Problem A

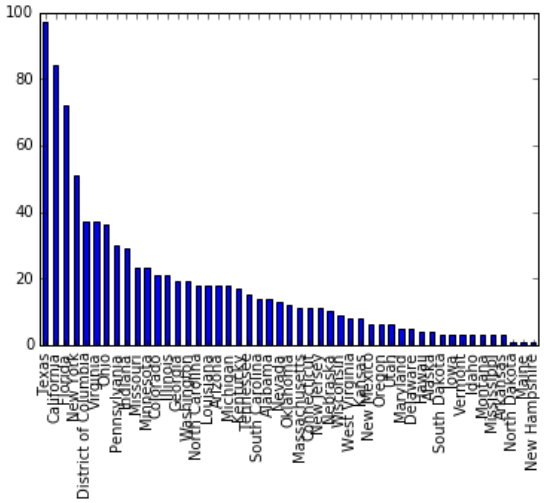
1. Summary Statistics:

Variable	Mean	Median	Mode	Standard Deviation	Missing Values (%)
First_name	N/A	N/A	Amy	N/A	0
Last_name	N/A	N/A	Ross	N/A	0
State	N/A	N/A	Texas	N/A	11.6%
Gender	N/A	N/A	Female	N/A	22.6%
Age	16.9961	17	15	1.45807	22.9%
GPA	2.98845	3	2	0.818249	22.1%
Days_missed	18.0111	18	6	9.62937	19.2%
Graduated	N/A	N/A	Yes	N/A	0

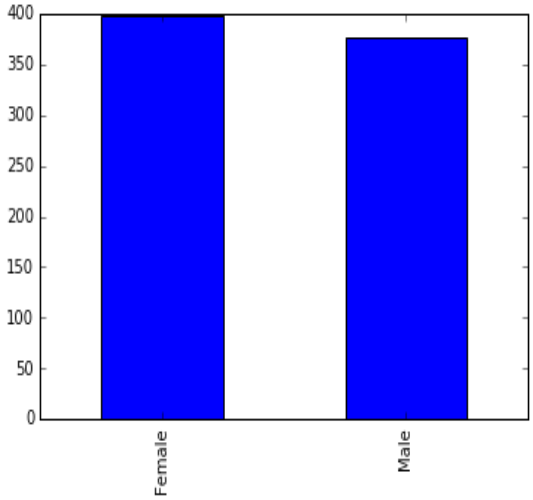
Note on N/A: non-numeric variables don't have mean, median and standard deviation.

Histograms:

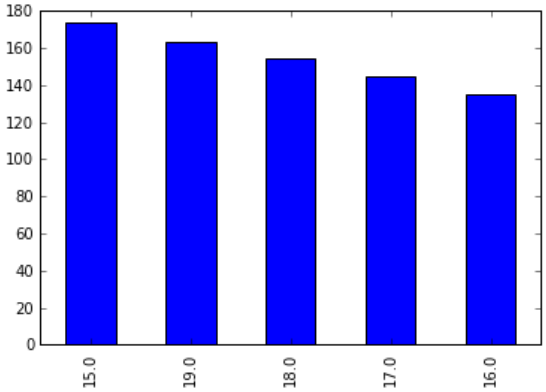
State



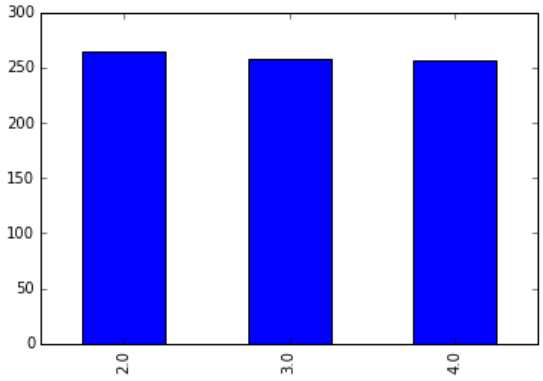
Gender



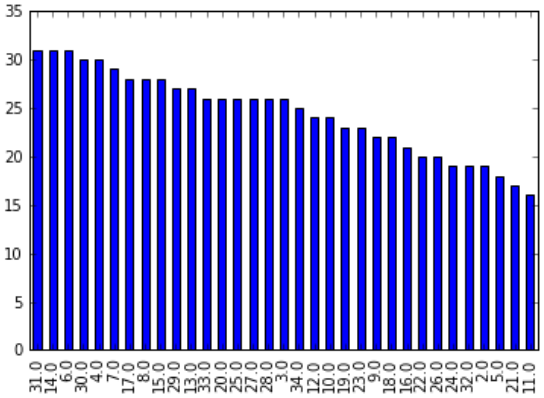
Age



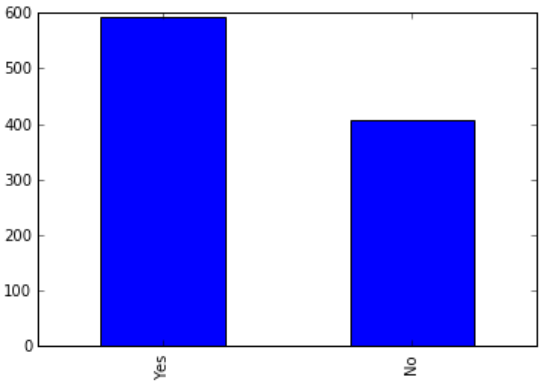
GPA



Days Missed



Graduated



2. I infer the Gender of each student by his/her first name if the data is missing (www.genderize.io). Please click here to see the result (gender values starting with small letters are from inference).

3. Similarly, I fill in the missing values of Age, GPA, and Days missed of each student:
   1) Fill in missing values with the mean of the values for that attribute. Please click here to see the result.
   2) Fill in missing values with a class-conditional mean (where the class is whether they graduated or not). Please click here to see the result.
   3) For a better method, we can fill in missing values with a cluster-mean. In specific, we may infer the missing value of an attribute of a student by the cluster-mean of that attribute, where the cluster contains students with same values in **all other** attributes. However, because sometimes clusters contain no student when using all the attributes, I enlarge the clusters stepwise by eliminating the following attribute one by one: Age, GPA, Days_missed, State, Gender, Graduated. This order is determined by the percentage of missing values of different attributes in descending rank (missing Gender is filled in as above). Please click here to see the result.

Problem B

1. Chris.
   The logit model is defined as:

$$\Pr[y = 1|x, z] = p = \frac{\exp(\alpha + \beta \cdot \ln x + \gamma z)}{1 + \exp(\alpha + \beta \cdot \ln x + \gamma z)}$$

   Taking derivatives with respect to x:

$$\partial p / \partial x = \frac{\beta}{x} \cdot p \cdot (1 - p)$$

   Here, p=probability of graduation, x=income, z=[all other variables], β= -0.109.

   Thus for small changes of x (with p almost unchanged), the effect of family income x to graduating probability p is negatively related to the level of income.

Thinking of Chris as 'an Adam' decreasing income $by 10,000 and David as 'a Bob' decreasing income by $by 10,000, since Adam and Chris has the same level of p, the one with higher family income (Bob) should experience smaller (less positive since $\beta<0$) partial effect of income change. As a result, Chris has higher probability of graduation (increased from 50%).

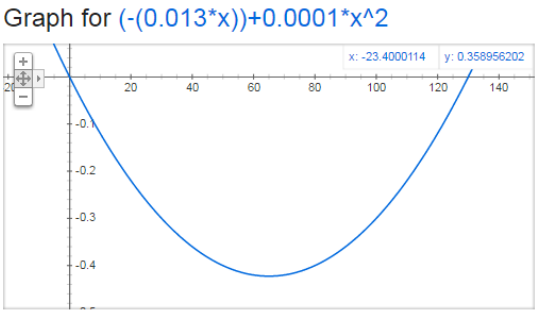
2. A)

Again, write the logit model as $p=\exp(\alpha+\beta\cdot x+\gamma z)1+\exp(\alpha+\beta\cdot x+\gamma z)$, where x=[Male, Female, AfAm, AfAm-Male]^T, β=[1.45, -2.11, 2.07, -0.872]. Thus the $\beta\cdot x$ component for:
African-American male=1.45*1+2.07*1+(-0.827)*1=2.693
Non-African-American male=1.45*1=1.45
African-American female=(-2.11)*1+2.07*1=(-0.04)
Non-African-American female=(-2.11)*1=(-2.11)

To conclude, the coefficient for AfAm_Male is just the interaction effect between race and gender. In other words, it is the **additional** effect after controlling for being Male and African-American. It does not mean that African-American Males are more likely to not graduate than African-American Females. In fact, other things being equal, African-American males (2.693) are more likely to graduate than both African-American female (-0.04) and Non-African-American male (1.45).

B)

The age effect is modeled by both a linear effect of Age (-0.013) and a nonlinear effect of Age_Sq (0.0001). Thus with other things being equal, when age increases, the probability of graduation decreases, but with diminishing marginal return, as depicted in the following graph:



However, it should be noticed that neither effect is significant at 5% level. So the age effect in general might be insignificant.

C)

To avoid multicollinearity, if students are classified as either male or female, I will drop one of the gender indicators (Male or Female).

I will also consider dropping Age_sq, which is highly correlated with Age. If Age becomes more significant (<5%) after this, then I may consider not including a non-linear effect of age.