You are given two data sets.

**StudentClass.csv** this is students’ performance in 10 core classes after joining college.

**StudentGraduate.csv** this is the students’ background information, and if this student graduate on time.

The ID of this two data set can be matched.

Data description:

ID ：

HS\_ENGLISH ： Number of semesters taking English in High school

HS\_MATH： Number of semesters taking Math in High school

HS\_HISTORY ： Number of semesters taking History in High school

HS\_LAB\_SCIENCES： Number of semesters taking Lab Science in High school

HS\_FOREIGN\_LANGUAGE： Number of semesters taking Foreign Language in High school

HS\_ART： Number of semesters taking Art in High school

HS\_ELECTIVES： Number of semesters taking Electives in High school

HS\_GPA: High school GPA

sex

WRITINGScore: SAT writing score

GraduateONTime : 1 graduate on time. 0 not on time

Major:

Research question:

1. Do students perform better in Morning class (before 12:00pm) or afternoon class?
2. Math, Physics and Statistics are considered quantitative classes. If a student take 2 or more quantitative classes in one semester, will he/she have lower average score?
3. Which Major has higher quantitative class GPA?
4. Can you combine two dataset and build a classification model to predict if a student can Graduate ON Time (Note: We cannot use the last 2 semester’s performance to predict the graduation, since it is too late to predict the graduation)