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

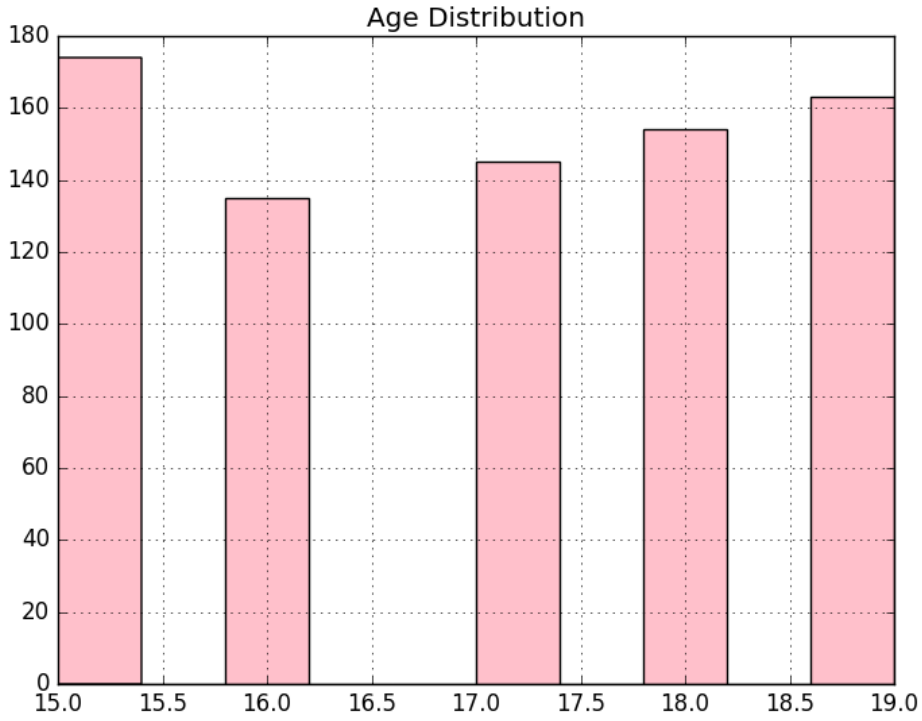
Field Name: Age
Mean: 17.0
Standard Deviation: 1.46
Median: 17.0
Mode: 15
Missing Value Count: 229

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

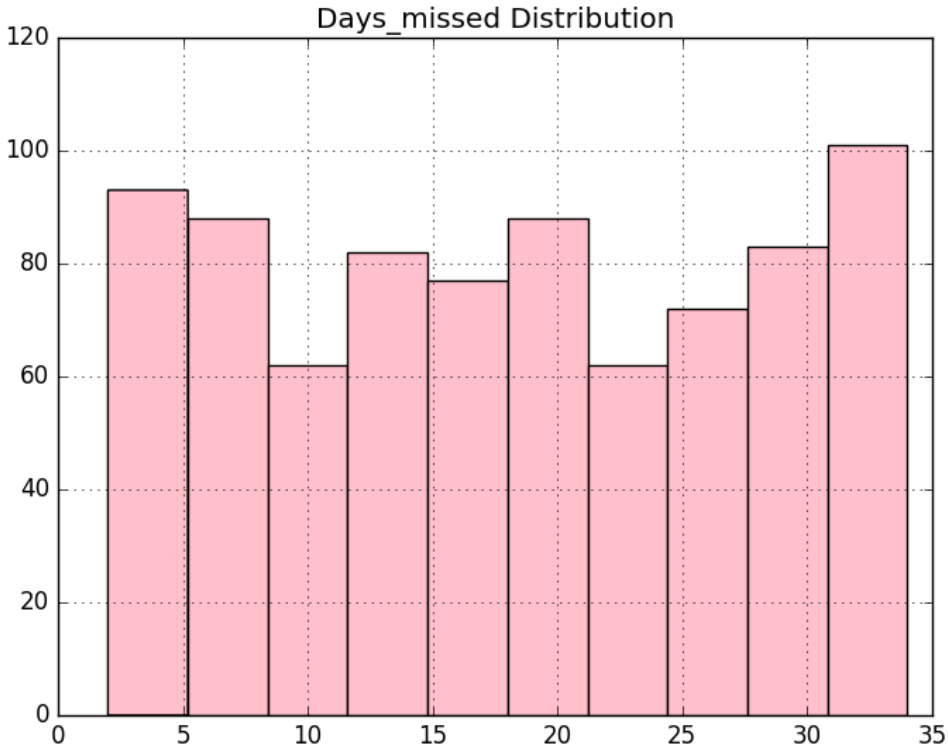
Field Name: Days_missed
Mean: 18.01
Standard Deviation: 9.63

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. Cannot tell based on the information provided. In order to infer the comparison of Chris and David from a logistic regression model with polynomial explanatory variables, we need more information on their other characteristics.
B. African American male students are more likely to graduate compared to African American male students and non African American students (including all gender). This implies that African American males are less likely to not graduate compared to African American females 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 quadratic 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. However, our analysis output indicates that both coefficients are insignificant; thus, there is not much information we can draw from it. In fact, a joint hypothesis has to be employment
D. One might argue dropping male or female. However, 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. Moreover, since both age and age squared are insignificant, I will drop them and run the model.