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

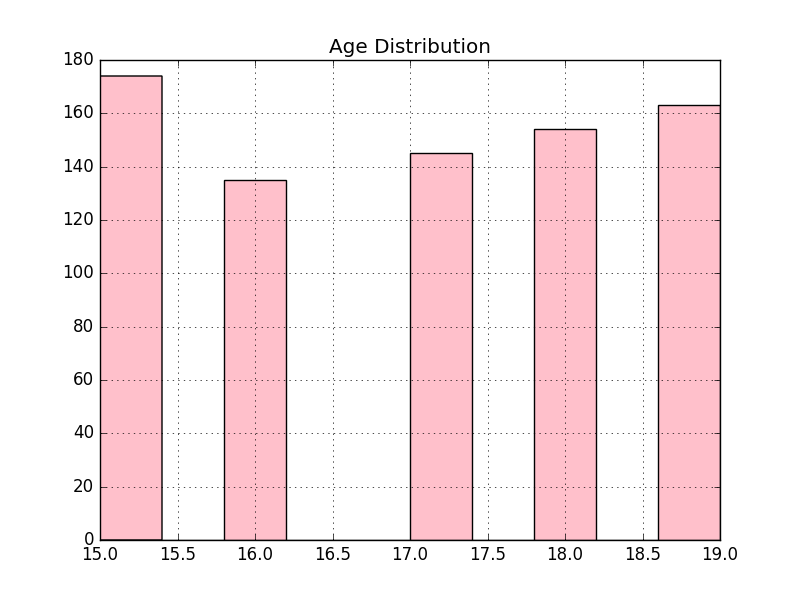
Mode: 6 14 31

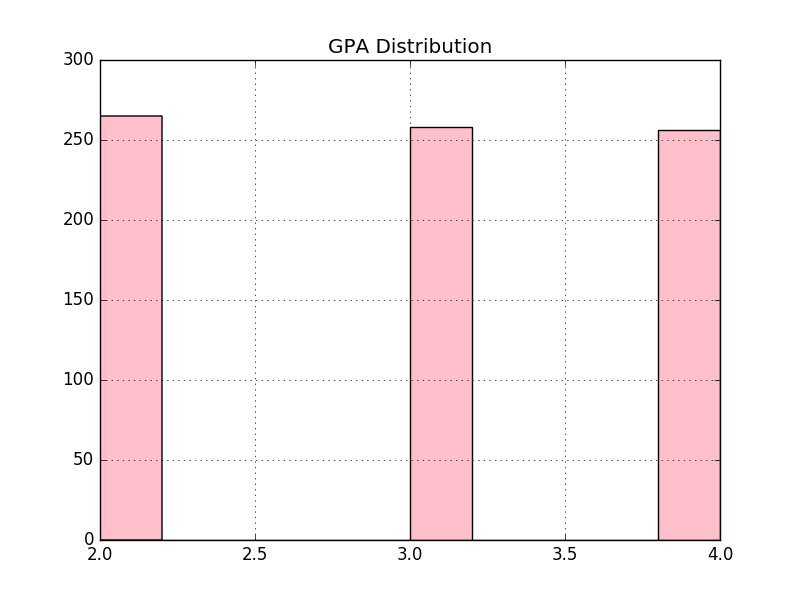
Missing Value Count: 192

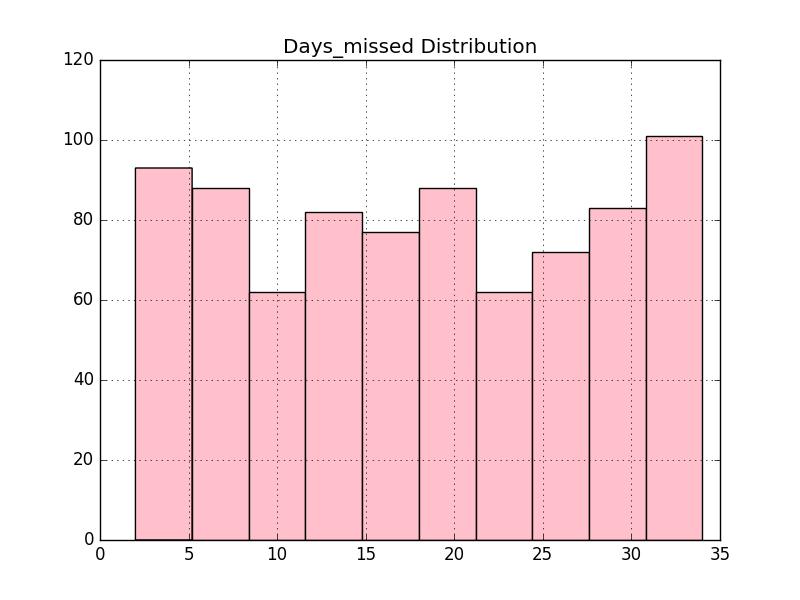
Field Name: Graduated

Mode: Yes

Missing Value Count: 0







Problem B

1. David has a higher probability of graduating. Because the effect of family income on one’s probability of graduating is significantly negative. While Adam and Bob have the same modeled probability of graduation, bob has a higher family income, which implies that compared to Adam, Bob’s other characteristics lead to a higher probability of graduation. For instance, Bob might live in a neighborhood that has a higher percentage of residents who have a college degree; Bob might be an African American while Adam is not.
2. 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.
3. 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.
4. 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. are there more categories other than female and male, if yes, the model existing is appropriate; otherwise, I will drop one of the two variables: female and male.