity: tổng số buổi học
- Percentage: tỉ lệ điểm danh

Dữ liệu chương trình học (dùng để tham khảo) (study_plan)

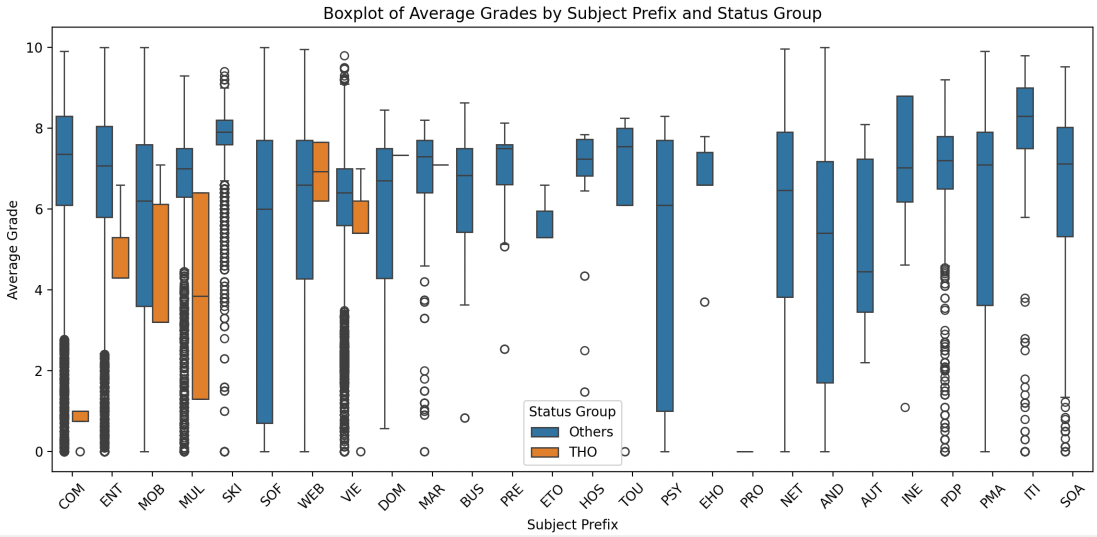
- Major: Ngành
- Specialized_major: Chuyên ngành
- Period_name: Kỳ học
- Subject_name: Tên môn
- Subject_code: Mã môn
- Num_of_credit: Số tín chỉ

Các charts đã plot được

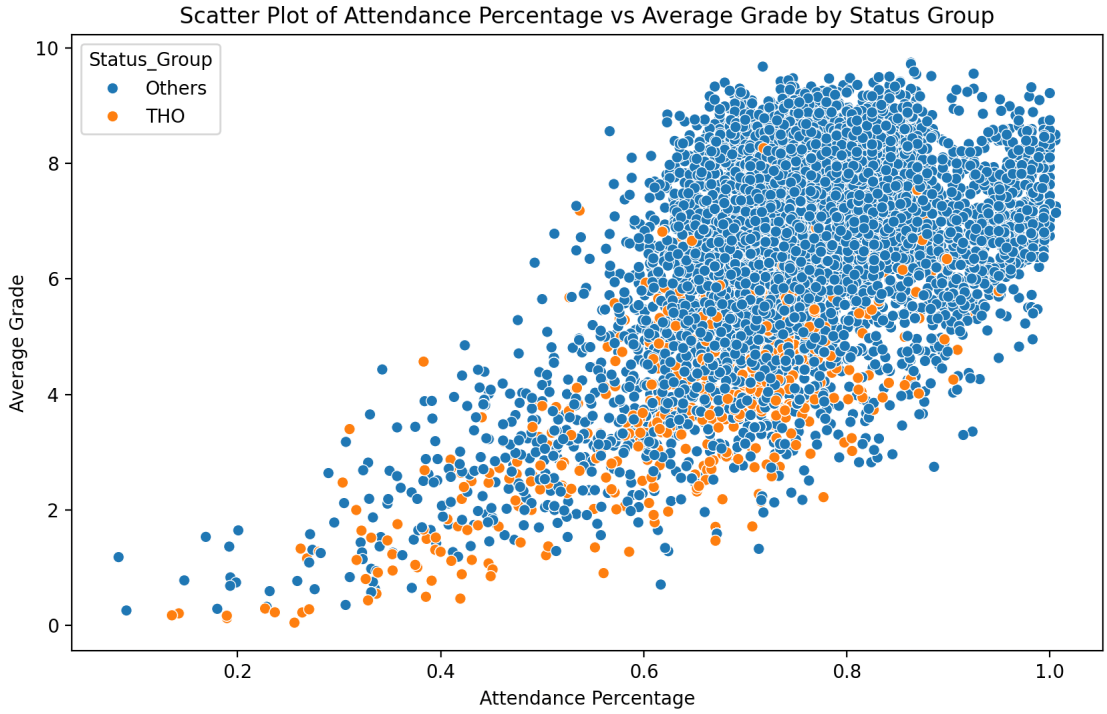




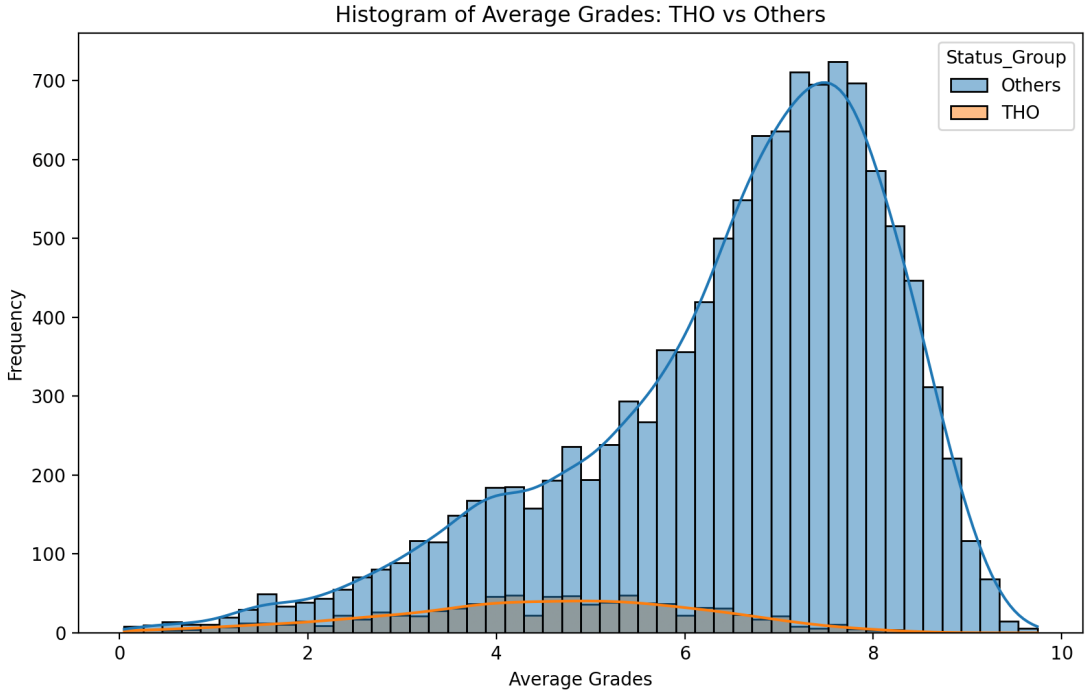
→ Điểm tổng trung bình có tính tương quan khá rõ để sử dụng trong mô hình

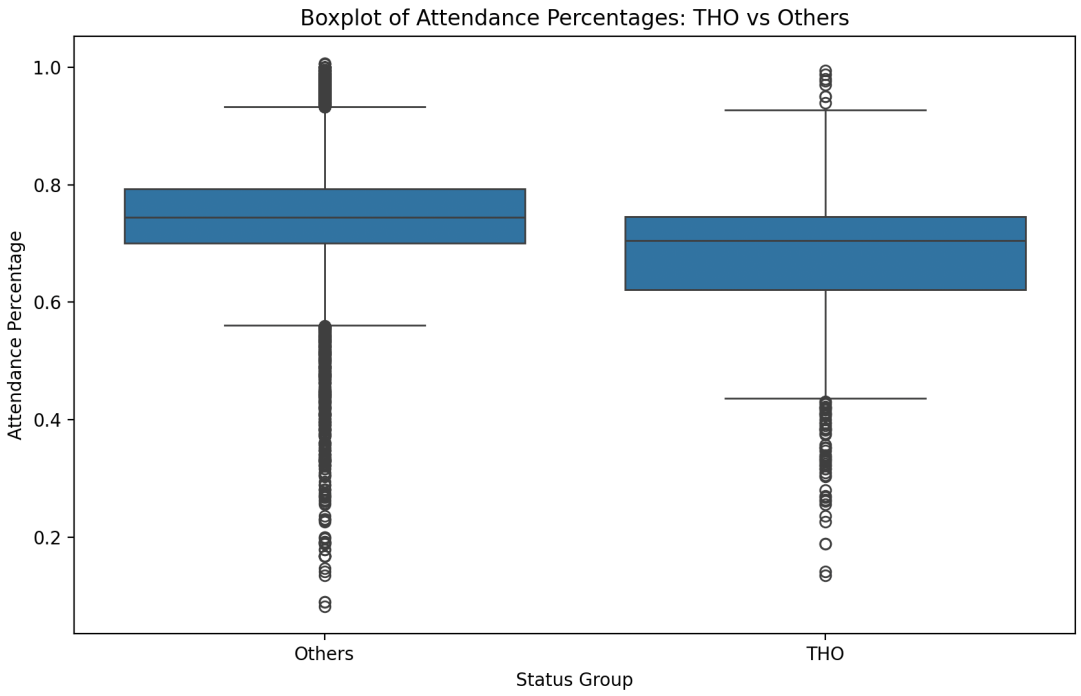


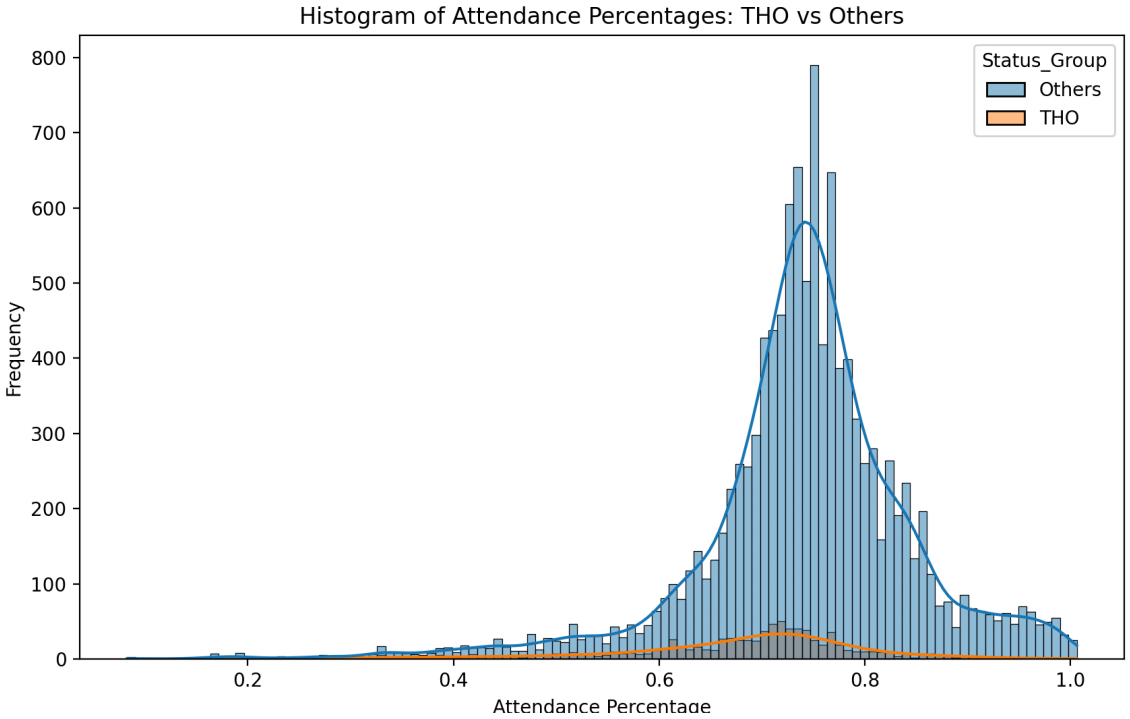
→ Nhóm môn MUL (các môn liên quan đến thiết kế) có tính tương quan khá rõ với quyết định nghỉ học của sinh viên

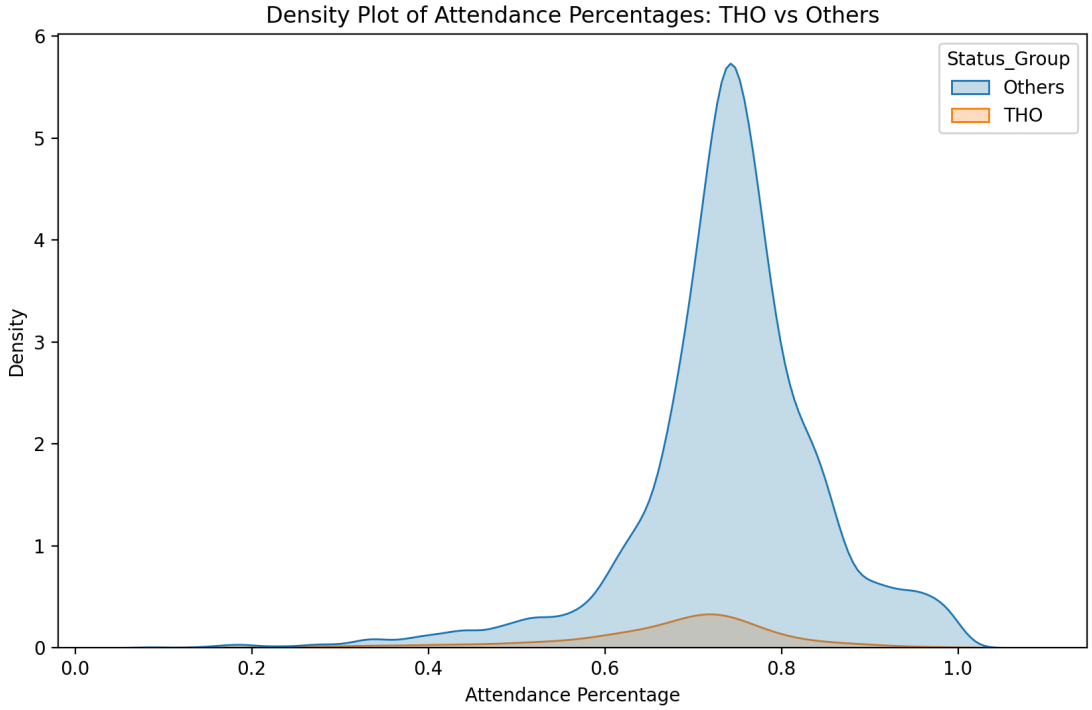


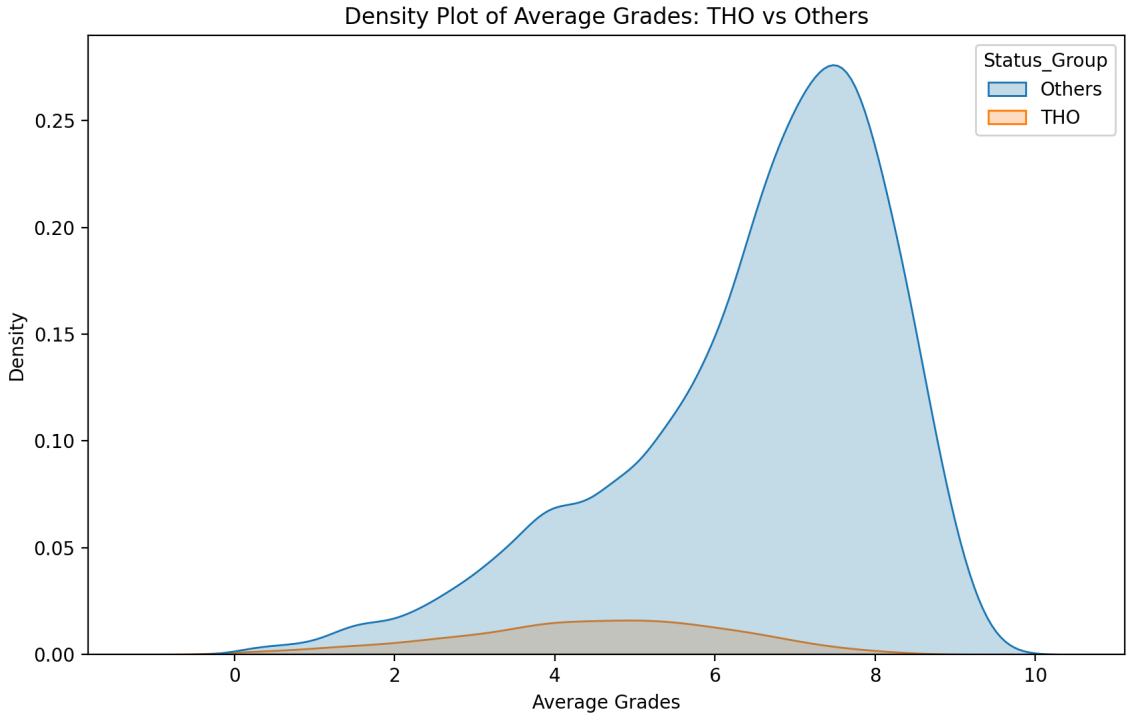
→ Tỉ lệ điểm danh chưa thực sự rõ rệt

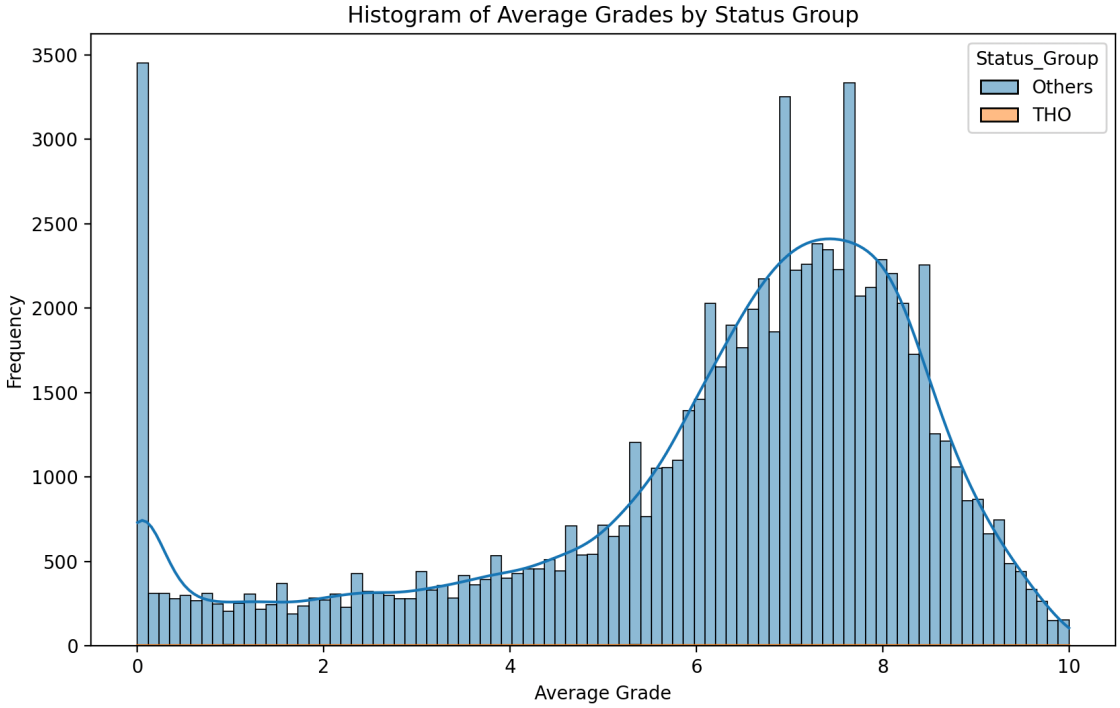

Histogram of Average Grades: THO vs Others
Status_Group
Others
THO
Frequency
Average Grades
0
100
200
300
400
500
600
700
0
2
4
6
8
10


Boxplot of Attendance Percentages: THO vs Others
Attendance Percentage
Status Group
Others
THO
0.2
0.4
0.6
0.8
1.0


Histogram of Attendance Percentages: THO vs Others
Status_Group
Others
THO
Frequency
Attendance Percentage
0
100
200
300
400
500
600
700
800
0.2
0.4
0.6
0.8
1.0


Density Plot of Attendance Percentages: THO vs Others
Status_Group
Others
THO
Density
Attendance Percentage
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
1.0


Density Plot of Average Grades: THO vs Others
Status_Group
Others
THO
Density
0.00
0.05
0.10
0.15
0.20
0.25
0
2
4
6
8
10
Average Grades


Histogram of Average Grades by Status Group
Status_Group
Others
THO
Frequency
0
500
1000
1500
2000
2500
3000
3500
0
2
4
6
8
10
Average Grade