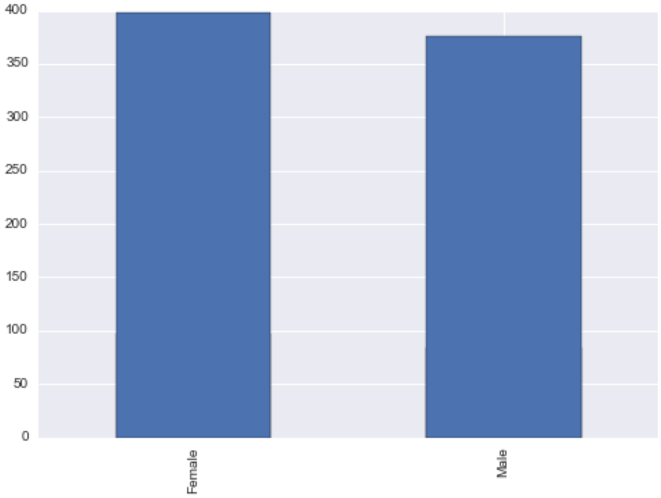
Abhi Gupta
CAPP 30254 HW 1

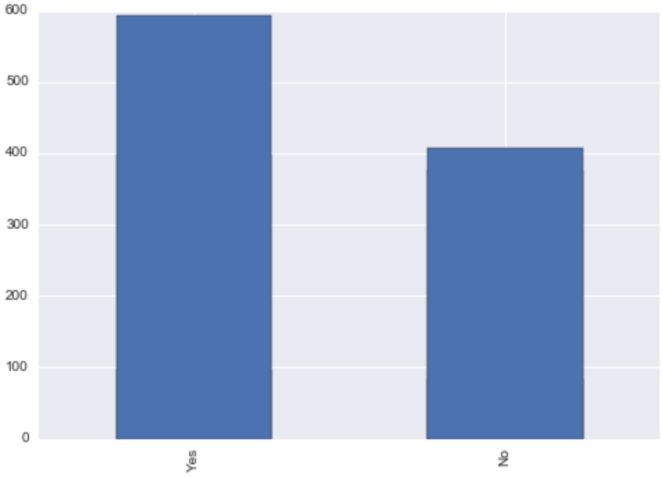
First, some summary statistics about the dataset:

	ID	First_name	Last_name	State	Gender	Age	GPA	Days_missed	Graduated
mean						16.9961	2.98845	18.0111	
median						17	3	18	
mode		Amy	Ross	Texas	Female	15	2	6	Yes
SD						1.45807	0.818249	9.62937	
num_missing	0	0	0	116	226	229	221	192	0

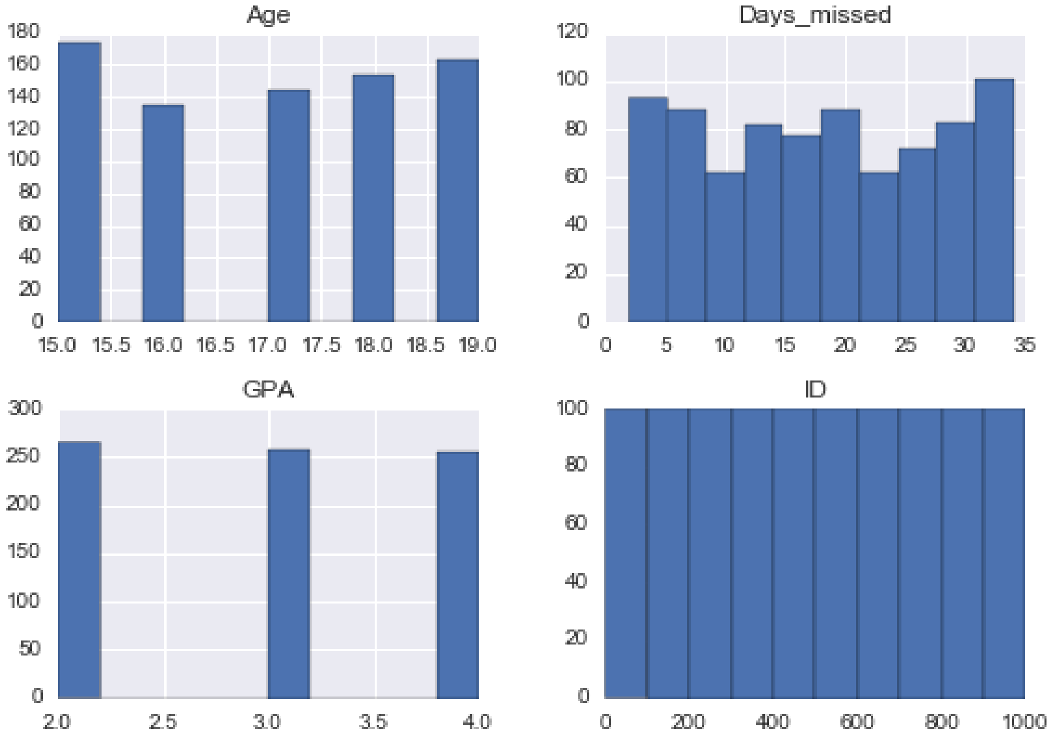
As can be seen, a not insubstantial number of the more auxiliary information (state, gender, GPA, days missed) are missing. The data that is complete is visualized below. This first plot shows that there are roughly equal number of men and women in the sample, leaving out those whose gender is not specified:

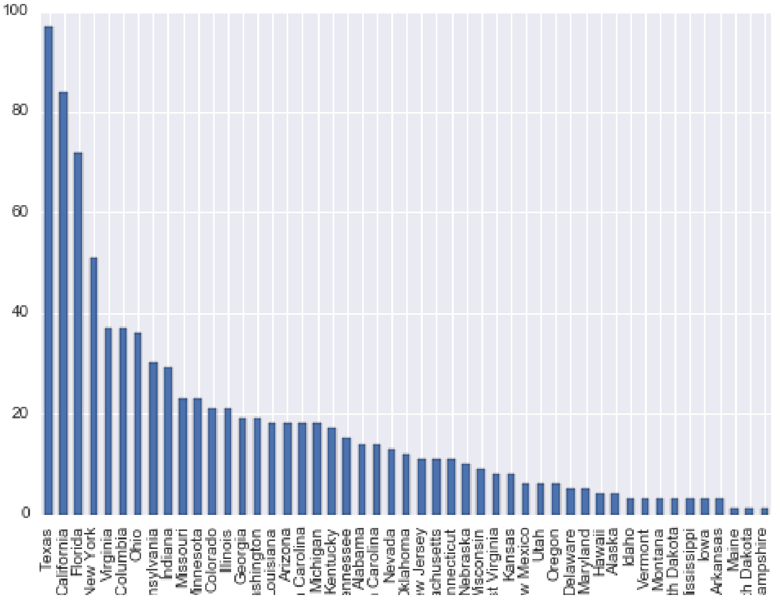


Unfortunately, about 40% of the students did not graduate.



The remaining fields (less first and last names) are visualized below. Note that everyone seems to have either a 2.0, 3.0, or a 4.0 GPA oddly enough, and that the first five or so states in the plot are home to the vast majority of students.





To impute missing values, a relatively standard and robust method is to use the mean of that category. This method can be improved by considering conditional means instead of unconditional means, as this makes better use of the data at hand. In this vein, I propose first using the most common state (Texas) to replace the missing values in the state category, and then compute means conditional on state, gender, and whether the student graduated in order to impute missing GPA, days_missed, and age values.

A. Consider 4 students, Adam, Bob, Chris and David. Adam and Chris share identical characteristics except for their family incomes. Bob and David also share identical characteristics (with each other, not necessarily Adam and Chris), except for their incomes.

Since the model is estimated on log income and not income itself, the drop from 50k to 40k has more of an effect than the drop from 200k to 190k. Since the coefficient on log income is negative, Chris will therefore be more likely to graduate than David.

A. The coefficient for AfAm_Male is negative. How do you interpret this? Does this mean that African-American Males are more likely to not graduate than African-American Females? What about relative to non African American males?

If we assume that students can always be identified as either male or female and either African American or not African American, the overall effect of a person's demographics can be seen by summing the coefficients that apply to them. As such, African American males' log odds change by (-.872+1.45+2.07)>0 vs African American females whose log odds change by (-2.11+2.07)<0. As such, African American males are more likely to graduate than African American females, and also more likely than males in general because the effect of their demographics on graduation, while positive at (2.07-.872)>0, is less than the effect that African American males receive from their demographics.

B. How do we interpret the difference in graduation probability between students of different ages? How do the variables in the model estimate such probability?

We can see the overall of a given student's age by summing the effects of age and age squared. On a normal range of ages (15-60, say), the signs and magnitudes of the coefficients indicate that older students are less likely to graduate. Because the squared coefficient is positive, the effect is convex; however, the effect is not positive for any age in the dataset.

C. Are there any variables in this model that you would choose to drop? Why or why not? Would you need more information in order to make this decision?

The age variables are the least significant, but since these are going to be at least moderately correlated over their likely range their low z values are not necessarily disqualifying. In order to assess the suitability of this model, the residuals and partial residuals ought to be examined. Some variables may prove to collinear or may need to be logged or squared or otherwise transformed in order to preserve the distributional properties of the estimated coefficients. A decent theory of why certain factors should or should not affect graduation probability would also help to further assess the model and guide further analysis.