Machine Learning pa1
CAPP 30254
Sirui Feng
siruif@uchicago.edu

Problem A

1.
Field Name: First_name
Mode: Amy
Missing Value Count: 0

Field Name: Last_name
Mode: Ross
Missing Value Count: 0

Field Name: State
Mode: Texas
Missing Value Count: 116

Field Name: Gender
Mode: Female
Missing Value Count: 226

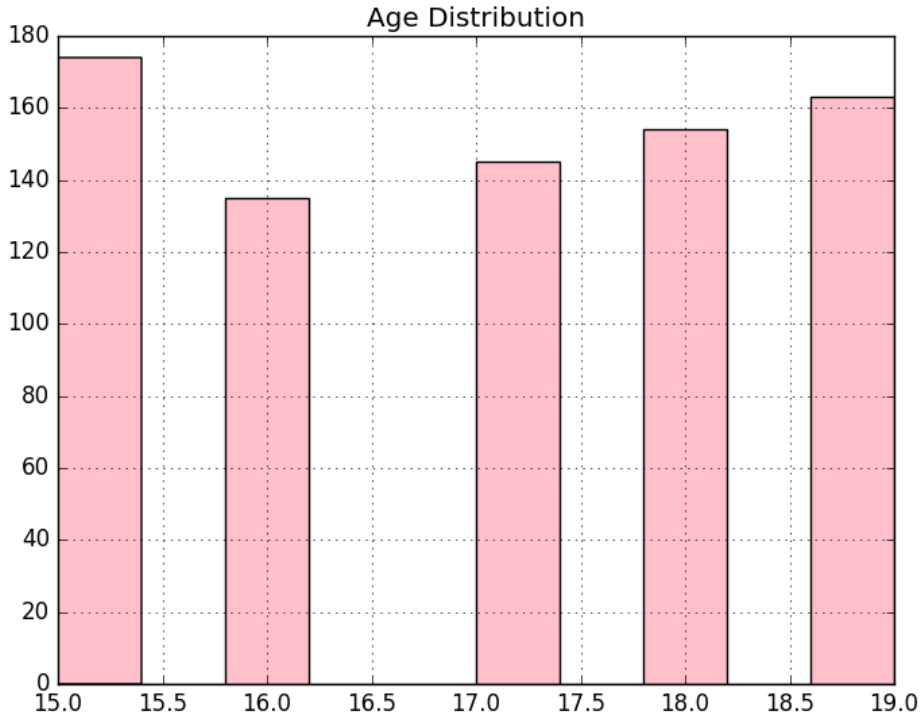
Field Name: Age
Mean: 17.0
Standard Deviation: 1.46
Median: 17.0
Mode: 15
Missing Value Count: 229

Field Name: GPA
Mean: 2.99
Standard Deviation: 0.82
Median: 3.0
Mode: 2
Missing Value Count: 221

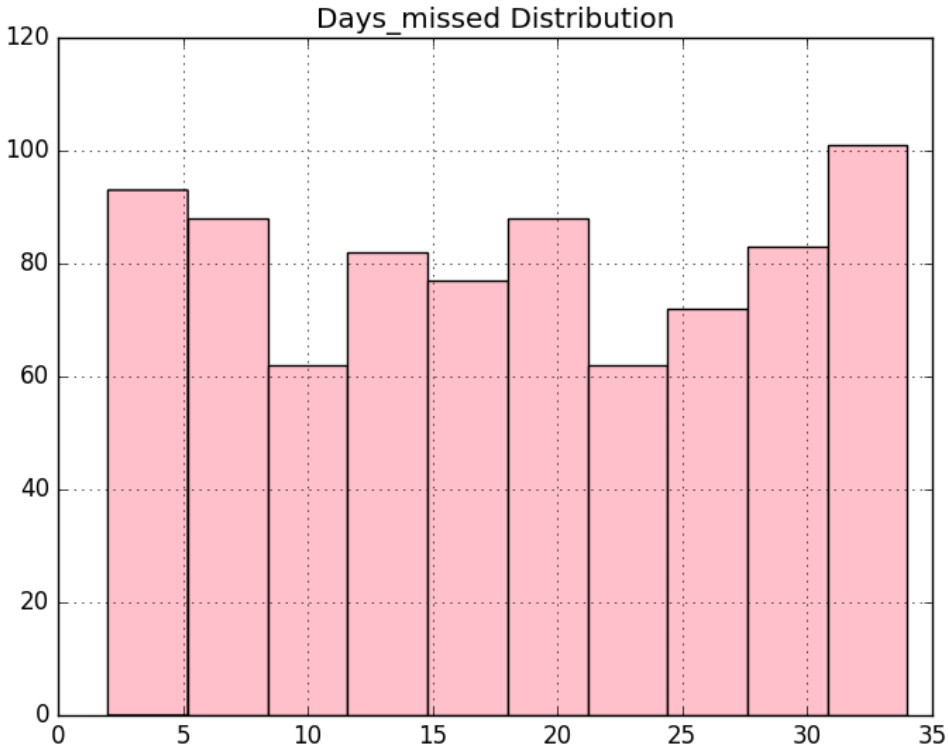
Field Name: Days_missed
Mean: 18.01
Standard Deviation: 9.63

Median: 18.0
Mode: 6 14 31
Missing Value Count: 192

Field Name: Graduated
Mode: Yes
Missing Value Count: 0




GPA Distribution
300
250
200
150
100
50
0
2.0
2.5
3.0
3.5
4.0


Days_missed Distribution
120
100
80
60
40
20
0
0
5
10
15
20
25
30
35

Problem B

A. They have the same probability of graduating. Adam and Chris share same characteristics and differ only by 10,000 of family income; Bob and David also share identical characteristics except for 10,000 of family income. Thus, we should expect the difference of likelihood between Adam and Chris and the difference between Bob and David being the same. And because Adam and Bob have the same probability, we should expect Chris and David have the same probability of graduation.
B. Holding other characteristics constant, an African American male student are less likely to graduate compared to male students and African American students. This does not imply that African American males are more likely to not graduate than African American females. Similarly, we need more information to compare African American males and non African American males.
C. The effect of age on the probability of graduation depends on one's age. Specifically, in this model, the variables age and age squared allow age to have a parabola effect on the likelihood of graduation – below a threshold, an increase of age is associated with a decrease in graduation probability; above that threshold, an increase of age is associated with an increase in graduation probability.
D. I would drop male or female. Because to show the gender effect, one of them should be left out as a base case. I would need more information about the categories of gender, i.e. whether there are more categories other than female and male, if yes, the model existing is appropriate; otherwise, I will drop one of the two variables: female or male.