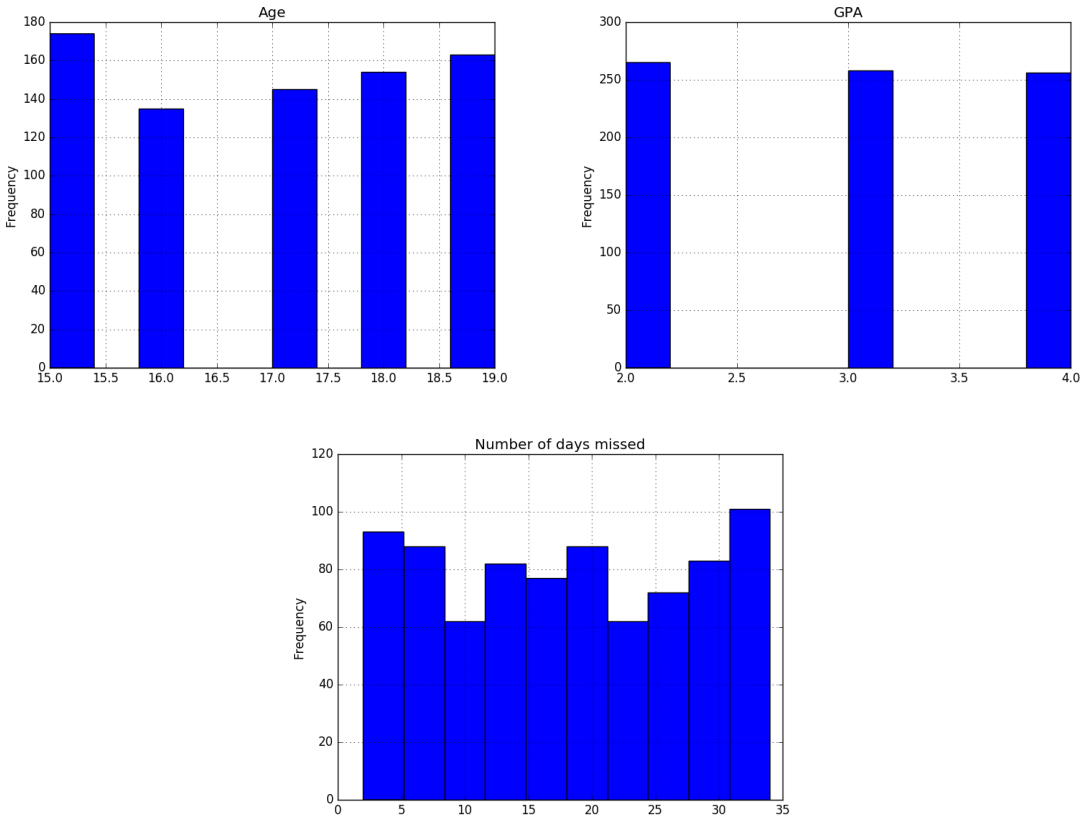
Problem A

Summary statistics and histograms for Age, GPA, and Days_missed[1]

Variable	Mean	Median	Mode(s)	Std. Dev.	Q1	Q3	Max	Min
Age (years)	17.00	17.00	15.00	1.46	16.00	18.00	19.00	15.00
GPA (points)	2.99	3.00	2.00	0.82	2.00	4.00	4.00	2.00
Days missed	18.01	18.00	6, 14, 31	9.63	9.00	27.00	34.00	2.00



To estimate the missing values of the variable *Age*, for example, a reasonable approach would be the following:

1. Take all the observations with no NaN's on *Age, State, Gender, Graduated, GPA*, and *Days_missed.* Call these 'complete observations'.
2. Take a subset of the complete observations to create a training model. Estimate how likely it is to have Age X given the rest of the variables (probably using a regression model).
3. Take a subset of the complete obervations and plug the average age. Calculate the mean squared error (MSE).

[1]Histograms are in the 'histograms' folder; datasets are in the 'dataframes' folder; code for generating this problem's data is in the 'code' folder.

4. Using that same subset, estimate Age using the training model. Calculate the MSE.
5. If the MSE with the estimation does better than the MSE where we just put the average age, then we found a better proxy for the *Age*. Otherwise, try to repeat the exercise with another subset of observations from the complete observations.

A similar approach could be implemented for the other variables, *GPA* and *Days_missed*.

(Problem B on next page)

Problem B
A larger data set than the one in the previous problem was used to build a logistic regression model that predicts the probability an individual student will graduate.

$$P(Graduation) = \beta_0 + \beta_1 Male + \beta_2 Female + \beta_3 \ln(FI) + \beta_4 age + \beta_5 age^2 + \beta_6 CC + \beta_7 AfAm + \beta_8 AfAm * Male$$

1. **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. Based on the coefficients provided, who would you think has a higher probability of graduating? What is your reasoning?**

   From the information provided, we know that:

   - P(Bob graduating) < P(David graduating)
   - P(adam graduating) < P(Chris graduating)

   From the logistic regression, we know that the marginal effect of FamilyIncome is not constant. Therefore, the marginal effect of a decrease of $10,000 will probably be different from $50,000 to $40,000 than it is from $200,000 to $190,000.

2. **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?**

   This coefficient indicates that being a male African American reduces on average the probability of graduation. Because of the way this regression is specified, probability of graduation for an African American male would take into account three covariate effects: Male (1.45), AfAm (2.07), and AfAm_Male (-0.872). Probability of graduation for an African American female would take into account: Female (-2.11) and AfAm (2.07). Therefore, this regression suggests that African American males are more likely to graduate than African American females.

   What about relative to non African American males?

   Probability of graduation for an non-African American male would onlye take into account the effect of Male (1.45). This means that African American males are more likely to graduate than non-African American males

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

   Holding everything else constant, any increase in age has a negative effect on the probability of graduation until age reaches 65. After 65, an additional year of age has a positive effect on probability of graduation.

4. **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?**

   I would drop either gender dummy variable, as there probably is a multicollinearity problem on data for using both males and females as dummys. Given the current specification, I would go for dropping females. If every observation of the dataset had a specified gender, this regression would probably have not have been possible to make.