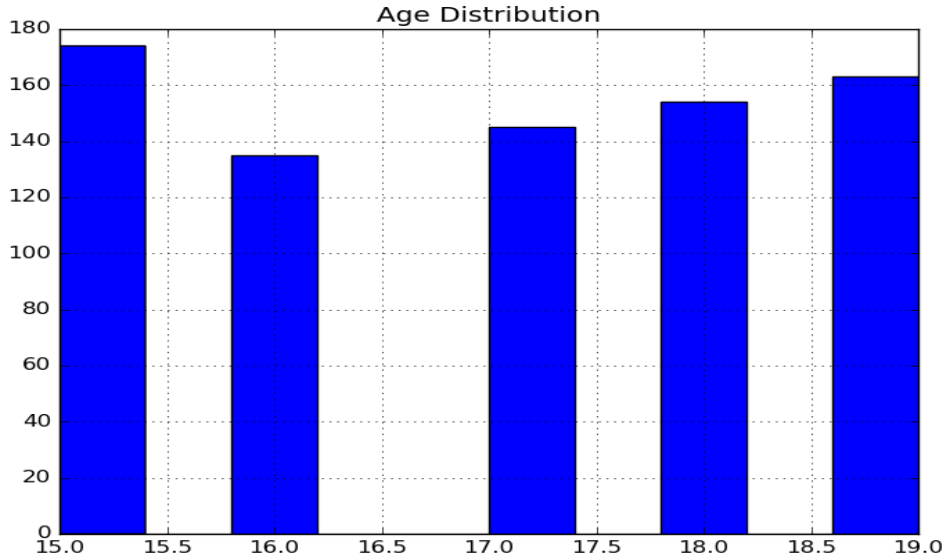
Paul Mack
HW1
4/4/2016

Summary of Variables of Interest:

A Link to the underlying code can be found here: https://github.com/pmack1/ML

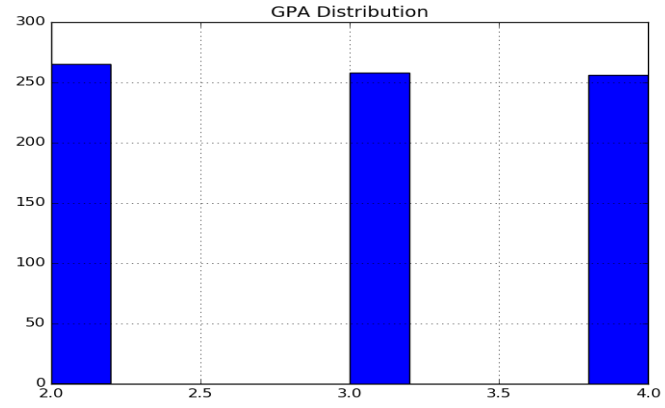
Age

	mean	median	mode	standard_deviation	null_count
0	16.996109	17.0	15.0	1.458067	229



GPA

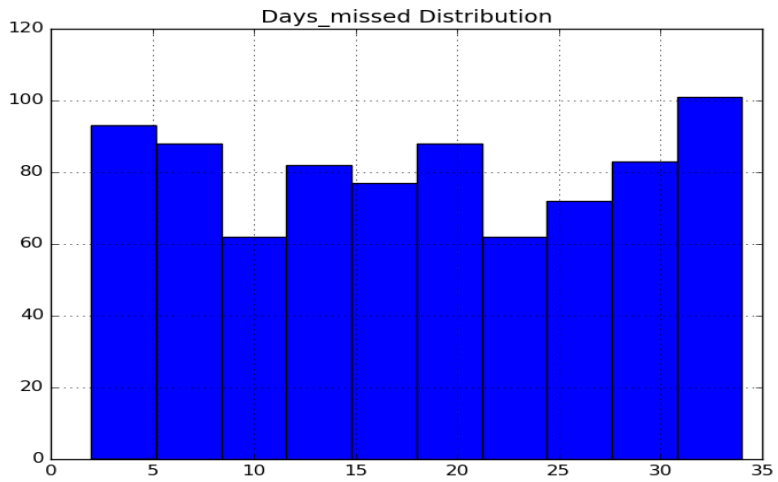
	mean	median	mode	standard_deviation	null_count
0	2.988447	3.0	2.0	0.818249	221



Days Missed

	mean	median	mode	standard_deviation	null_count
0	18.011139	18.0	6.0	9.629371	192
1	18.011139	18.0	14.0	9.629371	192
2	18.011139	18.0	31.0	9.629371	192

Days Missed had 3 modes: 6, 14, and 31



Filling in Missing Values:

All of the files below have had the missing gender fields filled in with the provided API. Additionally there are three .csv files that have had the "GPA", "Age", and "Days_missed" missing columns filled in in 3 ways:

1. "Student_Data_Filled_With_Mean.csv" fills in the missing values for "GPA", "Age", and "Days_Missed" with the total mean for each of these columns
2. "Student_Data_Filled_With_Conditional_Mean.csv" fills in the missing values for "GPA", "Age", and "Days_Missed" with the conditional mean of these columns on the condition that the student graduated
3. One other approach we could use for filling in the missing values is adding another condition to our conditional mean using our gender feature in addition to the graduated feature. That is implemented in the file "Student_Data_Filled_With_Conditional_Mean_With_Gender.csv"

Part II

We cannot tell whether Chris or David has the greater probability of graduating because we do not have enough information. We only know that:

A > C
B > D
A = B

therefore:
B > C
A > D

but we cannot compare C to D

A. The negative coefficient for African-American Male means that relative to African-American women, men are less likely to graduate. To compare Non-African-American Males to African-American Males you would need to compare the coefficients between Male and African-American Male.

B. The Age variable in our model contains a non-linear term, so as age increases the probability of graduating decreases up until a certain point until it will actually increase. However, in our case the positive coefficient on the non-linear term is small relative to the negative lineare coefficient, so the inflection point where the effect of age switches from negative to positive is likely outside the practical range of age values in the sample.

C. Including both Male and Female seems problematic as these values would be multicollinear and one should be dropped and used as the reference variable to the other, unless there are a lot of nonspecified values in the dataset and that is the comparison Male and Female is relative to. Also, age does not appear to be significant at the 95% confidence interval, but we may want to include it anyway as Age and Age_Sq might be significant if we jointly tested their significance